Ericsson REVIEW

THE TELECOMMUNICATIONS TECHNOLOGY JOURNAL

NO 1, 1997

ERICSSON

CONTENTS

No. 1 1997 • Vol.74

Cover: The display system in the JAS 39 Gripen is manufactured by Ericsson Saab Avionics. The sighting indicator enables the pilot to keep his gaze fixed in front of him and at the same time see important information in the indicator. The next generation of indicators will display the information directly in the visor of the pilot's helmet. Helmet displays of this kind may also be used in many different civilian applications.

CONTENTS
Previous issues

No. 1, 1996

No. 2, 1996

No. 3, 1996

No. 4, 1996

Ericsson Review © Telefonaktiebolaget L M Ericsson • Stockholm 1997 • Responsible publisher *Håkan Jansson* • Editor *Steve Banner* • Editorial staff *Eric Peterson, Eva Karlstein* • Layout *Paues Media* • Address Telefonaktiebolaget L M Ericsson S-126 25 Stockholm, Sweden • Fax +46 8 681 27 10 • Published in English and Spanish with four issues per year • Subscription one year USD 45

CONTRIBUTORS
in this issue

Mats Frisk

Seved Torstendahl

Gunnar Forsberg

Hans Ivarsson

Ove Lövenheim

My Spangenberg

Fredrik Thernström Strandh

Hans Brandtberg

Mats Frisk is the regional manager of marketing and sales of wireless messaging networks for the Americas and France at Ericsson Radio Messaging AB. He is also a member of the board of directors for the pACT Vendor Forum. He holds an MSc in Electrical Engineering from the Lund Institute of Technology. He has been with Ericsson since 1982.

Seved Torstendahl is currently the product manager for the open telecom platform at Ericsson Telecom AB. Since joining Ericsson in 1975, he worked on the APN 163, was responsible for the development systems for the APN 167 and EriPascal, and served as a member of the system management team for TMOS at Ericsson Hewlett-Packard Telecommunications. He holds an MSc in Physics from the Royal Institute of Technology in Stockholm.

Gunnar Forsberg is currently a consultant in fibre optic system design, participating in a multi-channel WDM project at Ericsson Infocom AB. Before becoming a consultant, he worked at Ericsson in several fibre optic-related projects. He holds an MSc in Electrical Engineering from the Royal Institute of Technology in Stockholm.

Hans Ivarsson managed the project to develop the EriOpto4 optical link. Today he works as a senior advisor for Strategic Planning, Products and Technology at Ericsson Research and Development AB. He is also Ericsson's co-ordinator of electrotechnical policy issues in the IEC and CENELEC, and represents Ericsson in the Electrotechnical Council of SEK – the Swedish national committee of the IEC and CENELEC.

Michael Lynn managed the electrical design subproject for the EriOpto4 project. He currently works with the ATM switch, system and hardware design business line at Ericsson Telecom AB. He completed the Electrical Engineering programme at Åsö upper secondary school in Stockholm.

Anders Lindström is the marketing manager of customer services for Mobile Telephone Systems PDC at Ericsson Radio Systems AB. Previously, he worked in business development and as the manager of a project that introduced the CMS 30 into the Japanese market. He has also worked as a regional manager of customer services in the USA. He holds a BSc in Electrical Engineering and a degree certificate in Marketing Management from the IHM Business School.

Olle Lövenheim is the manager of strategic planning for customer services for GSM, NMT and TACS at Ericsson. During his seven years with Ericsson, he has worked as market communications manager, product manager and marketing manager. He holds BSc degrees in Technology and Social Science.

My Spangenberg, who joined Ericsson in 1989, is the communications manager of customer services for Public Networks. Previously, she worked as the communications manager at LM Ericsson Data AB. She holds an MA in English as well as a degree certificate in Marketing Management from the IHM Business School.

Fredrik Thernström Strandh is currently the manager of service marketing and market co-ordination for Cellular Systems American Standards at Ericsson Radio Systems AB. Since joining Ericsson, in 1989, he has served as the manager of customer services for Land Mobile Radio Systems, of after sales for Mobile Telephony Systems, and of service business development for Cellular Systems American Standards.

Hans Brandtberg, who joined Ericsson in 1974, is currently a senior advisory specialist in the field of display and recon-

Michael Lynn

Anders Lindström

Peter Segerhammar

Claes Waldelöf

naissance systems at Ericsson Saab Avionics AB. Previously, he was manager of the Airborne Display Systems Design Group and of the Systems Development Section. He holds an MSc in Electronic Engineering.

Peter Segerhammar is the marketing manager at the Displays and Reconnaissance Systems Division of Ericsson Saab Avionics AB. Previously, he was programme manager for the first head-mounted display system developed within Ericsson. He has MSc degrees in Applied Physics and Electrical Engineering as well as in Bio-Medical Engineering.

Claes Waldelöf, who is currently the programme manager for helmet and head-mounted display systems for commercial and military applications, has been with Ericsson since 19xx. He has an MSc in Aeronautics Engineering from the Royal Institute of Technology in Stockholm. In addition, he has studied Aeronautics Engineering at Ecole Nationale Superieure de l'Aeronautique et de l'Espace in Toulouse, France, as well as Pedagogy at Linköping University.

FROM THE EDITOR

Steve Banner

"The future just ain't what it used to be". Although we hear this quotation most often used in a humorous context, it also alludes to our inability to correctly predict events to come.

For instance, I still recall a statement made to me two years ago, dismissing the Internet as unworthy of a great deal of attention from telecommunications companies because it was a medium used only by "eccentric engineers and college students". Admittedly, this remark was made at a conference social function, but its inherent scepticism is indicative of the conventional wisdom that prevailed among many senior business managers at that time. The exponential growth that has since taken place in the numbers of Internet applications and users with access to the World Wide Web is either indicative of a massive explosion in the numbers of certain sections of society, or a sign that the preceding viewpoint was somewhat awry!

Of course history is filled with similar examples of short-sightedness. The projection during the late 1940s that there would never be a market for more than a handful of different computer manufacturers in the world is but one example of how advances made in technological development have often been ahead of our ability to comprehend their possible future uses.

To return to the present, virtual reality devices such as the Ericsson head-mounted display systems that are described in this issue, make use of technology to synthesise real-world phenomena and create an artificial environment for the user. These systems could be used in a static situation to create a training environment for driver training, or in the real world to provide a night-vision view of a helicopter pilot's actual surroundings — to suggest only two of the vast number of possible applications of this technology.

The head-mounted display could also be used in the home to watch video-on-demand, delivered over an optical fibre. Of course the optical fibre has been with us for some time, being widely used in transmission networks of various types. But generally speaking, the costs associated with creating the prerequisite network infrastructure have not yet allowed optical fibre to fulfil its promise of providing many possible applications for multimedia services to the office, home or building. However the robust and reliable new EriOpto optical transceiver system, designed for use over a single fibre and able to be mass-produced at low cost, constitutes a significant step forward in the breaking down of barriers to widespread penetration of optical-fibre technologies.

Another technology often regarded as having reached its limits for application is that associated with wireless paging. However recent developments in the field of two-way paging have opened up many new applications for this type of communication, with the evolution into narrowband PCS and applications including two-way messaging, among others. The pACT protocol offers the efficient use of frequency spectrum and a cellular-based structure that allows its coverage to adapt and grow with the business needs of the operator.

Equally important to telecom operators' business operations are the flexibility and adaptability of its network elements. The ability to develop quickly and market new services and features is crucial to success. The Ericsson Open Telecom Platform for building and controlling telecommunications switching applications is designed to reduce the time to market of new products with new switching software, while allowing that software to be developed on standard, commercially available computer platforms.

Along with the rapid pace of technological development, market forces are also shaping the telecommunications industry. With worldwide trends towards deregulation and the liberalising of telecommunications markets, many new operators are entering the industry and numerous existing operators are entering new areas of operation. In such a fast-moving and competitive environment, operational support services - such as the comprehensive portfolio that Ericsson offers to network operators - can be vital to the business success of both new and existing operators, as described in this edition of Review.

As the contents of this issue demonstrate, the advance of technology offers enormously exciting promise of unique and affordable services that enable us to communicate better than ever before with our fellow human beings. Who knows, maybe our improved communication will encourage us to listen to the only group who has been able to show that they understood at least a part of the future - those "eccentric engineers"!

Steve Banner, Editor

Personal air communications technology – pACT

Mats Frisk

Advances in technology have allowed paging markets to evolve in recent years from simple one-way paging to more sophisticated services, such as two-way paging and mobile data. Eager to put these advances into practice, AT&T Wireless Services, Ericsson, PCSI, DATUM Telegraphic and others have joined forces to develop pACT, a new open standard for two-way paging and messaging services. pACT has its roots in cellular digital packet data.

The pACT protocol was developed mainly to enable compact, inexpensive devices with a long battery life to access reasonably low-cost, high-capacity network infrastructures. This two-way protocol enhances one-way paging, response paging, two-way paging, voice paging, telemetry and two-way messaging applications.

The author describes the pACT system, whose efficient use of valuable frequency spectrum and cellular-based structure ensure that its coverage is able to grow and adapt to operators' business needs.

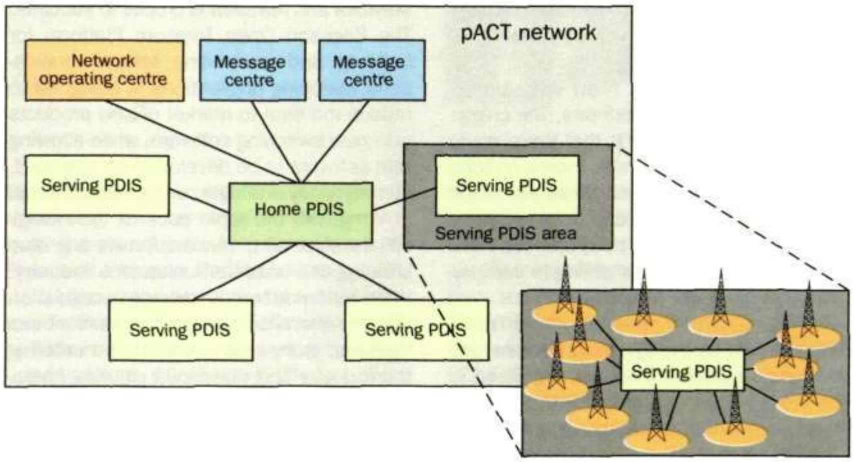


Figure 1
The PDIS forms the core of a pACT network. The home PDIS serves several serving PDIS which, by means of several radio base stations, provide radio communication to a given geographical area.

Narrowband PCS

In 1994, the Federal Communications Commission (FCC) set aside frequency spectrum for narrowband personal communications services (narrowband PCS, or NPCS) in the USA. The narrowband PCS spectrum, which consists of a 2 MHz outbound frequency spectrum and a 1 MHz inbound frequency spectrum, Figure 2, was then divided into 34 channels and allocated to five regions, 51 major trading areas (MTA), and 493 basic trading areas (BTA):

- 11 nationwide channels
- 6 regional channels
- 11 MTA channels
- 6 BTA channels

The FCC then held auctions for regional and national coverage, granting licences to 13 companies, most of whom were well-established US paging operators. All told, the operators paid more than USD 1 billion for rights to operate in the narrowband PCS spectrum, see Tables 1 and 2. When broken down, this amount roughly corresponds to USD 2-4 per person per MHz. Further auctions for the MTA and BTA channels will be held in 1997.

In Canada, too, authorities allocated (no auctions) the narrowband PCS spectrum within the same frequency band to various companies, including Cantel, LanSer Communications and Bell Mobility.

A great deal of money has been invested in the relatively narrow radio channels. The paging industry has seen substantial

Box A Abbreviations

API	Application program interface
ATM	Automatic teller machine
BHCR	Busy hour call rate
BST	Base station transceiver
BT	Both-way trunk
BTA	Business trading area
CDPD	Cellular digital packet data
CMIP	Common management information protocol
DRAM	Dynamic random access memory
ERMES	Enchanced radio messaging system
ERP	Effective radiated power
FCC	Federal Communications Commission
FPGA	Field programmable gate array
FTAM	File transfer, access and management
GMSK	Gaussian minimum shift keying
GPS	Global positioning system
GUI	Graphical user interface
HPA	High-power amplifier
IP	Internet protocol
IVR	Interactive voice response
LAPD	Link access procedure on the D-channel
LSM	Limited size messaging
MC	Message centre
MDLP	Mobile data link layer protocol
MIB	Management information base
MTA	Major trading area
NMS	Network management system
NPCS	Narrowband PCS
OSI	Open systems interconnection
pACT	Personal air communications technology
PCS	Personal communications services
PDA	Personal digital assistant
PDB	pACT data board
PDBS	pACT data base station
PDIS	pACT data intermediate system
POCSAG	Post Office Code Standardization Advisory Group
PSTN	Public switched telephone network
RF	Radio frequency
RRM	Radio resource management
SMS	Short message services
SMTP	Simple mail transfer protocol
SNMP	Simple network management protocol
SQL	Structured query language
TAP	Telocator alphanumeric protocol
TNM	Telecommunications network management
TNPP	Telocator network paging protocol
TRX	Transceiver
TSM	Time slot multiplexing
UDP	User datagram protocol
WWW	World Wide Web

growth in recent years (27% annual increase in the US between 1990 and 1995), and analysts expect this sector to continue to grow for several years to come. Nonetheless, some paging carriers are already running short of bandwidth, a dilemma that has them actively examining ways in which to add capacity to their networks. In dense markets, such as major US cities, the solution has been to make better use of existing spectrum, which is where narrowband PCS and personal air communications technology (pACT) play an important role.

The pACT protocol was originally created to address the demands of a growing market for narrowband PCS. The first pACT specification was released in October 1995. A second release, pACT97, followed in January 1997.

pACT services and applications – "one-way paging"

Narrowband PCS is primarily seen as a means of meeting paging operators' urgent need for increased capacity. As such, a "two-way" infrastructure is deployed mainly to support a "one-way" application.

In traditional one-way paging networks, such as POCSAG (Post Office Code Standardization Advisory Group), ERMES (enhanced radio messaging service, Box B), and FLEX (Box C), messages are delivered via simulcast distribution from several transmitters with very high output power – up to several hundred watts effective radiated power (ERP). Simulcast transmissions make paging networks very efficient in terms of low operational costs and of providing good coverage from a reasonably inexpensive infrastructure. However, because they do not make efficient use of bandwidth, simulcast transmissions fall short of meeting the huge demand for paging services. This is particularly true because one-way networks have no way of knowing where a subscriber is located at any given time. Thus, to ensure that messages reach a subscriber, network operators must transmit each message over every transmitter in the network. In the US, for example, approximately 7% of some 42 million paging customers (1996) subscribe to nationwide services. Sending a message to one such subscriber via a traditional one-way paging network requires the use of as many as 2000 transmitters. Obviously, this is a waste of expensive bandwidth.

The capacity of one-way paging systems varies in accordance with the busy hour call rate (BHCR), the average message length, and the customer service profile. Local or regional subscribers, for instance, do not require the same overall level of bandwidth as nationwide subscribers. However, in zones – that is, in markets with continuous coverage – many simulcast-oriented one-way systems have already reached a ceiling on capacity. In these zones, the only way to increase capacity is to add more frequency spectrum. Adding more transmitters improves coverage, but does not affect capacity. Unfortunately, in many countries and markets around the world, more frequency spectrum simply is not available for paging or for other wireless services.

Nationwide paging systems that encompass many zones support more users than local or regional systems. Likewise, the nationwide services require much more capacity than regional or local services. A nationwide numeric service often requires as much as 25-50 times more capacity than local single-market services. Alphanumeric services, which accomodate larger messages and character sets, require still another 5-10 times as much capacity. Thus, nation-

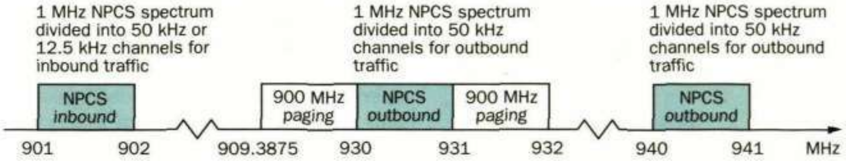
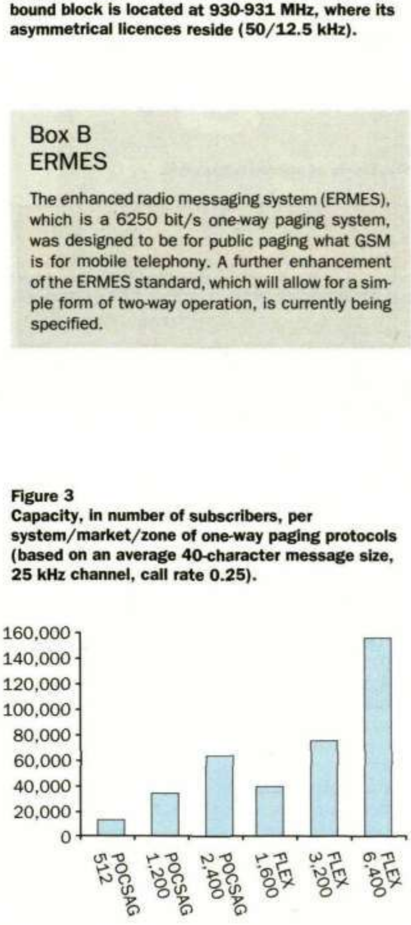

1 MHz NPCS spectrum divided into 50 kHz or 12.5 kHz channels for inbound traffic
1 MHz NPCS spectrum divided into 50 kHz channels for outbound traffic
1 MHz NPCS spectrum divided into 50 kHz channels for outbound traffic
NPCS inbound
900 MHz paging
NPCS outbound
900 MHz paging
NPCS outbound
901
902
909.3875
930
931
932
940
941
MHz

Figure 2
The outbound portion of the 2 MHz frequency spectrum is divided into two blocks. One block, at 940-941 MHz, provides symmetrical 50/50 kHz paired channels with the 1 MHz inbound block at 901-902 MHz (39 MHz fixed duplex space). The other outbound block is located at 930-931 MHz, where its asymmetrical licences reside (50/12.5 kHz).

Box B
ERMES

The enhanced radio messaging system (ERMES), which is a 6250 bit/s one-way paging system, was designed to be for public paging what GSM is for mobile telephony. A further enhancement of the ERMES standard, which will allow for a simple form of two-way operation, is currently being specified.

Figure 3
Capacity, in number of subscribers, per system/market/zone of one-way paging protocols (based on an average 40-character message size, 25 kHz channel, call rate 0.25).



Box C
The FLEX family of paging protocols

FLEX, ReFLEX and InFLEXion are paging protocols from Motorola.

FLEX is a protocol for one-way paging infrastructures. The siblings ReFLEX and InFLEXion are based on a two-way paging infrastructure. Unlike pACT, however, in order to locate subscriber devices in the network, the ReFLEX and InFLEXion protocols require a separate receiver (inbound messaging) network – normally with a very low transmission speed (800 bit/s).

The FLEX family of protocols is capable of supporting inbound transmission speeds of up to 9600 bit/s. However, to do so, while providing the same grade of service and maintaining the same geographical area that is covered by an 800 bit/s implementation, requires that many more receivers be deployed.

FLEX, ReFLEX, and InFLEXion are trademarks of Motorola Inc.

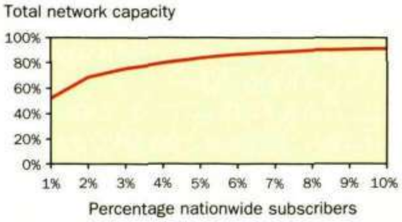
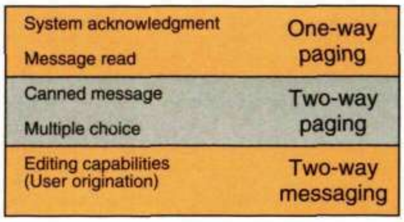


Figure 4
Channel capacity required to support a 1-10% nationwide subscriber base in a 100-zone one-way paging system. All remaining users are assumed to be local (single-zone) subscribers.

Figure 5
Five levels of acknowledgement.



System acknowledgment Message read	One-way paging
Canned message Multiple choice	Two-way paging
Editing capabilities (User origination)	Two-way messaging

wide alphanumeric services generally require 125-500 times more capacity than local numeric services. However, the operators of such services can typically only charge 5-10 times more than local numeric operators charge. Thus, although the demand for nationwide and alphanumeric services is high and growing, paging operators of simulcast-oriented one-way systems are reluctant to provide them, knowing that they can yield a greater return from investments in local numeric services.

Narrowband PCS and two-way protocols, such as pACT, give the paging industry new tools for increasing the capacity of one-way paging systems, and introduce possibilities for providing new or enhanced services. Like traditional mobile telephony systems, pACT increases capacity by reusing frequencies (providing theoretically unlimited capacity). Given higher airlink speeds (8000 bit/s) and knowing a subscriber's exact location, operators can increase capacity in a large zone 100-200 times, and even more in networks that are made up of several zones.

Table 1
Summary of nationwide narrowband PCS licences

Licences	Company	Winning bid, USD
50/12.5 kHz paired	AirTouch Paging	47,001,001
50/50 kHz paired	AT&T Wireless Services	80,000,000
50/50 kHz paired	AT&T Wireless Services	80,000,000
50/12.5 kHz paired	BellSouth Wireless (MobileComm)	47,505,673
50/50 kHz paired	Mobile Telecommunications Technologies	80,000,000
50/12.5 kHz paired	Mobile Telecommunications Technologies	47,500,000
50 kHz unpaired	Mobile Telecommunications Technologies (Pioneer's preference licence)	33,300,000
50 kHz unpaired	PageMart Inc.	38,000,000
50/50 kHz paired	Paging Network (PageNet)	80,000,000
50/50 kHz paired	Paging Network (PageNet)	80,000,000
50 kHz unpaired	Paging Network (PageNet)	37,000,000
		650,306,674

Table 2
Summary of regional narrowband PCS licences

Regions	Licence	Company	Winning bid, USD
3	50/12.5 kHz paired	AirTouch Paging	31,218,001
5	50/12.5 kHz paired	American Paging	53,621,666
1	50/12.5 kHz paired	Ameritech	9,500,000
1	50/12.5 kHz paired	Benbow PCS Ventures*	35,681,000
2	50/12.5 kHz paired	Insta-Check Systems*	8,000,013
3	50/12.5 kHz paired	Lisa-Gaye Shearing*	52,940,007
5	50/12.5 kHz paired	MobileMedia	53,669,092
5	50/50 kHz paired	PageMart Inc.	92,599,020
5	50/50 kHz paired	PCS Development Corp.*	151,544,001
		*40% bidding credit	488,772,800

To the end-user, one-way paging over a two-way network offers no real benefits; however, to operators the benefits are tremendous:

- Two-way networks enable paging devices to acknowledge that a message has been received. With this capability, operators can offer their customers "guaranteed delivery". However, because most subscribers assume they already have this service, operators may find it difficult to charge them more for it. In time, guaranteed delivery may become a fundamental service.
- To reach an intended recipient, operators do not have to send a message via every transmitter in the network. Instead, because they know the exact location of each subscriber device (subscriber devices register automatically as they move about in the network) messages are sent solely via the transmitter that is closest to the subscriber, freeing up a great deal of network capacity. Thus, all other transmitters in the network can be used simultaneously to serve other subscribers.

pACT-related services and applications – other services

The two-way pACT architecture affects more than capacity: it also enables providers to enhance services, as well as to provide completely new services and applications. For example, paging devices that contain a transmitter are able to transmit information back to the network, to someone else in the network, or to any other network.

The two-way paging and messaging paradigm is divided into five levels of acknowledgement, Figure 5:

- System acknowledgement – the subscriber device acknowledges reception of an error-free message. A transparent link layer acknowledgement, between the device and the network, enables guaranteed delivery service. The network stores and retransmits messages that the subscriber device has not acknowledged.
- Message read – when a recipient reads a message, the paging device transmits a "message read" acknowledgement back to the host system or to the originator of the message.

- Canned messages – the paging device contains several ready-to-use responses, such as "Yes", "I'm ready", etc, that recipients can use when they reply to an inquiry.
- Multiple choice – the originator of a message defines several possible responses to accompany his or her message. To reply, the recipient simply selects the appropriate response.
- Editing capabilities – some devices may be used to create messages. Editing capabilities vary from device to device; for example, some subscriber devices are managed by a few simple keys and provide only minor editing capabilities; other devices contain full-feature keyboards; and still others, such as pen-based screens, receive input in other ways.

Acknowledgement levels 1 and 2 represent an exclusive one-way service that complies with the current one-way paging model; however, it may be possible to enhance the functionality of these levels to provide longer message transfer, guaranteed delivery and various other broadcast services. Acknowledgement levels 3 and 4 offer additional opportunities for simple user interaction. Acknowledgement level 5 facilitates symmetrical two-way messaging applications.

Countless applications may be supported by a pACT network, Box D. Obviously, depending on the availability of frequency spectrum, some applications are more feasible than others. For instance, operators with only a 50/12.5 kHz channel are more likely to focus on applications that do not put great demands on the inbound (reverse) channel. Similarly, a pACT network, or any other network in which available frequency spectrum is limited, is not suited to applications with high volumes of data flowing in both directions, such as file transfer.

Another consideration is maximum battery lifetime (latency) versus maximum performance. Subscribers cannot expect batteries to last in devices whose applications are optimised to respond quickly. pACT provides support for either extreme: long battery life or short response times.

pACT system overview

The pACT system uses built-in network intelligence to manage an efficient, reliable flow of messages to and from sub-

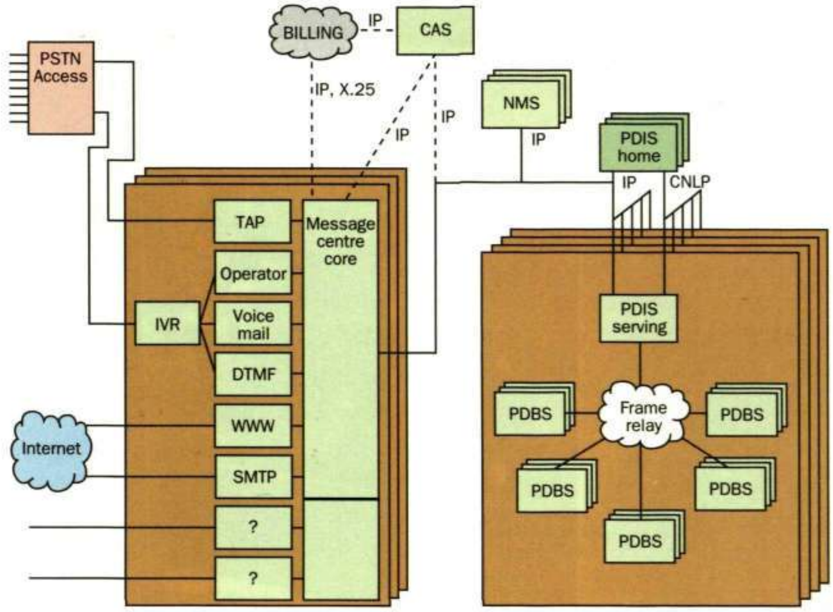


Figure 6
The message centre, which is the gateway or interface to the pACT network, permits several interfaces and applications, such as Internet applications, to be implemented. A pACT network may be configured in several ways, using message centres and sub-networks of various sizes.

Box D Services and applications in a pACT network

Enhanced paging applications
- On-demand local and national news information
- Emergency updates
- Content services
  - on-demand stock and financial information
  - interactive regional movie listings, local traffic updates
  - on-demand local and national weather information
  - on-demand sports scores and other event information
  - interactive airline information, services and reservations
  - gambling and lotteries
  - geographical- and profile-specific advertising
- Acknowledgement paging with response through operators, fax, WWW, and SMTP
- Mailbox manipulation
  - remote manipulation of voice mail and fax message forwarding
  - e-mail manipulation—auto-forwarding of header/size information allows subscribers to decide whether or not to select all or part of a message
  - downloading of fax messages for display on the pager, or to another device
- Pager-to-pager messaging
- Voice paging – delivery of voice messages to pocket-size devices

E-mail and messaging applications
- Devices that connect to portable data terminals (laptop and palmtop computers, PDAs, etc)
- Various devices with integrated chip sets for wireless communication
- Wireless modems

Fixed-point applications
- Vending machine monitoring
- Home monitoring and interactive services (electricity, gas, water, air conditioning, lights)
- Generic global positioning system (GPS) services
  - location service
  - vehicle system monitoring—emergency road assistance, auto-theft prevention, SOS service
- Utility substation monitoring (interactive facilities management)
  - electricity
  - gas, water (pipe pressure)
- Environmental monitoring and large equipment monitoring (information and error reporting)
- Road and remote commercial electronic sign information updates
- Point-of-sale and mobile automatic teller machines (ATM)

Combined asynchronous voice and data applications
- Response with text-to-voice and text messages
- Interaction with home or office voice mail system (defined filters, priorities, etc)
- Cost-based fee notification with option to download

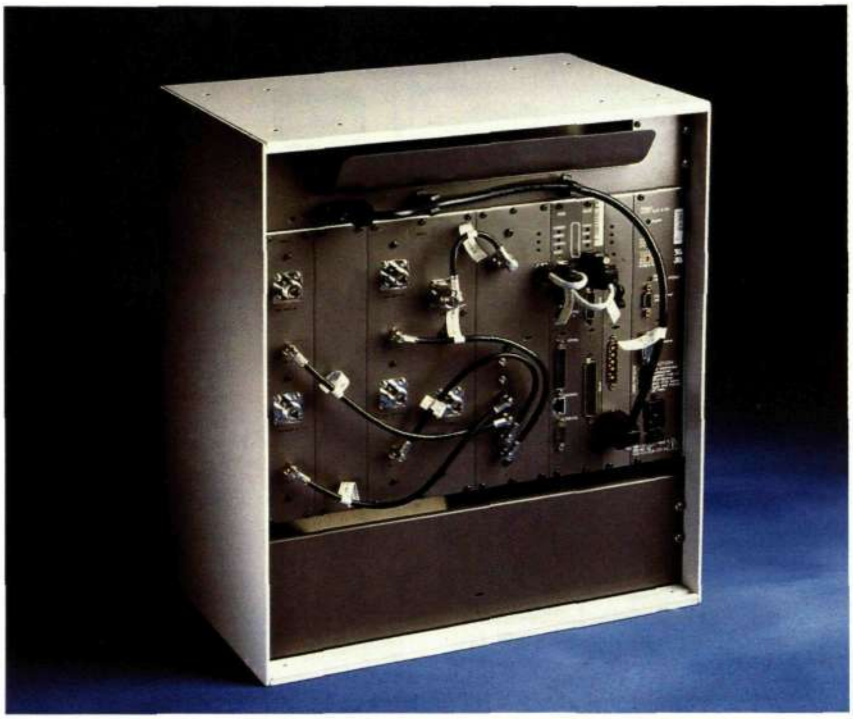


Figure 7
The Ericsson RBS 540 base station for the narrowband PCS pACT standard.

Box E
RBS 540 technical data

Physical	
Dimensions	18.90/17.32/ 11.81 inches
Weight	48 kg (106 lbs)
Power	
Power supply	120V AC or +24V DC
Power consumption	Maximum 250W
Radio specifications	
Frequency range	RX 901-902 MHz TX 930-931 or 940-941 MHz
Channel spacing	12.5 kHz
Receiver sensitivity (at 5% BER) with 1/3 convolutional code	-124.5 dBm
Transmitter output power	55W standard (150W optional)
Radio data transmission	8 kbit/s nominal
Alarms	Internal as well as external
Environmental requirements	0 to 50°C
Interfaces	
Physical	V.35

scriber devices. Cellular radio system design and roaming techniques enable pACT to determine precisely which base station is closest to a subscriber device each time communication takes place between it and the pACT network. A key feature of the pACT architecture is that network receivers are co-located with the transmitters.

The pACT system is built from several flexible modules that can be combined and configured in different ways to meet specific operator demands, Figure 6. Because the pACT network is based on the Internet protocol (IP), operators and application providers can take full advantage of existing tools, applications, and application program interfaces (API).

pACT provides secure service, including encryption and authentication – a method which ensures that messages are delivered solely to intended subscribers.

Ordinarily, pACT data base stations (PDBS) are connected to the serving pACT data intermediate system (PDIS) switch via a frame relay network. The serving PDIS is connected to the home PDIS switch. If necessary, the two switches may be located on the same hardware platform. In addition, several other standard servers are needed to operate the network, including the network management system (NMS) and the customer activation system. The PDIS and all associated servers run on standard Sun SPARC or Sun Solstice platforms. Various configurations of computer processor power and memory are available for the PDIS, depending on requirements for computing capacity and on how the requirements relate to traffic load and the number of subscribers in the network. More computing capacity may easily be added if necessary.

The fixed entry point into the pACT network is provided by one or more message centres, which initiate, provision, and connect pACT services to public and private networks as well as to the public switched telephone network (PSTN).

pACT system components

The RBS 540 – Ericsson's pACT data base station

The RBS 540, Ericsson's pACT data base station (PDBS), is located at the cell site and relays data between subscriber devices and the PDIS. The RBS 540, Figure 7, is a small, self-contained full-featured base station. Thanks to its small, easy-to-manage size, the RBS 540 can be installed and put into operation in a few hours. The base station was designed for an omnidirectional site with one radio carrier, but if additional capacity is necessary, it may easily be expanded to support two or three sectors.

The RBS 540 base station was designed with an eye to helping operators to keep their operating costs low. This means that the number of base stations (cell sites) must be kept at reasonable levels. The RBS 540 supports sectored solutions with high-gain panel antennas, and up to six receivers in order to take full advantage of receiver diversity and to provide the best possible coverage from each base station transceiver (BST). An adjustable power output up to 55W is used to balance the link (two-way services), and an optional high-power amplifier (HPA) can be added, delivering up to 150W to the antenna. The pACT protocol also contains several features that increase coverage.

By means of network management, each base station is provided with a set of radio resource management (RRM)

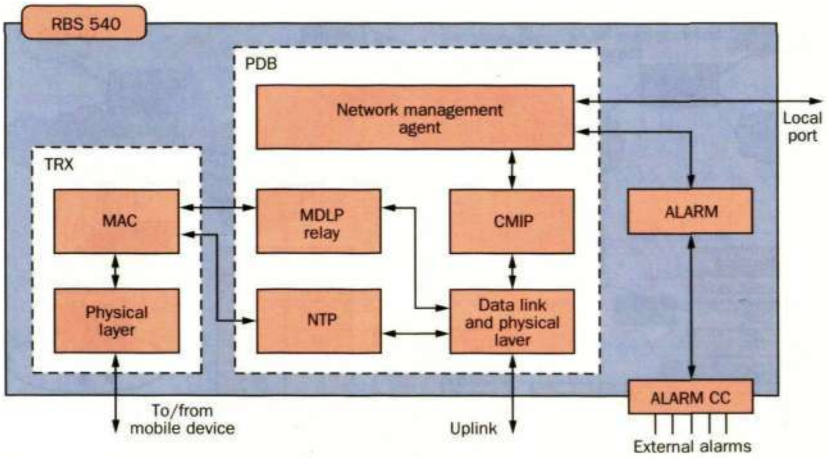


Figure 8
There are basically two main parts of the Ericsson RBS 540 pACT base station:
- the transmitter (TRX) and six receivers for handling radio communication with advanced signal processing;
- the PDB, which controls the flow of traffic and communication in the pACT network.

parameters. These are used to control traffic and maintain links to the network as well as to give orders to mobile terminals that access the channel. Cell transfer is determined by the mobile terminals, based on signal-strength measurements of nearby base stations.

The transceiver unit (TRX) contains a transmitter and six receivers. High sensitivity is achieved by means of six-level branch combining and maximum likelihood (Viterbi) decoding. That is, an input signal is sampled through each of six antennas (six-level branch). Discrepancies between the samples (micro-diversity) are then analysed (maximum-likelihood decoding), and a final signal is created by combining the samples' most pronounced common characteristics (maximum ratio). Three digital signal processors process the signals – one TMS320C50 and two field programmable gate array (FPGA) circuits: XC5210 and XC4025. The filter design, which provides high attenuation of adjacent bands, enables the base station to be co-located with other radio equipment. Duplex filters minimise the number of antennas that are needed. When additional transceiver units are connected to it, a single RBS 540 can support up to three independent channel streams.

The pACT data board (PDB), which controls traffic flow and handles communication with the pACT network, contains three computer processors (one MC68040 and two MC68360s). The RBS 540 contains 12 Mbyte flash memory for storing program files and 16 Mbyte dynamic random access memory (DRAM) for buffering and stack handling. A persistent management information base (MIB) stores information on the current configuration in flash memory, which enables operators to reconfigure the base station on-line.

Control of the RBS 540 is based on telecommunications network management (TNM) using the common management information protocol (CMIP). This form of control facilitates many useful functions, such as immediate alarm handling from the ALARM card. File transfer is usually accomplished by means of file transfer, access and management (FTAM) or, when in "boot mode", through the local port.

pACT data intermediate system

The pACT data intermediate system (PDIS) acts as the central switching site in a pACT network, routing data to and from the appropriate base stations. It also maintains routing information for each subscriber device in the network. There are two versions of the PDIS: the home PDIS, and the serving PDIS. Besides switching data packets, the home PDIS maintains a location directory and provides a forwarding service and subscriber authentication. Every subscriber is registered in a home PDIS database. The serving PDIS provides message forwarding, a registration directory and re-address services. Other services or functions are multicast, broadcast, unicast, airlink encryption, header compression (to minimise airlink use), data segmentation, frame sequencing, and network management.

The Ericsson PDIS runs on standard Sun SPARC or Sun Solstice computer platforms. Thanks to high-performance capabilities (serving PDIS: 2,000-4,000 packets/second; home PDIS: 10,000-15,000 packets/second), a single switch can serve the entire pACT network, especially if the network is small or must only meet initial capacity requirements. Flexible multi-switch configurations may also be provided to meet more demanding operator requirements.

Message centre

The message centre (MC) plays a very important role in pACT networks, initiating, provisioning and connecting pACT services to public and private networks, and to PSTN systems. Every message

Box F
Three kinds of messaging service

Multicast
The serving PDIS copies a multicast message to each base station for which members have registered a specific multicast address. This service may be customised to suit specific information services or individual subscriber groups.

Broadcast
The serving PDIS forwards a broadcast message to every base station within a defined area.

Unicast
The PDIS transmits a message regardless of whether a subscriber device responds or not. This service enables a device in a reverse-channel coverage hole to receive its messages.

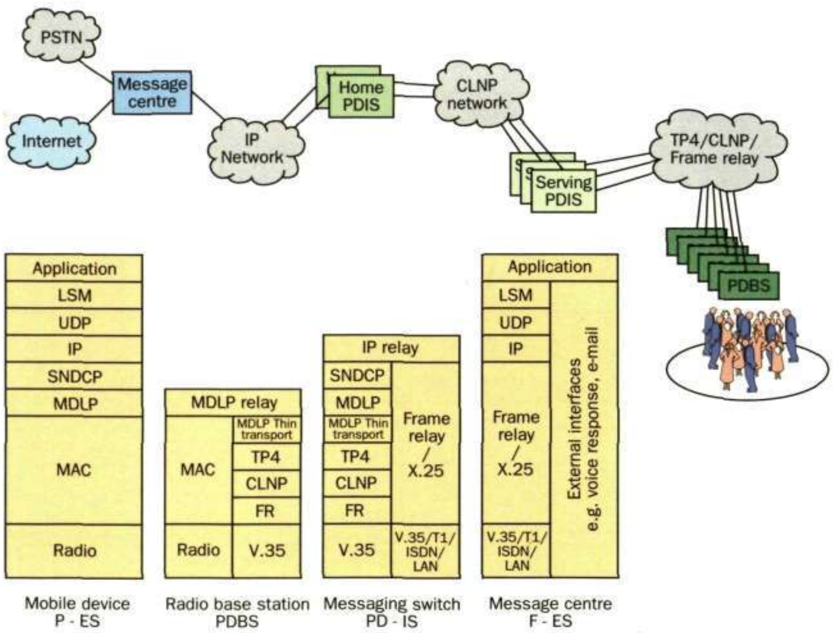


Figure 9
pACT protocol stack end-to-end for two-way message traffic between mobile users and users in external networks, such as the PSTN and Internet. The pACT protocol minimises unnecessary overhead over the air link.

passes through the message centre, whose functionality and applications vary according to network operator requirements. The core of the message centre is the message store, which handles virtually any data type and makes APIs accessible for building different kinds of applications; for example, interactive voice response (IVR), and voice or fax mail. The message store also provides functions for operation and maintenance, system monitoring, and event/alarm handling. Any of the message centre's databases can be queried via the structured query language (SQL).

The message centre also supports virtually any protocol. Typical protocols are the Internet protocol (IP), telocator alphanumeric protocol (TAP), telocator network paging protocol (TNPP), and X.400 (the ITU-T standard for message handling services).

Network management subsystem

The pACT network management subsystem (NMS) is a comprehensive solution that gives operators full control of every component in a pACT network. Because operator requirements for use involve a variety of components and protocols, the Ericsson pACT NMS supports the common management information protocol (CMIP) and the simple network management protocol (SNMP). The NMS is implemented on a modern, distributed, multiuser network management platform called the Sun Solstice Enterprise Manager. The platform includes a graphical user interface (GUI) and several general tools for developing network management applications.

The Ericsson pACT NMS displays logical or physical components. The components are colour-coded – to help summarise the status of sub-components – and displayed in clickable on-screen areas that operators can select to access them. Thanks to the graphical interface, operators do not need to know whether components are managed by the CMIP, SNMP, or by some other protocol.

The NMS contains a database component that permanently stores parameters, configuration data, and historical records of traps and performance data. A pACT network may contain more than one NMS.

Customer activation system

The pACT customer activation system gives pACT networks a complete account management service that enables customer service representatives to manage customer accounts and to dynamically activate pACT-related services for their customers. The customer activation system is easy to integrate into any network. Customer accounts are of two types: individual and business, where individual accounts are for single subscribers and business accounts are for multiple subscribers.

pACT end system (mobile terminals)

Mobile terminals vary from vendor to vendor. They range from simple pagers to sophisticated two-way messaging devices such as personal digital assistants (PDAs) or palmtop computers with wireless modems/embedded chip sets. Full application support is provided for many terminals by the pACT protocol stack with the limited size messaging (LSM) protocol.

Mobile terminals periodically check the designated forward channel for incoming messages. When they are not doing this they are usually in sleep mode – in order to conserve battery life. pACT provides an enhanced flexible sleep mode – which ranges from seconds to days – with simultaneous support of devices that are just "waking up" from sleep mode. Thus, manufacturers and operators may choose between performance (highest power consumption) and latency (maximum battery life) for different classes of device and service.

pACT system protocol stack end-to-end

The pACT protocol stack, Figure 9, features a design that is based on concepts and principles of the CCITT-X.200 and CCITT-X.210 reference models as well as service conventions for open systems interconnection (OSI). The new LSM protocol provides the functionality of a simple e-mail application protocol, such as the simple mail transfer protocol (SMTP), but is highly optimised for low-bandwidth channels. In addition, the LSM protocol *provides a platform for providing true* two-way messaging and datacom services. Examples of services are embedded response messaging for simple pager devices, true two-way e-mail connectivity, and multicast and broadcast messaging. In summary, the LSM protocol provides a migration path for future applications.

To the casual observer, the network layer might resemble the standard IP design for cellular digital packet data (CDPD). However, CDPD is TCP/IP centric, whereas pACT employs a very effective user datagram protocol/Internet protocol (UDP/IP) compression technique that compresses the standard 28-byte header to a single byte for over-the-air transmission. This attribute is important for providing a short alphanumeric messaging service; for example, when the message body is roughly the same size as the header. Consequently, the maximum number of subscribers that can be supported by the system is constrained not by protocol efficiency, but by service traffic. CDPD uses RFC-114, which is a technique for compressing the standard 40-byte TCP/IP header to an average of 3 bytes.

pACT's sub-network convergence layer provides an important new approach to encryption. To ensure that airlink bandwidth is not spent on resynchronising the encryption engines, pACT devices may employ a technique that automatically resynchronises the engines, even when the underlying layers fail to deliver a packet. This technique is important because it provides a mechanism by which multicast and broadcast services may be *encrypted*.

The CDPD link layer is optimised for duplex, whereas pACT uses two-way simplex. The pACT mobile data link layer protocol (MDLP) is an enhanced version of the link access procedure on the D-channel (LAPD), see CCITT Q.920, which allows subscriber devices to adopt strategies for auto-link reset and optimised power saving.

pACT airlink interface

The pACT backbone network is quite similar to a regular CDPD network. The main differences between the two involve functionality – mostly for extending battery life

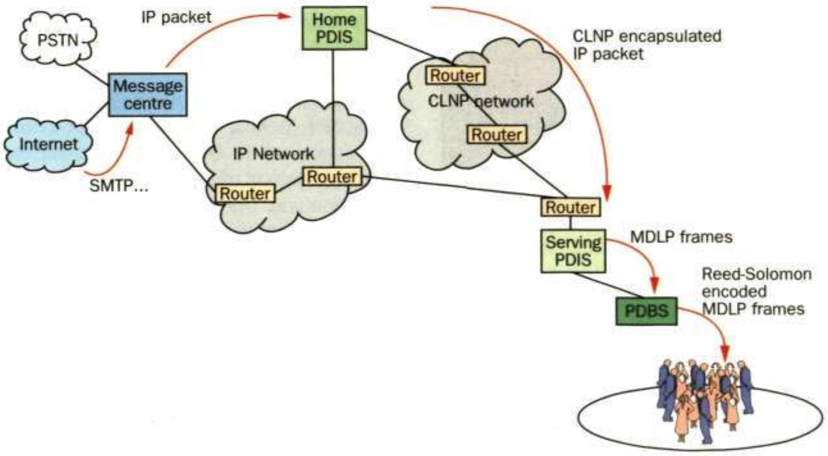
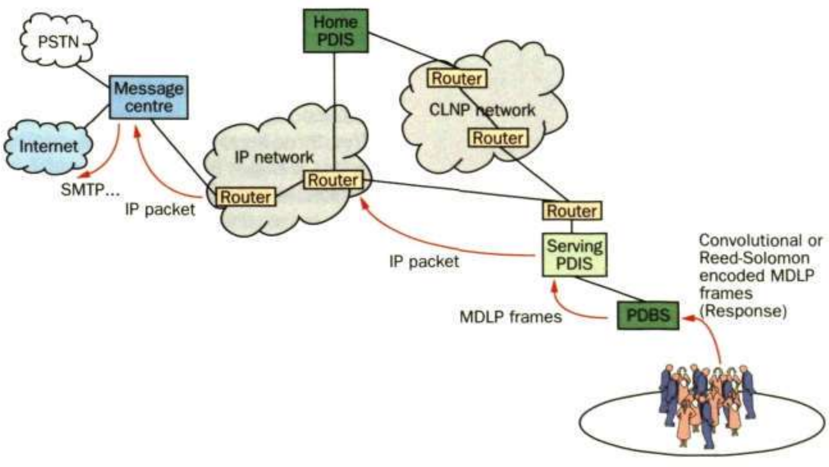


Figure 10
Network traffic flow – to mobile subscribers. A message is always routed via the home PDIS to the serving PDIS, which then delivers the message to the cell (base station) in which the subscriber is located.

Figure 11
Response traffic flow – from mobile subscribers. A message is routed via the serving PDIS to the appropriate message centre.



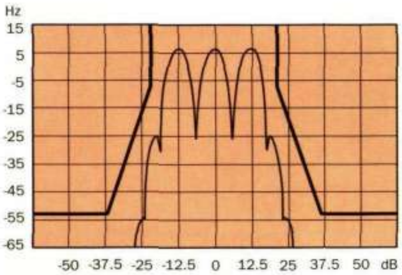


Figure 12
pACT power spectral density of the three most powerful carriers in a 50 kHz channel for GMSK sources: 3 x 50 W, 8 kbit/s, 2-level, BT = 0.5, h = 0.5.

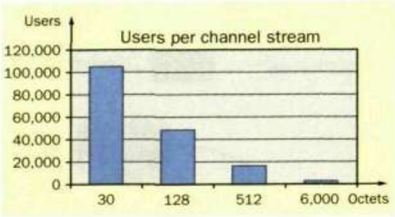
Box G
Example of capacity

Assumptions:

Busy hour call rate	= 25%
Channel	= 50 kHz/50 kHz
Base stations	= 100
Reuse pattern	= 12

Subscribers unevenly distributed in the coverage area. Approximately 1,900,000 subscribers (average message length 30 octets) or approximately 25,000 subscribers (average message length 6000 octets, voice messages approximately 40 seconds).

Figure 13
pACT single radio channel stream capacity. Busy hour call rate is 25%.



in subscriber devices – and features such as group messaging, broadcast, and unicast. Despite the similarity, however, the pACT air interface really does not compare with CDPD. The air interface was developed with very stringent device requirements in mind. While the CDPD air interface was optimised for standard IP-based applications, the pACT protocol shortens and reduces the number of transmissions and contains an efficient sleep mode for conserving batteries. Moreover, although base station power may be 100 times higher than mobile terminal power, which is typically less than 0.7 W, the pACT performance specification enables operators to balance the forward and reverse link coverage.

A 50 kHz channel may accommodate up to three individual radio frequency (RF) carriers. Each base station is assigned a particular 12.5 kHz channel. pACT also supports time slot multiplexing (TSM), which means that in addition to its distinct radio frequency, each base station is assigned a duty cycle and repetition state. The duty cycle and repetition state fulfil requirements for the reuse pattern which, in turn, meets the criteria of the carrier-to-interference ratio. Typical reuse patterns are 9:1, 12:1 and 15:1. An operator with two adjacent 50 kHz channels may drop the guard band (6.25 kHz) to pick up an extra RF carrier – for a total of seven RF carriers.

The airlink interface uses Gaussian minimum shift keying (GMSK) modulation. The both-way trunk (BT) value is 0.5, with a symmetrical transmission speed of 8 kbit/s in both directions. (The BT value is 0.3 for uplink in single 12.5 kHz configurations.) Reed-Solomon (63.47) coding is used on the forward link, and convolutional coding (2/3, 1/2, and 1/3) is used on the reverse link (variable-rate Viterbi coding).

The three-level convolutional coding, which is a major feature of the pACT air interface, is connected to the subscriber device power-control feature. Together, the two play an important role in increasing base station coverage: the coding levels make it possible to trade capacity for reliable transmission capabilities (increased coverage) at the outer fringes of the cell. Power control management helps reduce substantial losses in capacity, especially in cells with heavy traffic loads.

The raw bit rate is always 8 kbit/s in each direction. When a mobile terminal is close to the base station and has a very good radio path, it uses minimum output power and as few coding bits as possible (2/3). As the mobile terminal moves away from the base station, the transmitting power level is gradually increased until it reaches the maximum level allowed. From that point on, nothing more can be done to sustain coverage, except perhaps to add more coding bits in the transmission frames.

Coding bits are added in two steps: "more coding bits" (1/2), followed by "most coding bits" (1/3). Adding coding bits in the transmission frames has the same effect as lowering transmission speed – in either case capacity is traded for coverage. pACT does not trade capacity for coverage unless absolutely necessary. Operators who want to maintain symmetrical throughput in a network or in a particular area do so by building dense networks, where subscriber devices are able to register with another base station before coding bits must be added in order to sustain coverage.

The cellular design of the pACT network enables network operators easily to add capacity where it is needed most. Operators may either split the cells of heavily loaded base stations, or add more base stations.

pACT economics

As with any two-way infrastructure, pACT is more expensive to deploy and operate than regular simulcast-based one-way systems. In order to provide end-users with transparent service offerings, two-way paging devices will most likely require some kind of operator subsidy. That is, based on individual operator requirements and market-specific price levels, operators should anticipate investing 3-7 times as much to deploy a two-way network – which provides the same coverage as a one-way network – as they would to deploy the one-way network. Similarly, operators should expect to pay 5-8 times as much to operate a two-way network. Nonetheless, thanks to the tremendous gains in capacity that are won by the two-way infrastructure, operators are able to serve more subscribers at a lower cost per subscriber.

In two-way systems, a much larger customer base is required to yield a lower average cost per subscriber than the lowest cost per customer in a one-way system, Figure 14. Operators of a two-way infrastructure who exceed this critical point have a distinct competitive advantage. Not only are these operators able to provide one-way paging services for less than their competitors, but they are also able to reap new revenues, by providing new and enhanced services via the same network infrastructure.

There are two basic reasons for investing in a two-way infrastructure:

- Operators need to increase the capacity of geographically large one-way systems without adding frequency spectrum.
- Operators want to provide customers with new or enhanced services.

Small, local operators whose sole objective is to provide regular one-way services in countries or regions where frequency spectrum is not limited are unlikely to benefit from investing in two-way technology.

With the introduction of broadband PCS into the cellular industry, pricing is likely to fall dramatically in coming years. Monthly cellular billing charges will drop to new lows, and the price of basic paging services, which is already quite low, will also continue to fall. The current price differentiation will most likely diminish but not disappear altogether.

pACT Vendor Forum

pACT has received a great deal of attention from the paging industry. Just months after the pACT Vendor Forum was formed in April 1996, more than 30 companies had joined it, including AMD, AT&T Wireless Services, Cantel, Casio, Ericsson, IBM, LanSer, NEC, Panasonic, PCSI, Retix, SEMA Group, Sharp, and US Robotics.

In its charter, the pACT Vendor Forum states that its role is to enable worldwide support for, and to promote the adoption of, the pACT two-way narrowband PCS standard. Forum members are dedicated to the rapid development of a variety of pACT products and services that are designed for open systems, and that ensure multi-vendor compatibility and interoperability. Membership is open to any vendor, or prospective vendor, of products or services that comply with the pACT standard.

Conclusion

Despite the introduction of packet data services and short message services in various cellular technologies, multi-service platforms cannot be optimised for every target group. Dedicated networks for packet services still play an important role, and narrowband networks and related services will continue to emphasise their particular strengths. In head-on competition, cellular providers will find it difficult to compete with the low cost, small device size, long battery life, and low monthly service fees that paging services offer.

pACT is primarily intended for the North American narrowband PCS market. However, pACT will work in any frequency band that is set aside for two-way paging and messaging services. Besides pACT, other protocols for the narrowband PCS frequencies include the family of FLEX protocols from Motorola.

pACT increases capacity by reusing frequencies. By combining this feature with high airlink speeds, operators can increase capacity in a large zone 100-200 times – even more in networks that consist of several zones.

The pACT system is built from several flexible modules that can be combined and configured in different ways to meet specific operator demands. Because the pACT network is based on the Internet protocol, operators and application providers can take full advantage of existing tools, applications, and APIs. pACT also provides secure service, including encryption and authentication. Very few systems can provide millions of customers with the same cheap, easy-to-use service as pACT, especially when applied to a total channel bandwidth that is measured in kHz.

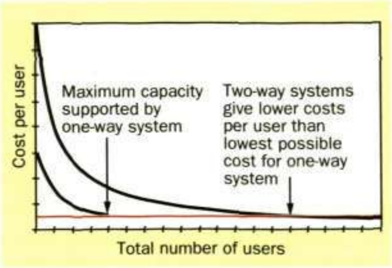


Figure 14
Comparison of network operator costs per subscriber for one-way and two-way systems.

References

1 pACT Specification, Release 1.1, October 1995, AT&T Wireless Services.
2 pACT 97 specification. Ericsson, PCSI and Datum Telegraphic
3 RBS 540 Product Description. Ericsson Radio Messaging.
4 Industry Report, Wireless Communication – Volume 1. John Adams, Wessels, Arnold & Henderson

Open telecom platform

Seved Torstendahl

The open telecom platform (OTP) is a development system platform for building telecommunications applications, and a control system platform for running them. The platform, whose aim is to reduce time to market, enables designers to build – from standard, commercially available computer platforms – a highly-productive development environment that is based on the programming language Erlang.

The OTP also permits application designers who program in C, C++, Java and other languages to take full advantage of sourced components. Moreover, the OTP allows designers to consider costs when matching computer platforms with requirements for processing power and component availability.

The author outlines the OTP system architecture, its tools and building blocks, while describing the strengths of the open telecom platform as either a development environment or as a target system.

The open telecom platform (OTP) described in this article is primarily intended for new applications that combine a need for reliable, high-performance telecom characteristics with a need for using externally sourced hardware and software components. This includes ATM-products for access networks and data communications, such as Internet protocol-related traffic.

Multipurpose platform

The open telecom platform is a development system platform for building, and a control system platform for running, telecommunications applications. However, the OTP is not a monolithic platform, but is made up of sets of tools and building blocks, which include:

- Erlang – a programming language complete with compiler, debugger, and other development tools (as a rule, the tools and building blocks that are required for Erlang are already implemented in the language, Box D);
- SASL – systems architecture support libraries (SASL) contain functions for building fault-tolerant distributed applications;
- Mnesia – a real-time fault-tolerant distributed database management system (DBMS);
- the Erlang run-time system;
- sourced programs;
- standard, commercially available operating systems;
- computer hardware.

Different packages of OTP-related components have been defined to suit different user needs. Applications development packages, for example, contain all the tools and building blocks that designers need to develop applications for specific computer systems. Package types that are meant to be integrated into an embedded application system (target system) make up a basic part of that system and contain the building blocks whose functions are needed at run-time. Examples include a database, or a simple network management protocol (SNMP) agent.

OTP development environment

Many of the application programs for the OTP are written in Erlang. Application programmers will find that several tools in the OTP development environment support the tasks of developing and testing applications. For example, the OTP development environment contains:

- tools for translating into stub code the interface specifications given in the form of C header files, simple network management protocol-management information base (SNMP-MIB) definitions, and abstract syntax notation number one (ASN.1) definitions;
- an Erlang-to-Emacs mode that facilitates editing Erlang programs;
- the Erlang compiler and debugger for testing modules;
- a test coverage tool, and a hot-spot finder.

Some programs are written in other languages, including C, C++ and Java. In such cases, the application programmer uses the tools provided by the supplier of the target system, or the tools that are available in the development environment.

Target systems

An OTP that consists of a control system with one or more processors may be configured as a target system. If the system comprises more than one processor, then the processors must be able to communicate with one another through a logical network. Some processors, equipped with secondary storage and various I/O units, supervise and control the general system by means of functions in the systems architecture support libraries (SASL)—the OTP's central component.

A communication mechanism joins each processor to the distributed system. If desired – depending on requirements for reliability – the processors and the

Box A
Abbreviations

API	Application program interface
ASN.1	Abstract syntax notation number one
BOS	Basic operating system
BSD	Berkeley software distribution
DBMS	Database management system
HTTP	Hypertext transfer protocol
I/O	Input/output
IP	Internet protocol
JAM	Joe's abstract machine
MIB	Management information base
MMU	Memory management unit
OTP	Open telecom platform
RCS	Revision control system
RPC	Remote procedure call
SASL	Systems architecture support libraries
SNMP	Simple network management protocol
TCP	Transfer control protocol
UDP	User datagram protocol

communication mechanism may be duplicated. The OTP handles all the basic telecommunications requirements for the control system (real-time, fault-tolerance, live software upgrades, distribution), thereby allowing application designers to fully concentrate on the unique aspects of their own work. Moreover, the OTP can easily be ported to several different commercial operating systems.

System architecture

All application systems that are built on the open telecom platform (OTP) adhere to a basic architectural structure, Figure 1.

Bottom layer
Commercial computer systems make up the bottom layer. The system hardware in this layer may also be designed in-house, if the manufacturing cost and the volume of delivery can be justified. This is merely an architectural view; in real systems, the bottom layer contains many computers which may be of different types.

Middle layer
Support for telecommunications requirements is provided by a robust real-time distributed database management system (DBMS); basic support for handling software and reporting events; an extensible SNMP agent; a Web server; and a library of routines for interworking between applications written in C and Erlang.

Top layer
All applications have access to Mnesia and SASL. The SNMP agent and the Web server may also invoke functions that are provided by the applications in this layer.

Interfaces
Three interfaces are provided: an interface to OTP software; an interface to the operating system; and an interface between applications written in different languages. In terms of logic, the third interface applies to the application level, but is implemented in the OTP and in the operating system.

Application programs
The OTP includes a set of application program interfaces (API), as well as rules and guidelines for writing application programs. It also includes teaching materials, documentation standards, and sample programs that show how application programs are designed.

Most programs are written in Erlang. However, application programs for time-critical parts may be written in C.

Sourced programs
The OTP defines how sourced programs, which include protocol stacks and management applications, may be incorporated into a system and made to interwork with programs that were developed by application programmers.

SASL
The systems architecture support libraries (SASL) contain basic software that supports system start/restart, live system software updates, and process management. The basic operating system (BOS)—a predecesor to SASL – was used for several years in the Mobility Server and other systems.

Mnesia
A real-time fault-tolerant distributed DBMS that supports fast transactions for the telecommunications application, and a query language, called Mnemosyne, for handling complex queries.

SNMP support
SNMP provides run-time support through an extensible agent, and development support by means of agent/sub-agent design principles and an MIB compiler.

Web server
The Web server permits data that refers to Erlang functions to be collected via Web pages.

Erlang run-time system
The basic system that supports the execution of Erlang programs. The Erlang run-time system includes the Erlang abstract machine, which executes intermediate code, the kernel, and standard libraries.

Computer platforms
The operating system and computer hardware consist of standard commercially available systems. Testing environments, for example, use conventional workstations.

OTP component packages are defined to suit various purposes. Devel-

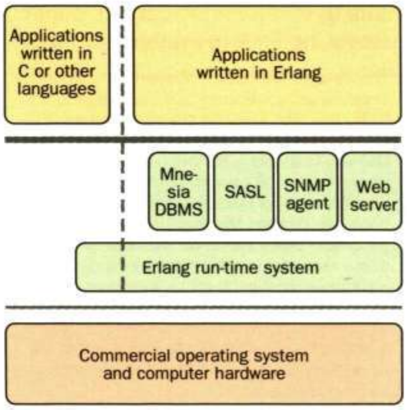

Applications written in C or other languages
Applications written in Erlang
Mnesia DBMS
SASL
SNMP agent
Web server
Erlang run-time system
Commercial operating system and computer hardware

Figure 1
The OTP system architecture.

Box B
Reducing time to market

Shortening time to market entails much more than simply reducing the time it takes to write program code. The time spent on the other phases of a project must also be reduced.

One way to do this is to have program designers begin their work at the early stages, or at high levels, of design – for example, starting with the functional specification. Moreover, design phases can be broken down into short incremental steps, where each step is fully integrated into and tested at the level where it will be implemented. The initial step, for example, might consist of starting all blocks. Next, critical parts can be added. Finally, management functions and error handling can be added. This way, problems can be detected and corrected early on.

Box C
Getting the most out of Erlang's easy-to-port functionality:

In any part of an application written in C, do not include functions that relate to a specific operating system. Otherwise, gather these parts into a cluster, or else mark them clearly.

Recommendation:

Use the C compiler that is used for the Erlang runtime system.

Table 1
Operating systems and computer platforms on which the OTP may be run

Microprocessor architecture	Operating system
Intel Pentium	Linux, Solaris 2, Windows 95, Windows NT
Motorola 680X0	Vx Works
Sun SPARC	Solaris 2

opment packages include software for developing Erlang programs, as well as libraries and applications. The user organisation purchases the actual computer platform, such as a Sun workstation with the Solaris operating system, a PC with a Windows operating system, or some other computer system for which an OTP development environment is available.

Packages for embedded systems include libraries and applications, but do not contain the compiler and debugger. These packages also include references to agreements that have been reached with vendors of computer systems. Under the terms of these agreements, various licensing fees are paid automatically when specific brands of hardware are procured, or when specific versions of software are copied.

Strengths of the open telecom platform

The open telecom platform (OTP) runs on many computer platforms, and caters to applications that have been designed in Erlang as well as in standard programming languages. Moreover, the OTP is designed to support typical telecommunications requirements for robustness, smooth software upgrades, distribution, and real-time functionality.

Time to market

A major objective of the OTP is to provide a highly productive environment for designing telecommunications applications. This objective is supported by:

- the Erlang programming language – a high-level functional language that supports distribution, fault detection, and recovery;
- building blocks
  - robust, real-time distributed DBMS;
  - support libraries for creating applications that can report errors, restart themselves when errors occur, and update themselves with new versions of software when ordered to do so;
  - an extensible SNMP agent and a Web server that is closely integrated with the database.

Multiple computer platforms; easy to port; up-to-date computer technology

To accomodate different microprocessor architectures and operating systems, several versions of the OTP exist, see Table 1.

The effort required to support many different operating systems is manageable, mainly because most support for non-Erlang languages is obtained directly from the vendors of computer systems. Only a minor part of the Erlang run-time support is dependent on the computer platform. Most Erlang run-time support, and all parts of applications that are written in Erlang, are independent of system hardware and related operating systems. Moreover, "Joe's abstract machine" (JAM) code is directly executable on any platform provided that the Erlang code does not explicitly make use of a specific feature of the operating system. Applications written entirely in Erlang are easily ported, and in most cases their load modules (compiled code) are directly executible on other computer systems.

Hardware and software from external suppliers

Sourced hardware and software play a decisive role in reducing time to market. Very often, software vendors can offer special applications at prices far below the cost of developing an in-house solution. Likewise, because computer systems change rapidly, rather than investing considerable resources to develop a computer board, Ericsson can offer a wide range of attractive solutions that are based on existing, commercially available systems. To this end, the Ericsson external technology provisioning process is used to enforce competitive terms for external hardware and software, and to forge ties with reliable vendors.

Software

Software from external vendors falls into three categories:

- Applications – which include the upper layers of protocol stacks.
- Operating system (OS) kernels.
- Software that is directly coupled to hardware or to the OS kernel – for example, device drivers and protocol stacks.

To handle support and maintenance, sourced software should be imported as object code. Because malfunctions can cause a system to crash, operating system software and software that is directly coupled to hardware or to the OS kernel must be thoroughly tested and proven reliable. Some protection may be provided for applications by running them as

Box D
The Erlang programming language

Background

When introducing new technology, real-time systems often lag behind other systems. Indeed, many real-time systems are written in the programming language C. Some languages, however, have been developed specifically for programming concurrent real-time systems. Some of the best known examples are Ada, Modula2 and PLEX.

Today, declarative programming languages such as Prolog or ML are used for a wide range of industrial applications. These languages drastically reduce the total volume of source code in applications, as well as the efforts that are required to design and maintain them. However, declarative languages were not primarily designed to be used in concurrent real-time systems.

Erlang, which is a small but extremely powerful language for programming concurrent real-time, fault-tolerant distributed control applications, combines important attributes of declarative languages with constructs for supporting concurrency, distribution, and error-detection. It is an expressive, high-level functional programming language without pointers – a feature that greatly simplifies design and testing.

Built on experience

Erlang has been in use at Ericsson for more than five years. To date, many hundreds of thousands of lines of Erlang source code have been written, demonstrating the language's suitability for use in large projects. A key to achieving very high productivity, Erlang has been used for some years in several Ericsson products, including Mobility Server, ISO Ethernet, and NETSIM.

Functions, modules and processes

Erlang programs are made up of functions that have been grouped into modules and packaged as software products. Functions spawn processes – the executing elements of an Erlang system – that are very lightweight and that enable fine-grained concurrency.

Processes communicate by sending and receiving messages. Communication with the external non-Erlang world is conducted through ports (which behave like processes). Processes may also be linked to each other in order to detect software errors.

A built-in distribution mechanism enables designers to create a system whose processes may run on different computers. Erlang allows fault detection and recovery in a distributed system, and the OTP software implements customisable schemes for recovery after faults.

Reference:
"Concurrent Programming in Erlang" by Armstrong, Virding, Wikström and Williams.

Very high-level functional/declarative language

Symbolic data representation

Support of massive lightweight concurrency (parallelism)

Support of designing distributed, non-homogeneous systems

Permits tailored-to-fit fault recovery schemes in distributed systems

No pointers, no memory leaks

No fixed sizes or limits

Easy to interface other software and hardware

Permits software to be updated while running

Modular concept for structuring applications

Easy to create reusable libraries

Figure D
Characteristics of Erlang.

separate processes, preferably with memory management unit (MMU) protection. Application software is best managed through its source code, where software management requirements can be adapted.

Hardware

The OTP is designed to use different microprocessor architectures and computer boards. For example, standard workstations or PCs are used for development, and standard computer boards are normally used for inclusion in embedded (target) systems.

Computer technology is evolving very rapidly. Every nine months, PC vendors design a new generation of boards with faster processors, more and faster memory, and new integrated I/O circuits. Competing at this pace with in-house solutions would be very costly for Ericsson. However, by relying on external processor hardware, Ericsson can transfer their need for staying abreast of developments to computer vendors. Other advantages of using standard computer boards in the OTP are that every new board, from low- to high-end, can be made available, and that development costs can be shared with the rest of the market. Thus, Ericsson draw on the expertise of external vendors, while remaining focused on their own core areas of business.

Erlang and other languages

Applications that use the OTP can be implemented in Erlang or in any other programming language. The choice of language is governed by the characteristics one hopes to derive.

Programming in Erlang shortens development time and provides support for designing robust distributed real-time applications. Support is provided in the form of

- libraries of ready-to-use components;
- guidelines for using the components;
- guidelines for designing the applications that provide desired characteristics.

By means of careful system design, and by applying the "90/10" rule (fine-tune the 10% of Erlang code that occupies 90%

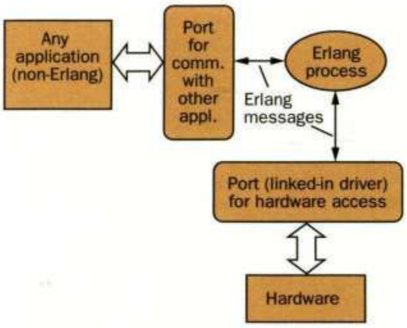

Any application (non-Erlang)
Port for comm. with other appl.
Erlang process
Erlang messages
Port (linked-in driver) for hardware access
Hardware

Figure 2
The port mechanism in Erlang.

Box E
Reliability

A key requirement of the OTP is to provide a robust environment for implementing fault-tolerant telecommunications applications. By robust environment is meant a system that provides specified services despite the presence of hardware faults or software errors. The minimum requirements of a robust environment are error detection, damage assessment, and error correction. Each of these aspects is necessary for meaningful error-recovery at run time.

Other requirements of the control system are as follows:

- The "outage" time (that is, the time during which services are unavailable) must be short.
- While error recovery operations are under way, only a minimum of services may be disrupted. This implies that some kind of "hot" or "warm" standby solution must be used.

Hardware considerations

Hardware is made fault-tolerant by duplicating resources; for instance, by means of redundant CPUs and dual communication paths. The provision of a redundant CPU structure is more important than synchronous duplication of CPUs, since identical CPUs executing in parallel do nothing to guard against software errors – that is, given the same error, both machines would fail at the same time.

Software considerations

Reliable software systems are built by carefully structuring the system so that an error in one part of the system may not affect its other parts.

of processing time), it is possible to obtain good run-time performance from Erlang.

Development in other programming languages (usually C) is sometimes necessary for reducing execution time. Highly optimised C code is very efficient, and may be required in time-critical parts of applications. The OTP offers different ways of integrating applications written in C into applications written in Erlang or in mixed programming languages.

One reason for using C is that a market already exists for applications or components written in that language, including applications for implementing communication protocols. Obviously, being able to use existing products is a desirable feature. Thanks to a component in the OTP and to an adaptation unit written in C, these parts can be integrated into the software management principles used by the OTP.

There are several ways of creating interfaces between applications written in Erlang and other programming languages:

- through the port mechanism;
- using the socket library;
- through the C/C++ interface generator;
- by enabling C programs to imitate an Erlang node.

The port mechanism

According to the port mechanism, which is the standard mechanism for interfacing applications, Figure 2, a port on the Erlang side of the interface is perceived and behaves as an Erlang process. Here, all communication consists of standard Erlang messages. Software units written in another language perceive these messages as communication lines to another program. Ports may also be used for accessing hardware or low-level features directly.

The socket library

Delivered as part of the standard Erlang distribution, the socket library is a low-level mechanism that enables non-Erlang processes to communicate with Erlang processes by means of the transmission control protocol/Internet protocol (TCP/IP). The Erlang program is responsible for understanding and complying with the protocol that is used on top of the TCP.

The C/C++ interface generator

Thanks to the C/C++ interface generator, functions written in C and Erlang can easily communicate with one another. The interface generator includes conversion routines on both sides of the interface for encoding and decoding transferred data. This enables applications written in Erlang to access and manipulate C data structures. To use the conversion routines, some programming is required on the C side of the interface. All underlying communication is conducted using the port mechanism, or the socket library.

C programs that imitate an Erlang node

Erlang data can be encoded to byte sequences in a format known as the Erlang external format. A library of useful functions has been provided to enable programs written in C to decode byte sequences of this kind. The library has been extended to facilitate communication with a distributed Erlang node using the Erlang distribution protocol. Because they behave like an Erlang node, C programs that decode byte sequences of Erlang code are called C nodes. C nodes, which have very limited functionality, are not visible to Erlang, and are therefore said to be hidden.

OTP support for high reliability

An Erlang system that is run as a task by a host operating system is called an Erlang node. In the OTP, processes in different nodes communicate as easily with one another as processes within the same node. Erlang processes may also be linked together. Should one of the processes die, due to an error, then a signal is sent to each process that is linked to it. Since error signals can be trapped by supervisory processes, highly reliable layered systems can be designed.

The Erlang programming technique of dividing computations into "supervisory" and "worker" roles can be employed to build a robust system architecture. SASL provide patterns that facilitate design according to these rules.

The system structure consists of a critical "safe kernel" that must always be correct, and an application area where the requirement for correctness is somewhat less stringent. The safe kernel is provided by SASL.

Erlang has a real-time garbage collector, very few operations with side effects, and no pointers. Thus, when Erlang is used, a large class of problems commonly associated with programming real-time systems is eliminated.

Updating sourced software

Ordinarily, operating system software and device drivers can only be replaced by rebooting the processor. Thus, where continuous operation is a requirement, system software may only be updated in systems that make use of multiple processors. In a multiple-processor system, the system software is replaced on one processor while the other processor continues to execute as usual.

By contrast, even in single-processor systems, individual applications can usually be replaced while the system is running; however, the applications themselves must typically be terminated, updated, and then restarted. Again, if this method is not acceptable, then a multiple-processor system may be necessary.

Tools in the OTP

In this context, tools are programs that are used to develop software. A brief description of the tools included in the OTP follows below.

Application monitoring.

The appmon program has two main parts: the node window, which shows an overview of the applications on all known nodes; and the application window, which shows the process tree of each application.

Because both windows run on one node (called the server node), graphics need only be installed on one node. The monitoring programs then use small information agents on each monitored node.

Development process

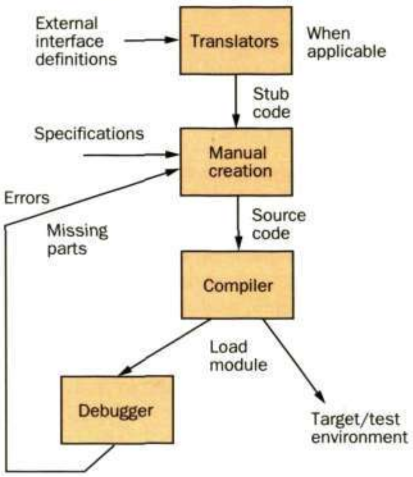


Figure 3
Applications may be developed from interface definitions (C header files, ASN.1 specifications, CORBA/IDL interface definitions). After the source code has been written, compiled and tested, it is reiterated until satisfactory results are reached.

ASN.1 compiler

The ASN.1 standard, which is used for describing data representations in communication, defines unambiguously the interpretation of transferred data units. An ASN.1 compiler is provided for translating interface definitions in ASN.1 notation into stub code together with calls to translation routines for packing and unpacking transferred data.

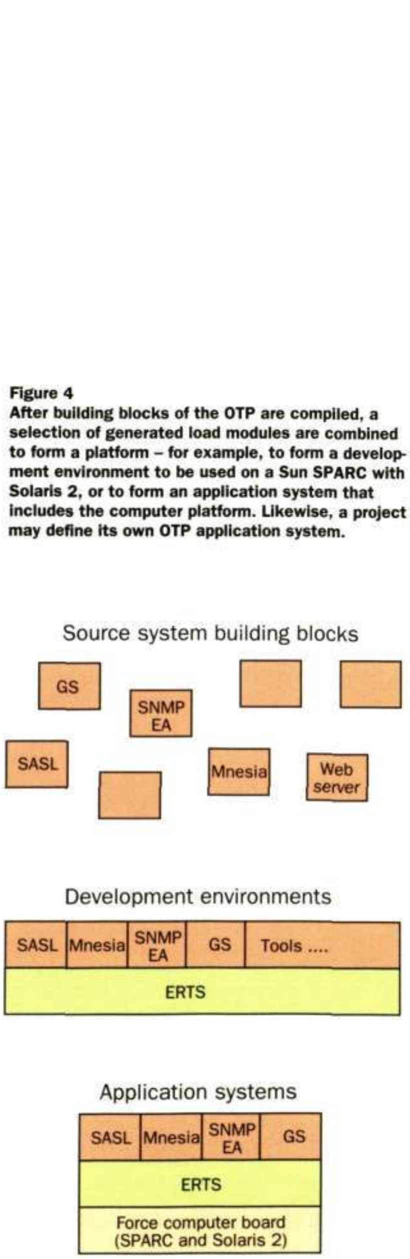
C interface generator

This tool facilitates communication between programs written in C and Erlang. An interface specification received in the form of a header file is translated into stub code. The stub code may then be used for calls between Erlang and C to convert data representations.

Compiler

The Erlang compiler translates Erlang source code (text files) into object code that is independent of the CPU in use. This means that the resulting code may be loaded onto any computer platform

Figure 4
After building blocks of the OTP are compiled, a selection of generated load modules are combined to form a platform – for example, to form a development environment to be used on a Sun SPARC with Solaris 2, or to form an application system that includes the computer platform. Likewise, a project may define its own OTP application system.



that supports Erlang without the code having to be changed or recompiled.

Coverage tester
The coverage tester is used to ensure that all code in a module has been executed during a test.

Cross-reference tool
The exref tool is an incremental cross-reference server that builds a cross-reference graph for every module that is loaded into it. A great deal of information, including graphs and module dependencies, may be derived from the cross-reference graph. The call graph is represented as a directed graph in text.

Debugger
The Erlang debugger provides mechanisms for visualising what happens when code is executed in specified modules, or when processes crash. Because it allows breakpoints to be set, and because it performs sophisticated tracing, the debugger interferes with the behaviour of the system being debugged.

The debugger is mainly used to locate errors in code (bugs). However, designers may also use it to learn about, or to better understand, applications that were written by other designers.

Because Erlang is a distributed concurrent language, conventional debugging techniques do not apply. Thus, the Erlang debugger provides mechanisms for attaching to, and interacting with, several processes simultaneously, including local processes and processes located in other Erlang systems in a network of Erlang systems. Every process that runs code in debugged modules is monitored, and continuous information on the status of any given process may be displayed.

Erlang mode for Emacs
Emacs is a widely-used text editor whose behaviour can be customised. A definition file is distributed with the OTP that allows the Emacs editor to recognise Erlang syntax, and that helps designers to write well-formatted, syntactically correct code. The Emacs editor is not included, but may be obtained free of charge from many sources on the Internet.

Graphics interface for Erlang
The graphics system is a general graphics interface for Erlang. The interface is easy to learn, and may be ported to many different platforms. If future applications in Erlang are written to the same graphics API, then it will be possible to run each supported platform without having to change a single line of the application code.

Profiler
The profiler is used to determine which parts of code, "hot spots", occupy the most CPU time. This tool is useful for optimising programs.

SASL
The systems architecture support libraries (SASL), which are used as building blocks by applications that are running, provide usage rules for designing robust applications that may be started, stopped, and restarted, and that can report errors or events.

SNMP MIB compiler and instrumentation
The OTP provides support for using the simple network management protocol (SNMP) for operating and maintaining applications on the platform. SNMP support consists of

- run-time support – in the form of an extensible agent;
- development support – in the form of a management information base (MIB) compiler and a programming model for implementing MIBs.

The extensible agent uses two mechanisms – dynamic MIB loading and subagent handling – that provide an environment where MIB modules can be loaded/unloaded in an efficient plug-and-play fashion. The sub-agent concept also supports distributed applications; that is, different MIBs can be implemented at different nodes. If necessary, the transaction mechanism for SNMP set requests may be customised.

The MIB compiler may be used to generate the prototype instrumentation of MIBs automatically. The results may then be incrementally refined and tuned. This feature enables designers to start developing manager applications at the early stages of a project.

Support is also provided for using the Mnesia DBMS together with the SNMP tool kit. This means that Mnesia tables can be read and manipulated with SNMP, and that an SNMP table can be implemented as a Mnesia table. The

SNMP tool kit provides an API for implementing application-specific MIB modules, including a mechanism for sending traps.

Trace tool
The debugger-tracer has a graphical front-end. When used in a running system, it causes little disturbance to the system. Trace tool supports distributed debugging, which means it may be used to debug Erlang processes running on the same system as well as processes that are running on several remote systems. The remote systems may each run on different operating systems and CPUs. Also included is a simple line-oriented interface to the trace functions.

xerl
A complete integrated Erlang environment, xerl is an X Window-based interface to the standard Erlang shell. The interface may be used to compile, edit, debug and run Erlang. Moreover, because xerl is sensitive to Erlang syntax, it checks and formats code as it is written.

yecc parser generator
A parser generator for Erlang that is similar to yacc. From a grammar definition (input), the generator produces Erlang code for a parser.

Other tools
At present, tools for managing different versions of software are not delivered as a part of the OTP. Nonetheless, a management tool, such as the revision control system (RCS) or ClearCase, is highly recommended.

Building blocks in the OTP

When the OTP is used to create an embedded system, some of its components are retained by the delivered system. OTP run-time applications are also available as building blocks in the development environment (Figures 3 and 4).

Erlang virtual machine
The Erlang virtual machine runs on top of a host operating system. The Erlang run-time system frequently runs as a single process in the host operating system. The Erlang virtual machine provides the following support for Erlang programs:

- Consistent operating system interface on all platforms.
- Memory allocation and real-time garbage collection, which effectively eliminates memory leaks.
- Lightweight concurrency and support of thousands of simultaneous tasks.
- Transparent cooperation between all computers in the system.
- Location and encapsulation of run-time errors.
- Supervision of run-time code as it loads or is replaced, and while it is linked.

The structure of the Erlang virtual machine allows it to be easily incorporated into new operating systems. The Erlang virtual machine is the only building block that was not written in Erlang.

Kernel
The kernel is always the first application to be started. In a minimal OTP system, the kernel is one of only two applications. The kernel application contains the following services:

- application_controller
- auth
- code
- error_logger
- file
- global_name_server
- net_kernel
- rpc
- user

Box F
Components of supporting design of robust applications

Component	Description
Standard libraries	The standard libraries contain Erlang modules, such as lists, "ordsets", and "dict". An important property of these libraries is that the code they contain is free of side effects. Thus, calls to functions in these modules work the same in every environment.
Behaviour	In this context, behaviour is a formalised pattern of design that can be reused to build new applications. Behaviour is implemented in the form of call-back modules: a call-back module must export a specific set of functions that are called by the system to evoke a particular pattern of behaviour. Examples of behaviour are: - gen_server – a behaviour that implements a generic client-server machine; - gen_event – a behaviour that implements a generic event-handling machine; - supervisor – a behaviour that implements a specific style of supervisor-worker programming which allows supervision trees to be constructed dynamically.
Applications	Applications are named collections of objects that are used for encapsulating system components. The basic properties of an application are that it can be named, loaded and unloaded, as well as started and stopped. The aim of an application is to provide a mechanism that enables users to divorce themselves from the internal details of the application, and to think instead about the relationships between applications.

The kernel application also includes standard *error_logger* event handlers.

Mnemosyne – the query language

The Mnesia DBMS is organised as a basic layer that takes first-order predicate logic with Erlang data types as values. Recursion, negation, and a simple constraint system are also supported. The evaluation is set-oriented. In practice, users have a choice of several query shells in which to type their queries and receive their answers.

Views of data are displayed according to the rules defined in the modules. Each module corresponds to a file and is compiled and stored in the database itself. The modules and rules are declared in a schema. In some instances, even when the table is used in the module, the schema may be changed by simply recompiling the module.

Mnesia real-time DBMS

In most applications some data must (a) be stored safely and (b) remain easy to access. Requirements vary a great deal; for example, the amount of data may range from a lot to very little. On the one hand, the loss of data from a system crash may be acceptable, whereas in other instances, data must survive practically any kind of system failure. Similarly, some applications require near-instantaneous access time – perhaps only fractions of a millisecond – whereas in other instances a longer access time may be acceptable. Moreover, some applications require a valid prediction of the access time – real-time access.

In short, despite a broad spectrum of requirements, it must be easy to update data without introducing inconsistencies, especially when several corresponding changes are made. Also, for operations and maintenance purposes, operators must sometimes be able to access data by means of complicated and non-standard queries.

In response to these requirements, the Mnesia DBMS was developed with the following basic properties:

- Data is accessed in two basic ways:
  - through an API for programs;
  - using a query language, called Mnemosyne, for humans
- Read and write access is protected by means of transactions. This method gives each user (program or person) the impression that he is alone in the system. It also bars inconsistencies from being introduced into the data (status control) when several corresponding changes are made.
- Data may be distributed transparently over several Erlang nodes. Important data is copied to several nodes; less important data may reside at only one location. The distribution of data over nodes cuts access time drastically.
- Data is declared to reside in RAM, on secondary storage, or both.
- If a hardware or software node crashes, the data that it contained may be reconstructed from redundant copies of the data stored at other nodes, or from logged updates. Assuming that the data at a particular node is replicated at other live nodes, then the addition or removal of that node will not be observed by users in the system.
- Data is organised in tables. This provides *a sound theoretical basis, and a* broad range of well-known, proven methods for modelling the data.
- The tables, their organisation, location, storage type, and other aspects are declared in a schema. Views of data are supported from the API and from the query language.
- An optimising query-language compiler and evaluator is available for use by operators and in the API.

OS monitoring

The OS monitoring application defines the following services:

- disksup – which checks the available disk space and sends an event if a stated threshold has been passed;
- memsup – which checks the available primary memory and sends an event if a stated threshold has been passed.

Read, write and search from the API

From the basic API, a program – with or without a transaction – can read and write table entries or merely search a table. Multiple, related accesses that involve writing without a transaction are strongly discouraged, since they can severely damage data. Nonetheless, this option has been provided to give immediate access to experienced programmers who need to perform concurrency control themselves. Each time the data is read without a transaction a snapshot is taken.

A query language has been provided to facilitate complicated queries, such as when data from many tables needs to be combined. The query language, which is mainly based on first-order logic, is accessed through list comprehension. A list comprehension is a functional language construction that is well-suited to the Erlang programming language.

SASL
The systems architecture support libraries (SASL) are designed for building embedded real-time Erlang applications, and include a set of standard design templates that can be used to solve common programming problems on application start, restart and supervision.

SNMP extensible agent
The OTP facilitates use of the SNMP for operating and maintaining applications on the OTP. SNMP support consists of run-time support in the form of an extensible agent, and development support in the form of an MIB compiler and a programming model for implementing MIBs.

The extensible agent uses two mechanisms that provide an environment where MIB modules can be loaded/unloaded in an efficient plug-and-play fashion. These mechanisms include dynamic MIB loading and sub-agent handling.

Sockets
Sockets are the Berkeley software distribution (BSD) UNIX interface to communication protocols. Various protocols may be accessed through sockets. The socket module provides an interface to the BSD UNIX sockets. The udp module supports user datagram protocol/Internet protocol (UDP/IP) sockets.

Standard modules and libraries
The Erlang programming environment contains several standard reusable software modules. The functional program paradigm on which Erlang is based greatly facilitates reusing software.

Many standard modules are specially adapted to the needs of concurrent distributed systems. The remote procedure call (RPC) module, for example, allows designers to program a remote procedure call in one line of source code. Examples of standard modules are lists, "ordsets", "gen_server" and "gen_event".

Web server
The Web server is a hypertext transfer protocol (HTTP) daemon implemented in Erlang. Access to the server looks up a requested Web page and sends it to the browser. The page may fetch data from a database table or from a function call. Integration is very efficient.

Conclusion

The open telecom platform (OTP) enables designers to build and run telecommunications applications on a broad range of standard, commercially available hardware and software platforms. The OTP also allows designers who program in C, C++, Java and other languages to integrate sourced components – protocol stacks, APIs, I/O units and drivers – into their applications.

The OTP comes with an exhaustive collection of tools and building blocks, such as the programming language Erlang, systems architecture support libraries (SASL), a real-time database management system (DBMS), an extensible SNMP agent, and a Web server.

As a design environment, the main strengths of the OTP are: time to market; compatibility with many different computer platforms and sourced components; up-to-date hardware and software technology; and reliability. What is more, the OTP permits designers to consider costs when they match computer platforms with requirements for processing power and component availability.

As a target enviroment, the OTP meets all basic telecommunications requirements. It has a real-time distributed control system that is fault-tolerant, and that can handle software upgrades while it is running. In addition, the OTP can easily be ported to several different commercial operating systems.

References

1 Däcker, B.; Erlang – A new programming language. Ericsson Review 70 (1993:2), pp. 51-57.

Box G
Sockets

A socket is a duplex communications channel between two UNIX processes, either over a network to a remote machine, or locally between processes running on the same machine. A socket connects two parties, the initiator and the connector. The initiator is the UNIX process that first opens the socket. It issues a series of system calls to set up the socket and then waits for another process to create a connection to the socket. When the connector makes its connections, it also issues a series of system calls to set up the socket. Each process then continues running. The communications channel is bound to a file descriptor that each process uses for reading and writing.

EriOpto – A new family of optical transceivers

Hans Ivarsson, Gunnar Forsberg and Michael Lynn

Recent decades have witnessed telecommunications switching and transmission equipment evolving to become increasingly complex in terms of functionality, while steadily decreasing in physical size. The emergence of optical fibre as an inexpensive means of transferring large amounts of data has opened vast potential markets for service providers who are able to employ fibre in their distribution networks. However, stringent requirements for reliability and robustness are placed on the optical transceiver equipment in such applications.

Desiring to buy an optical transceiver, but finding no existing technologies that met their specifications, Ericsson set out to design and industrialise a suitably robust and reliable product that could be mass-produced at low cost.

The authors describe the EriOpto – a flexible optical transceiver system over a single fibre – as well as the innovative and successful approach to its design, which called on many parts of the organisation to work together across functionally and geographically diverse boundaries.

Ericsson have designed a family of compact optical transceivers (TRX), called EriOpto. The primary feature of the transceiver module is that it may be placed flush with the front panel of a printed board assembly (PBA) and still fulfil requirements for electromagnetic compatibility (EMC).

The EriOpto is a duplex optical transceiver that uses optical multiplexing technology to transfer data on a single fibre. While the data transfer rates may vary within the EriOpto family, references in this context are to the EriOpto4, whose data transfer rate is 184 Mbits/s. Data transfer is achieved on 820 and 1300 nm wavelengths.

The EriOpto module may be adapted to several transmission speeds and used for a wide variety of applications, Figure 1. Examples include:

- connections between an asynchronous transfer mode (ATM) switch and a local area network (LAN) server;
- internal transmission links for switching equipment;
- LANs that require bidirectional high-

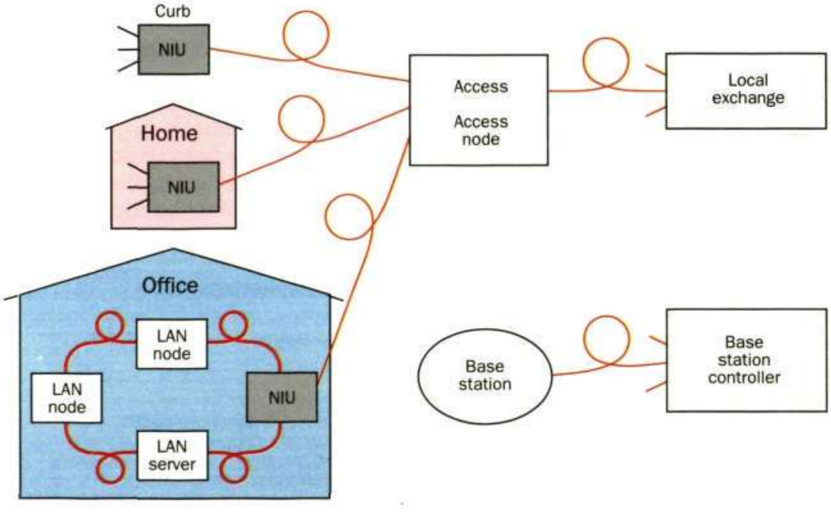


Figure 1
Some applications of the EriOpto link.

NIU Network interworking unit

EriOpto-link

speed data transfer over a single fibre;
- optical fibre to the curb;
- optical fibre to the home;
- optical fibre to the office;
- transmission links to radio base stations or to other remote equipment.

Project overview

Wanting to buy an optical transceiver, but finding that no existing technologies met their specifications, Ericsson set out to design and build a robust and reliable transceiver of their own. Consequently, a project team was assigned to develop and industrialise a duplex optical transceiver that could be mass-produced at low cost. This objective was broken down into several specific requirements that went into the project specification.

- The project must be completed within 12 months.
- Once completed, working prototypes of the optical module must be delivered to internal customers.
- On project completion, final products must be in production – these products must be robust, reliable, easy to use, and competitively priced; they must consume very little power, have a small footprint (to facilitate mounting them at the front of a printed board assembly), and fulfil requirements for electromagnetic compatibility and electrostatic discharge (EMC/ESD).
- Equipment must be developed that can test more than 10,000 transceivers a year.
- The final production test must take no longer than four minutes to complete.

While drawing on Ericsson's knowledge of quality improvement as well as their long experience of designing and manufacturing telecommunications equipment, the project team applied several innovative techniques in designing and manufacturing the optical transceiver. The two chief characteristics of the project were that it represented a total implementation of concurrent engineering, and that it produced a fully-functional demonstrator transceiver by the end of the preliminary study. Incidentally, inventions within the project resulted in eight patent applications, and several pending applications.

Working methods

A mixture of competencies

The main obstacle facing the project team was their tight development schedule, which necessitated the decision to design the optical transceiver while simultaneously developing a production line that could manufacture more than 10,000 optical modules per year.

Prior to the project, much of the technological know-how related to optical transceivers was still at an early research level. Therefore, the research as well as

Box A Abbreviations

ASIC	Application-specific integrated circuit
ATM	Asynchronous transfer mode
BER	Bit error ratio
CKE	Clock enable
CMOS	Complementary metal-oxide semiconductor
COB	Chip on board
DC	Direct current
DLD	Dark line defect
EMC	Electromagnetic compatibility
ESD	Electrostatic discharge
ETSI	European Telecommunications Standards Institute
FIT	Failure unit
HP-VEE	Hewlett-Packard Visual Engineering Environment
IEC	International Electrotechnical Commission
I/O	Input/output
LAN	Local area network
LE	Loop enable
LED	Light-emitting diode
MIL-STD	Military standard
MTTF	Mean time to failure
NIU	Network interworking unit
PBA	Printed board assembly
PECL	Pseudo emitter coupler logic
PIN	Photodiode (Pindiode)
PLL	Phase-locked loop
PROM	Programmable read-only memory
RH	Relative humidity
RX	Receiver
SMD	Surface-mounted device
ST	Connector, ST is a trademark of AT&T
THB	Temperature-humidity bias
TRX	Transceiver
TX	Transmitter
VCXO	Voltage-controlled crystal oscillator
WDM	Wavelength division multiplexer

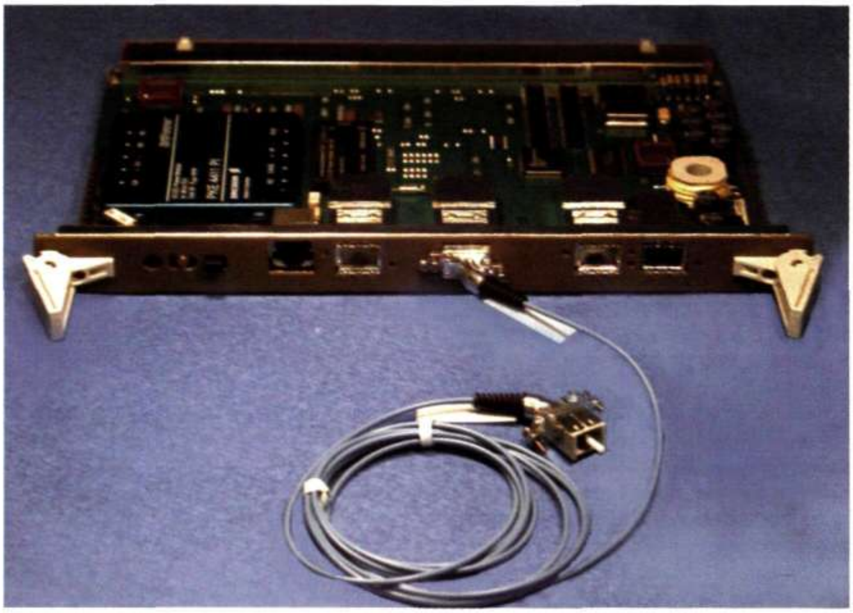


Figure 2
EriOpto4 mounted at the front of a printed board assembly.

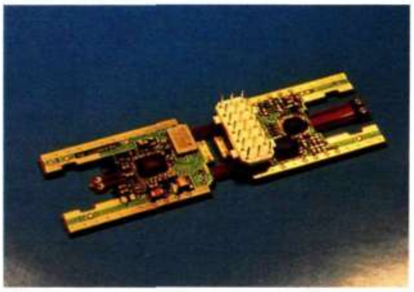


Figure 3
Picture showing COB version mounted on the flex-rigid board.

technical and product development had to be carried out in parallel. This was made possible thanks to the communications applications that were available via the Ericsson corporate networks. A broad range of competencies – from research centres, application labs, technology development centres, product development centres, patent departments, component departments, procurement, and supply and installation departments – were called upon to collaborate from many different parts of Europe (England, Finland, Norway, Sweden and Switzerland). While this approach was not without risk, the project managers reported that it proved to be a real success.

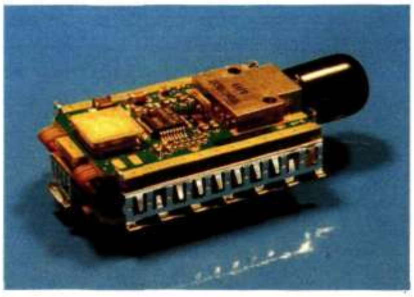


Figure 4
Picture showing SMD version of the transceiver, PBA folded around the grounding shield.

Throughout the project, managers actively encouraged team members to share their knowledge, skills and experience with one another. This interaction helped keep the team motivated, and propelled the project forward.

Early on, the project reached a crucial milestone by deciding how the module's design would facilitate manufacturing and system requirements for optimal production, installation and maintenance.

Patent handling

A very proactive approach was adopted for handling the inventions that resulted from the project's various activities. A patent engineer worked with the team to ensure that no opportunities in this area went overlooked. As a result, eight new patents were filed – a remarkable feat considering the relatively small size of the project.

Results

Using technical specifications as a roadmap, the EriOpto project succeeded in meeting each of its goals – despite some very stringent requirements. For example, the design project successfully reduced the time needed to perform the final test of a completed EriOpto unit from approximately eight hours to the stipulated goal of just four minutes. Similarly, the cost of manufacturing the optical transceiver modules was reduced very significantly.

System description of the optical transceiver link

Background

In recent years, manufacturers have worked energetically to reduce the physical dimensions of robust optical modules that fulfil requirements for EMC/ESD. Their aim has been to produce optical modules that can be mounted at the front of a printed board assembly (PBA). Notwithstanding these efforts, prior to the EriOpto project, externally-available optical modules failed to meet Ericsson's requirements for

- robustness – physical as well as in terms of EMC/ESD;
- low power dissipation;

- small footprint;
- ease of handling.

Therefore, Ericsson set out to create their own optical transceiver. To date, two versions of the EriOpto have been designed: the EriOpto40, and an improved version, the EriOpto4N.

The EriOpto optical transceiver is a generic opto-mechanical building block that is well adapted to the Ericsson front connector concept. As such, it may be used in a variety of communications applications. The design of the EriOpto modules and cables reduces cabling volumes, extends transmission distances, and fulfils EMC/ESD-related requirements. Moreover, as it is easy to use, the EriOpto saves designers valuable time (time to market) when they design system products.

Mechanical design

Currently, a leading development tendency within telecommunications technology is towards the miniaturisation of components and circuits for optical transceivers. However, the aim of the EriOpto project went much further than this: besides designing an optical module whose small physical dimensions enable the module to be mounted at the front of a printed board assembly, Figure 2, the project produced a robust unit that fulfils requirements for EMC/ESD and can be manufactured at low cost.

The module's superior quality was achieved by applying several new ideas. For example, the solution to EMC/ESD-related requirements involved integrating a shielding plate into the board design. Another idea made use of a new kind of printed board, called the flex-rigid board, Figure 3, whose middle section consists of a flexible plastic foil. In this design, components are mounted on only one side of the board, which simplifies the production process further.

One side of the board is assembled adjacent to the shielding plate. The other side is folded over by bending the plastic foil to create a separate assembled board unit, Figure 4. In this way, the shielding plate separates the receiver and transmitter from one another.

A further function of the shielding plate is to shield the complete module together with its covers, thereby creating three screening cages. The first is situated between the shielding plate and the lower part of the cover; the second is situated between the shielding plate and the upper cover; and the third screening cage is situated between the lower and upper covers.

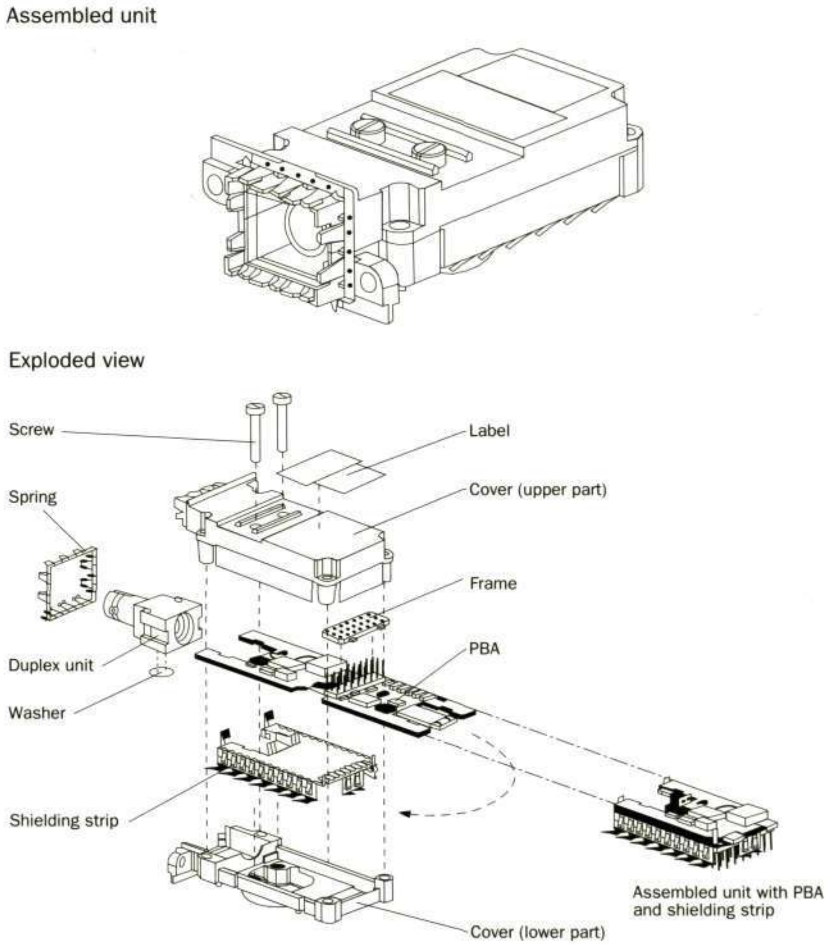


Figure 5
Exploded view of EriOpto4 module.

The design also channels heat away from the components on the board to the upper and lower covers, which dissipate the heat into the air. A spring, Figure 5, grounds the complete module to the front panel.

A pull-relief clamp is designed to protect the connector against any force acting on the optical cable, Figure 6. The optical cable with connector fit into ETSI-miscellaneous (ETS 300 119-3) racks and cabinets.

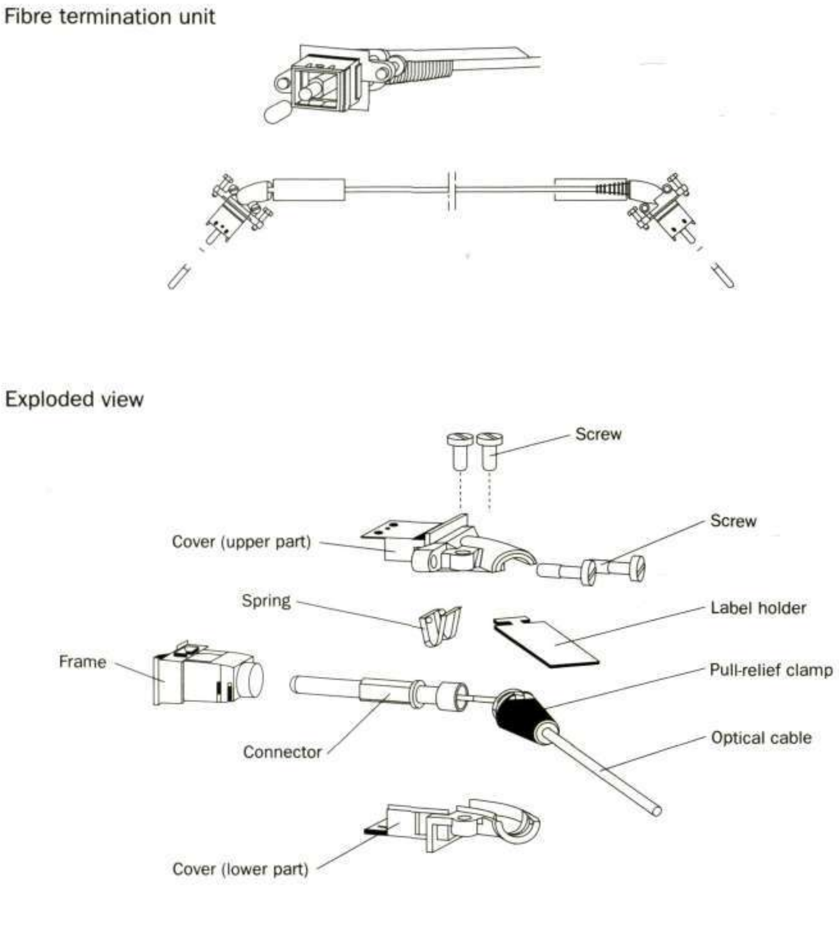


Figure 6
Exploded view of optical cable with connector.

The transceiver module

The EriOpto4N module is built around two full-custom application-specific integrated circuits: EriOpus-TX and EriOpus-RX. These ASICs, which use high-performance BiCMOS technology, make up the transmitter and receiver parts of a physical link interface for a 184 Mbits/s optical transceiver.

The transceiver module, Figure 7, houses a light-emitting diode (LED), a photodiode (PIN) a transmitter (TX), a receiver (RX), a wavelength division multiplexer (WDM), and a physical receptacle for the connector for the optical cable assembly. The connector meets requirements that are compatible with an ST connector.

Optical data is sent using the LED and is received using the photodiode. The WDM separates the different wavelengths that are used for transferring data in each direction.

EriOpus-TX

EriOpus-TX is an electrical-to-optical transmitter circuit for use in relatively short-length (less than 200 metres) optical-fibre, high-speed (184 Mbit/s) data communication links. A stronger light source may be added to adapt the transmitter for use over longer distances. The transmitter circuit also contains logic for the testing of full-speed dynamic electrical input signals as well as the static loopback of clock and data signals, Figure 8.

The key design features of the transmitter are:

- extremely low power dissipation (typically 230 mW);

Box B
General technical data on the 2x184 Mbit/s EriOpto

Transmission	Duplex
Bit rate	2x184 Mbit/s
Bit error ratio (BER)	$<10^{-16}$
High-speed electronic input/output (I/O)	Compatible with differential pseudo emitter coupler logic (PECL)
Low-speed electronic I/O	Compatible with complementary metal-oxide semiconductors (CMOS)
Data and clock output	Data is latched by clock
Data and clock input	Clock latches data
Clock enable input (CKE)	0; data is not latched by clock
ALARM output	High at loss of synchronisation

Power consumption	EriOpto4O: typically 1.1 W EriOpto4N: typically 0.45 W
Optical technical data	
Wavelength	1300/820 nm (direction-dependent)
Fibre	Multimode 62.5/125 or 200/230
Optical output power	Typically -17.5 dBm (62.5/125)
Optical sensitivity	Typically -26 dBm (BER = 10^{-16}, 62.5/125)
Guaranteed link length	200 m for 62.5/125; 30 m for 200/230
Connectors	Only at link ends

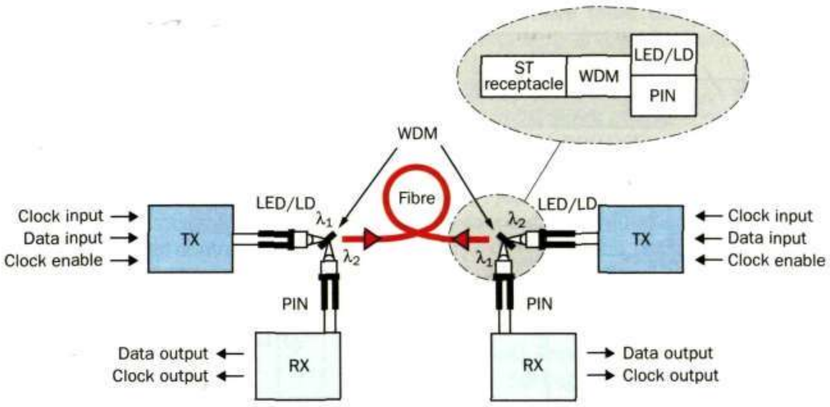


Figure 7
EriOpto4-link block diagram.

- operation from a 3.3 V nominal power supply;
- a high-current drive for interfacing LED types (short wavelength: 820 nm, long wavelength: 1300 nm);
- digital control of the high- and low-current drive levels via 16-bit programmable read-only memory (PROM) code;
- very fast switching times to permit a 184 Mbit/s data transfer rate;
- very low pulse-width distortion to permit full recovery of data, while preserving the direct current (DC) balance of the communications link;
- a low percentage extinction ratio – to ensure an adequate signal-to-noise ratio as well as a low bit error ratio;
- an active signal – to enable operators to turn the LED drive off.

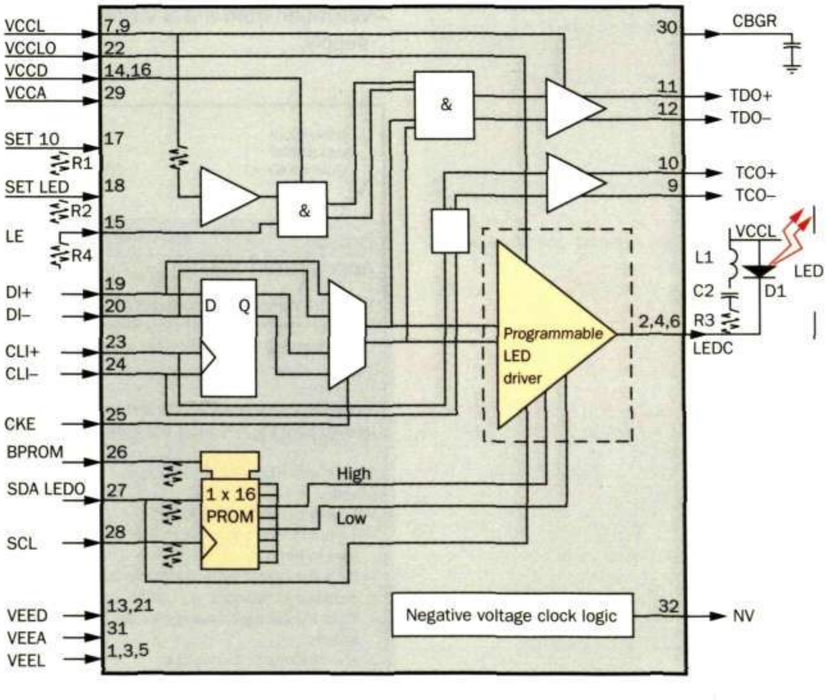
The final design of the LED driver is based on a series switch. In the logic low state, the current in the LED is much reduced, which reduces overall power consumption.

The high- and low-current LED drives are selected by digital control through 16-PROM bits. When the transmitter and LED are first put into operation, the electrical drive levels are adjusted until they produce the desired optical power levels for high and low data transfer rates. Once the test system has properly adjusted the settings, it then burns fuse links in the PROM to make them permanent. Finally, to complete the process, and to prevent further fuse links from being burned, the test system causes a protection bit to blow.

EriOpus-RX

EriOpus-RX is an optical-to-electrical receiver circuit for use in relatively short length (less than 200 metres) optical-fibre, high-speed (184 Mbit/s) data communication links. The receiver circuit is based on a single chip and includes clock extraction. The input signal is provided via a photodiode that is coupled to an

Figure 8
Block diagram of the optical transmitter circuit (TX).



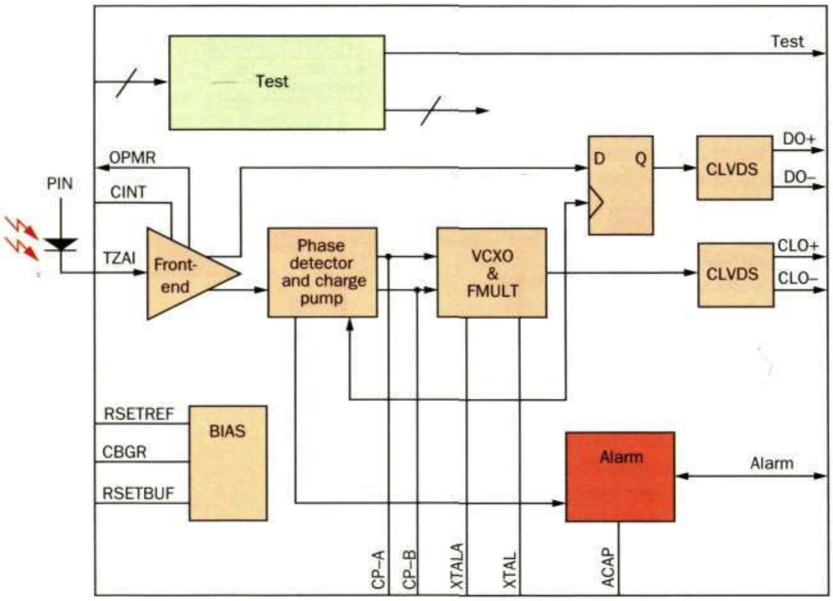


Figure 9
Block diagram of optical receiver circuit (RX).

optical fibre. The receiver circuit contains logic for testing full-speed dynamic electrical input/output signals as well as the static loop-back of clock and data signals, Figure 9.

The key design features of the receiver are:

- extremely low power dissipation (typically 220 mW);
- operation from a 3.3 V nominal power supply;
- an interface to various photodiode types (short wavelength: 820 nm, long wavelength: 1300 nm);
- very fast switching times to permit a 184 Mbit/s data transfer rate;
- very low pulse-width distortion at the output;
- very low jitter.

Box C
Accelerated ageing

The most common method used for demonstrating long-term reliability of electric and fibre-optic components is called *thermally accelerated ageing*.

The relationship between life and temperature is derived from the Arrhenius equation,

$t_1/t_2 = \exp[(E_a/k)(1/T_1 - 1/T_2)]$
where

- t_1 and t_2 represent the life;
- T_1 is the test temperature (absolute temperature in Kelvin);
- T_2 is the operating temperature (absolute temperature in Kelvin);
- E_a is the activation energy for the failure mechanism;
- k is Boltzmann's constant.

The main building blocks of the receiver are the front-end transimpedance amplifier and the clock recovery function. A DC bias current is provided by the EriOpus-TX (the transmitter). The current is input to a photodiode, which modulates the current and then sends it to the transimpedance amplifier circuit. This circuit amplifies the signal and then sends it to a low-pass filter, which limits the signal bandwidth.

Clock recovery is accomplished by a phase-locked loop (PLL), whose main components are a phase detector and a voltage-controlled crystal oscillator (VCXO). The phase detector, which is an XOR phase detector with a charge pump for driving the external loop filter, contains alarm circuitry for monitoring synchronisation.

The VCXO is a fundamental-mode oscillator with a 46 MHz crystal for the resonant circuit followed by a frequency multiplier. The frequency multiplier multiplies the frequency in two stages, converting the fundamental frequency from 46 to 184 MHz. In each stage, the frequency passes through a self-tuned (frequency-locked-loop concept) bandpass filter; is doubled (*non-linear frequency doubling*); and then passes through a limiting amplifier.

The ASICs for the transceiver were developed according to specification by Ericsson Components. The transmitter part was designed in England and the receiver part was designed in the Stockholm area. Overall control of this sub-project rested in Stockholm.

Quality and reliability

Qualification program

Electric and fibre-optic components manufactured by Ericsson for use in telecommunications equipment must work properly for the life of the system in which they are used (20 years). Accordingly, these components must undergo a battery of qualification tests.

Fibre-optic components

Because their long-term reliability is well understood and documented, long-wavelength (1300 nm) LEDs and photodiodes are considered mature devices. By contrast, the reliability of short-wavelength (800-900 nm) LEDs remained an open issue until quite recently. The chief argument against these components was that defects in their crystal structure – known as dark line defects (DLD) – caused them to emit less output power. Today, manufacturers have more or less overcome this problem, and evaluations of manufacturer test data

(<200 nf/h at $T_{junction}$ = +60°C within 20 years of operation)

indicate that 800-900 nm LEDs are suitable for telecommunications applications. Manufacturers are responsible for performing the qualification tests of duplex units, which contain both a photodiode and an LED.

Qualification tests

The most important tests conducted on plastic packaged microcircuits are: temperature cycling, humidity test, and accelerated life test (Box C). Nevertheless, as it is difficult to design a relevant procedure for testing the operational life of a printed board assembly, a decision was made to test the operational life of an EriOpto40 module instead. Due to their nature, however, fibre-optic components cannot withstand the same high testing temperatures as microcircuits. And an accelerated life test at +85°C for the customary duration of 2000 hours represents only approximately 2.5 years of operation at +40°C. Thus, to simulate 20 years of operation, the testing period must be extended (Box D).

The manufacturer chosen to assemble the EriOpto40 modules runs a series of reliability tests on the chip-on-board (COB) mounting technique that is used

Box D
Failure rate

Failure rate is defined as nano failures per hour (nf/h), or 10^9 failures per hour. This is also expressed in failure units (FIT), where one FIT equals one failure per 10^9 hours. The failure rate, which usually varies with time, is represented by the bath-tub curve, Fig. D.

Infant mortality may be reduced with proper burn-in and screening procedures.

Life tests are designed to verify that components will not wear out prematurely. Only random failures may be accepted (constant failure rate).

Mean time to failure (MTTF) is often used to describe the lifetime:

MTTF = γ

where γ is the failure rate.

MTTF is used when the failure rate is constant in time. Otherwise, the reliability parameters median life (time to 50% wear-out failure) and dispersion (measure of the spread in lifetime) are commonly used.

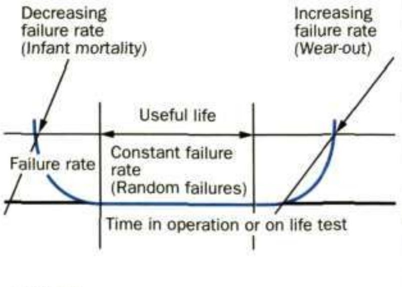

Decreasing failure rate (Infant mortality)
Increasing failure rate (Wear-out)
Useful life
Failure rate
Constant failure rate (Random failures)
Time in operation or on life test

Figure D
The failure rate, which usually varies with time, is represented by the bath-tub curve.

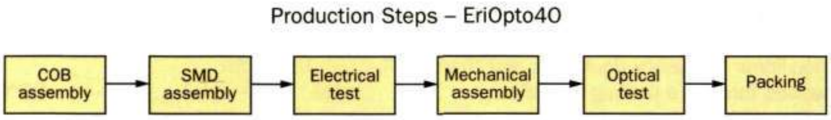


Figure 10
Production flow for the COB design.

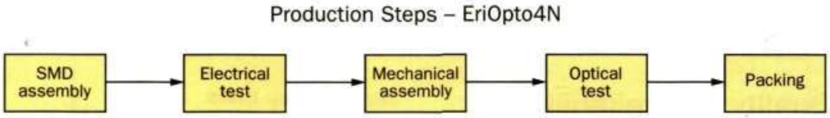


Figure 11
Production flow for the SMD design.

for this type of module. Thus far – in terms of meeting telecommunications requirements – the results have been favourable (Tables 1 and 2).

Production

Each version of the module has a unique construction. The EriOpto4O features a chip-on-board (COB) and a surface-mounted device design, whereas the EriOpto4N employs solely an SMD design. In either version all components are assembled on the primary side only.

In brief, the process for assembling surface-mounted devices is as follows:

- The solder paste, including flux on the conductive pattern of the printed circuit board, is screened.
- Components are then placed on the board with a pick-and-place robot. The main discrete components in the EriOpto4 are the smallest of their kind on the market (size 0402, which is approximately 1 mm x 0.5 mm).
- The complete circuit passes through a reflow oven at a temperature of 210°C, which is adequate for soldering the components to the board.
- The module is cleaned and all residue is removed.
- Connectors are soldered manually; a plastic frame is inserted; and a shield and an optical transceiver are mounted.

Before delivery, each fully-assembled unit is thoroughly tested to check transfer functions and optics as well as to measure distortion and the attenuation band.

Table 1
Qualification tests for printed board assemblies

Test	**Condition/method**	**SS/c[1])**
Temperature-humidity bias (THB)	+85°C/85% relative humidity (RH), 1000 hours.	20/0
Temperature cycling	-55°C to +125°C, 500 cycles	37/0

[1]) SS = sample size, c = acceptance criteria

Table 2
Qualification tests for EriOpto4O modules

Test	**Condition/method**	**SS/c[1])**
Vibration	IEC 68-2-6, 10-500 Hz, 10g	5/0
Shock	IEC 68-2-27, 100 g, 6 ms	5/0
Robustness of termination	IEC-749-II.1	2/0
Resistance to soldering heat	IEC 68-2-20	5/0
Solderability	IEC 749-II.2.1 or IEC 68-2-54	5/0
Marking permanence	IEC 749-IV.2 Method 2	2/0
ESD (human body model)	MIL-STD-883C, Method 3015 minimum requirement is 500 V	2/0
Temperature-humidity bias (THB)	+85°C/85% RH, 1000 hours.	15/0
Temperature cycling	-40° to +85°C, 500 cycles	15/0
Operating life test	T_{case} = +85°C, >2000 hours.	24/0

[1]) SS = sample size, c = acceptance criteria

These tests are an integral part of the production flow. The final production process, which includes testing and assembly, takes less than five minutes to complete, Figures 10 and 11.

Testing

A criterion for meeting the goal of low-cost mass production was that an automated testing process be implemented to shorten testing times to no more than four minutes per module.

The EriOpto module is a self-contained optical subsystem, complete with clock recovery and power supply filtering. Thus, in the design phase, designers of a large board – such as a fibre-to-the-home application that incorporates the EriOpto module – need not be experts in fibre-optics or in high-frequency design. Similarly, in the testing phase, the board test need not involve a complex optical test system.

The implications here are great; that is, compared with designers who must design an integrated optical subsystem from scratch, those designers who make use of the EriOpto module are able to cut board-related time to production and speed up service roll-out dates by as much as six months.

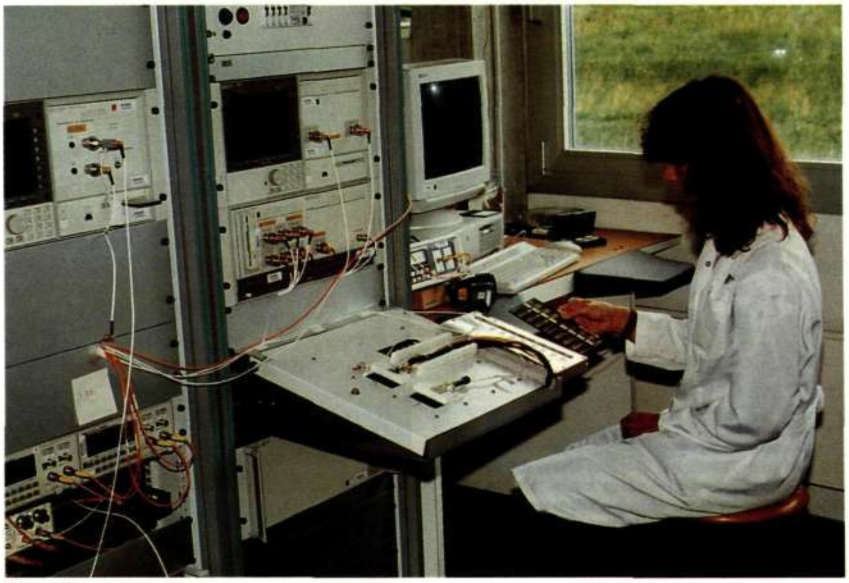


Figure 12
Test equipment (at Valtronic, Les Charbonnières, Switzerland) consisting of two joined instrument cabinets with a protruding operator's test table and a PC with a bar code reader.

Production test

To facilitate mass production it was necessary to develop automatic test equipment, Figure 12. A distinguishing feature of the test equipment is that no special skills are required to operate it. Instead, test operators need only scan the bar code label of each test object with a bar code reader. The software in the test equipment then takes over, automatically supplying the test program with necessary parameters, values, and so on (Box E). The test program features the Hewlett-Packard Visual Engineering Environment (HP VEE), which is a user-friendly man-machine interface, with support for localising faults before the boards are assembled into complete modules, Figure 13.

In the past, the many optical connectors that must be manually connected and disconnected posed a major inconvenience for test operators. The test equipment, however, uses just one optical connector for connecting to each test object. Moreover, the connector is placed on a moveable sledge, which further increases ease of use, Figure 14.

Given that (a) the test interval never exceeds four minutes, (b) the final assembly takes 30 seconds or less to

Figure 13
Test operator with bar code reader.

Box E Elements of the production test

Current consumption
Impedance of input termination
Electrical output amplitude
Current due to differential output data signals
Current due to differential output clock signals
Photodiode bias
Bit errors in looped condition
Alarm output
Clock enable input (CKE)
Loop enable input (LE)
LED disable input
Automatic adjustment of LED current to nominal current
Optical output power at nominal LED current
Automatic adjustment of LED current
Optical output power at chosen LED current
Current consumption
Total power consumption
Rise and fall time of optical output signal
Pulse-width distortion of optical output signal
Extinction ratio of optical output signal
Bit errors at optical output signal
Optical power level for BER = 10^{-9} (Receiver)
Increase in BER due to crosstalk from LED driver
BER measurement for 30 s, at higher optical level
PLL capture range
Jitter transfer function and jitter tolerance
Alarm output in the absence of optical input
Photodiode currents under varied conditions

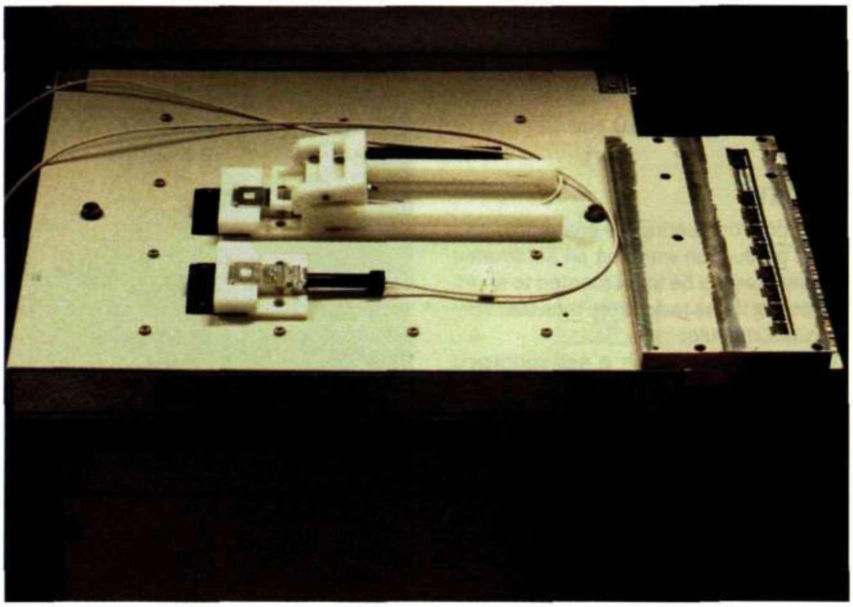


Figure 14
The movable sledge can be seen in the middle of the test table. When the operator slides the sledge to the left, the connector easily mates into the test object (seen left of the sledge). When a new test object is to be tested, the sledge is positioned to the right and the test object is electrically connected to a conventional lever-operated electrical test connector.

complete, and (c) the test station is operated round the clock, then a single test station may be used to test more than 100,000 modules per year. To accommodate larger production rates, additional test stations could be used in parallel; for example, five test stations would be needed to test half a million units per year.

Conclusion

In many ways, the EriOpto project represents a significant step forward for the Ericsson research and development organisation. The project succeeded not only in designing a completely new optical transceiver system in accordance with strict technical specifications, but also – as a result of very demanding time constraints – pioneered several new working methods and approaches.

The flexible design of the EriOpto family of products promises much potential for the future evolution of communications technologies and applications.

The results of the project were thus not only very beneficial from a technical standpoint, but also extremely valuable in terms of the organisational experience that was gained for future operations.

Box F
Future applications of the EriOpto

Fibre types

Multimode

- 200/230 µm step index, mechanically robust fibre
- 200/230µm gradient index, mechanically robust fibre
- 62,5/125 µm

Single mode

9-125 µm

Data transer rates up to 1.454 Gbit/s

Narrowband

0-1 Mbit/s

Narrowband is built into some curcuits.

Broadband

155 Mbit/s, 184 Mbit/s, 622 Mbit/s, 737 Mbit/s

Several different types of circuit are available for 155-737 Mbit/s, some of which contain narrowband functionality.

OPERA will provide circuits that consume less power for 155-737 Mbit/s.

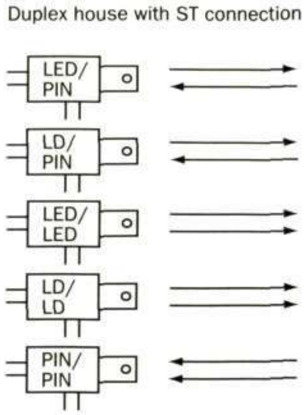


Figure F1
Duplex house with ST connection.

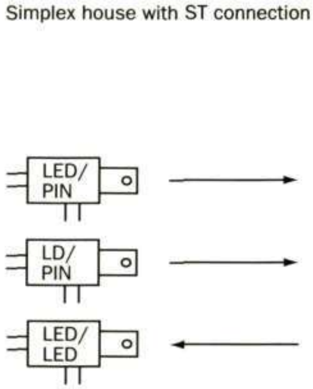


Figure F2
Simplex house with ST connection.

Increased competitiveness through Ericsson services for telecom operators

Anders Lindström, Olle Lövenheim, My Spangenberg and Fredrik T. Strandh

Growing competition and high subscriber penetration in the deregulated telecommunications market are forcing operators to become more competitive. To succeed under such conditions, operators need to differentiate their offerings to subscribers, and to make their operations more cost-effective. Today, operator success depends on being able to increase revenue while reducing costs and enhancing productivity.

The authors describe Ericsson services for telecom operators. These customer services comprise four main areas of support to different phases of operation – professional services; implementation and integration; maintenance and support; and customer training – including specific tailored-to-fit solutions to specific operator needs.

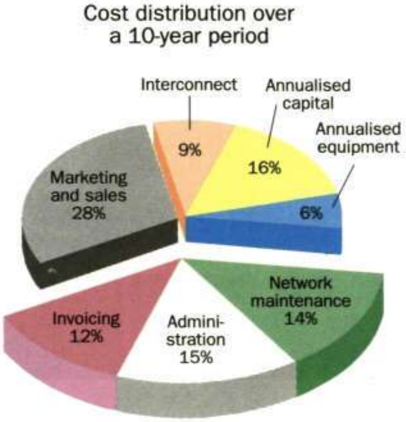


Figure 1
In 1992, the US Federal Communications Commission issued a report on the cost structure of new personal communications services operators. A further development of the estimates is shown in this diagram.
The capital cost share is probably valid for most western world operators, while interconnection costs vary considerably – due to the differing nature of competition in the wireline markets.
The remaining parts will also vary from case to case. However, the diagram indicates the proportions of operator costs. Network maintenance, customer administration and billing make up considerable part of the total cost.

Cost-efficiency in every phase of network development

Since the financial implications of a network vary throughout its lifetime, the long-term allocation of costs plays an important role in operator competitiveness. The costs of network maintenance, billing, and customer administration represent a big part of operator budgets. Fortunately, the potential for improvement in these areas – where improvement is synonymous with increased competitive strength – is considerable. In tackling the issue of cost-efficiency, operators must consider distributing over several years the costs of their original capital investments, network maintenance, and customer care, Figure 1.

Services for new and established operators

The Ericsson services for network operators enable operators to maximise the benefit of their investments in technology and to increase long-term cost-efficiency and competitiveness. Together with the operator, Ericsson analyse the operator's activities, processes and expertise, and plan programmes that complement and enhance overall operations. The scope of the analyses is adapted to the status of the operator organisation.

New operators need quick coverage as a base for gaining market shares. After getting established, they must improve capacity and service performance – thus striving to retain subscribers.

New operators, as well as established operators who are taking on new roles or branching into new markets, are supported by solutions that speed up roll-out and minimise financial risks. For these operators, a complete business analysis might be necessary. Established operators might require help in optimising and fine-tuning specific parts of their organisations and networks.

The customer services programme complements Ericsson's traditional offering of system support services, and matches new operator and end-user requirements.

Customised service solutions

The customer services offering is a concept for integrated service solutions that are tailored to fit the needs of individual operators. Service solutions cover every phase of network development, from initial planning to ongoing operations. By means of the customer services programme, Ericsson ensure that the complete network – not just traffic-carrying functions, but also billing and customer care – are up and running from the very start, thereby facilitating a rapid flow of revenue.

The services, which apply to all major digital and analogue technologies and standards, as well as to telecommunications networks, include advice on network performance and planning, and hands-on operation and maintenance.

The long-term objective of the service commitment is to make operators more successful by:
- reducing time to market;
- cutting overall cost;
- improving service order activation;
- ensuring more efficient customer care.

The Ericsson services for network operators enable them to expand their subscriber service applications in a step-by-

Box A

Abbreviations

BSS	Business support system
HLR	Home location register
MMIS	Maintenance management information system
O&M	Operation and maintenance
OSS	Operations support system
SIM	Subscriber identity module
VPN	Virtual private network

Figure 2
Ericsson's complete range of service offerings and customised solutions meets specific operator needs. Current examples of service include planning and optimisation projects in the USA, spare parts management in the UK, radio frequency change projects in Japan, a network surveillance and maintenance project in Italy, and training in South America and Hungary.

step fashion. They also enable operators to move key end-user services, such as call centre and point of sale, as well as customer care, to specialised applications. Thus, a high level of automation is achieved, speeding up customer processes and making each investment more productive and cost-effective.

The Ericsson service portfolio for network operators comprises four service areas:

- Professional services – for helping network operators to plan network and business operations.
- Implementation and integration services – for helping network operators to implement and install nodes or complete networks.
- Maintenance and support services – for helping network operators to operate and maintain networks and end-user services.
- Customer training – for helping network operators to establish and develop necessary competencies.

Planning ahead with professional services

Ericsson provide customised solutions to network and business operations. With these solutions, operators reduce their costs and optimise network use and performance. This strategy improves cost and revenue levels by decreasing churn, and by giving operators something besides tariffs with which to compete effectively. Ericsson consultants recommend strategies for planning networks, and help operators to define appropriate characteristics for new or expanding networks, conduct network performance evaluations, evaluate procedures and tools, and to create network models.

Examples of current service offerings are:

- business operations support and consulting;
- network operation consulting;
- network performance evaluation and optimisation;
- systems integration and support.

Business operations support and consulting

Business operations support and consulting is a service offering that comprises solutions to billing, customer care, point of sale, and fraud/churn management. All solutions are independent of network technology.

Ericsson ensure that network and support systems can interwork with one another, and that upgrades of billing and customer care systems can keep pace with releases of new technology and growing networks. Ericsson work exclusively with leading third-party vendors and subcontractors. The business operations support and consulting services enable operators to realise the full business potential of their networks.

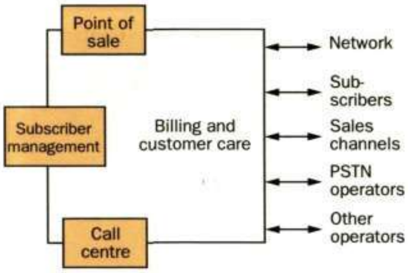


Figure 3
Active business operations support and consulting services help operators to give their subscribers improved levels of service. The Ericsson service solutions enable operators to recommend subscription packages, create flexible rating plans, and to activate new subscriber services more rapidly.

Network operations consulting and support

The services network operation consulting and network management system consulting are used to evaluate and re-engineer network operations as well as to introduce efficient support systems.

A typical offering of network management system consulting includes proposals for an operation and maintenance (O&M) organisation, procedures, and tools as well as for training paths. Once each of these aspects has been evaluated, the Ericsson consulting team recommend areas of improvement and change.

Examples of support systems are the maintenance management information system (MMIS), operations support system (OSS), and business support system (BSS). Customised statistical report packages can be complemented with consulting services for planning networks, changing radio frequencies, and administering the OSS and BSS.

The MMIS supports field maintenance by optimising O&M economy through the efficient handling and management of work orders and trouble reports. With the help of the management system, operators can establish cost-effective routines for field maintenance and for controlling the O&M organisation. The system also includes functions for keeping track of equipment in the network and in stores. The MMIS interfaces with the operations support systems of operator networks, allowing alarm and trouble ticket data to flow from the OSS to the MMIS.

Getting started with implementation and integration services

The Ericsson services facilitate implementing, installing and integrating nodes and networks. Network operators can

Box B Network performance and optimisation projects

Case 1, Brazil.
Tremendous subscriber growth in Brazil has caused operators to re-evaluate their networks in order to increase capacity and optimise network performance. Currently, Ericsson are assisting several Brazilian operators with network performance evaluation services.

After field measurements are taken, computers are used to analyse coverage efficiency, reliability, and network quality. Then, based on the data collected, an optimisation plan is developed and implemented. Action plans include changes to cell parameters, reduced (or increased) power levels, adjustments of antennas, and the re-tuning of cells.

Case 2, Portugal.
In 1995, the Portuguese telecom operator Telecel decided that, due to the demand for more sophisticated telecommunication services and because of a growing number of new competitors, it was time to re-evaluate their operations. They turned to Ericsson for advice on how they could improve all aspects of their service and network functions.

"The network performance review has been a very valuable tool. Ericsson's evaluation and their documentation revealed professionalism and a high level of support. The modifications that we implemented based on recommendations from the network performance review have been extremely useful to Telecel. We have improved both network performance and O&M procedures."

Voice of a Portuguese Operator

Case 3, Italy.
Telecom Italia was looking to improve quality of service in their network. Pending deregulation and the threat of increased competition made this essential. The network synchronisation review provided a natural starting point, as synchronisation quality has a major influence on quality in digital networks. Ericsson's review of Telecom Italia's AXE 10 exchanges showed a very well-managed network; still, an average of 10 improvements per exchange were recommended and implemented.

"This has been an opportunity to improve the reliability of the synchronisation network. The very well-structured final report allows us to prevent errors and to maintain the network in an effective way."

Voice of an Italian operator

Figure B
Customer service consultants discuss the business opportunities that arise from improving network performance.

apply these services to optimise their investments in technology and reduce time to market.

Ericsson's commitment comprises a complete package of services, such as site installation engineering, equipment installation and commissioning. In addition, Ericsson also manage comprehensive turnkey projects that handle everything from project management, civil construction and site engineering to installation. Also offered are service packages for introducing new nodes, such as home location registers (HLRs), and for adding to existing software or to network equipment.

Box C
Rapid time to market for nationwide networks in Argentina

In Argentina, Ericsson recently built two large wireless networks in record time. Less than ten months after the contracts were signed, the two networks, which span the entire country, were up and running successfully.

The combined area covered by the networks exceeds 2.8 million square kilometres. The projects involved the installation of 48 mobile services switching centres and 1200 radio base stations. Supervision and maintenance is handled centrally by Ericsson operations support systems (OSS).

All planning and installation of the networks was managed by Ericsson, who also commissioned the equipment.

Figure C
Installation engineers erect links for a new nationwide wireless network.

Systems integration (multi-vendor and multi-network)

Multi-vendor services comprise the integration of Ericsson equipment with equipment purchased from other suppliers, as is often the case when switches and network management equipment are purchased from different manufacturers.

Multi-network services are meant to ensure the operational compatibility of systems that are based on more than one network; for example, where a private virtual network is created from various parts of several otherwise separate networks.

Box D
Performance-based system support in Hong Kong

Service quality is crucial in extremely competitive markets such as Hong Kong. At present, ten operators offer wireless services to the demanding subscribers in the Hong Kong area. By mid-1997, the wireless penetration rate is expected to approach 25 per cent.

One of the most successful wireless operators in the growing Hong Kong market is *SmarTone*, with 259,000 subscribers reported in their network at the end of September 1996. SmarTone started operations in 1993.

In 1995, Ericsson and the SmarTone consortium signed a new type of support contract, in which Ericsson guaranteed certain levels of network performance. The support fees, which are paid as a percentage of operator turnover, are directly related to defined indicators of network performance – that is, to how well Ericsson provide support. Thus, when the network performs well, SmarTone and Ericsson each derives greater income.

The support agreement, which strengthens SmarTone's position in the competitive Hong Kong market, marks SmarTone and Ericsson's commitment to a long-term relationship that is based on clear, mutual objectives.

Figure D
Thanks to a service level agreement, in which Ericsson assume full responsibility for network performance, Hong Kong-based SmarTone can maintain a slim support organisation and provide superior quality to their subscribers.

Hands-on operation with maintenance and support services

The Ericsson services for operators introduce complete solutions to network operation and maintenance. By implementing these solutions, operators optimise the use of their network equipment, increasing network availability and performance as well as call quality.

Ericsson assume full responsibility for hardware and software maintenance and support activities, such as fault identification and analysis, and on-site operations support. Ericsson's commitment comprises routine and emergency support. Through traditional support activities, Ericsson help operators throughout the world to maximise system performance. Moreover, Ericsson have newly begun to introduce other types of services, including performance-based contracts and solutions that involve outsourcing O&M completely.

Complete responsibility for O&M

The network operation and maintenance service offering enables operators to outsource the full spectrum of network-wide O&M services, such as network management, field maintenance, and spare parts handling. Ericsson assume complete long-term responsibility for:

- the technical performance of the network;
- managing all staffing and training needs;
- implementing processes, procedures, routines, and support systems;
- handling spare parts.

With tailored-to-fit working methods and interfaces, operators benefit from highly-efficient day-to-day operations that comply with defined levels of quality. Moreover, operators are free to focus on their core activities, such as marketing, sales and customer care.

Network management and field maintenance

Network management services include fault management, configuration management, and performance management. The field maintenance organisation executes preventive and corrective maintenance actions in the network, and is responsible for the flow of spare parts.

Output from these network-wide O&M services are defined quality of service in the network and reports of network status and performance. The operator benefits of the Ericsson network operation and maintenance offering are the best possible performance for the money (price/performance) and much improved overall competitiveness.

Building competence with customer training

Ericsson customer training programmes apply the technical knowledge of operator staff in the best possible way. Training comprises programmes for managers and for operations and engineering staff, as well as certification and systems overview courses for sales staff.

Ericsson offer a wide range of computer-based and instructor-led courses, including telecommunications basics, on-the-job training, and high-level technical training. The courses cover every phase of network development.

Ongoing training activities enable operators to raise their internal levels of com-

Box E
A joint approach to staying ahead of the competition

In Italy, Ericsson recently built an international virtual private network (VPN), allowing the customer to further develop their business through a global offer. Ericsson was given the assignment of supporting this endeavor with new technologies, transfer of know-how, and through a range of services.

Establishing operations quickly – by choosing the right end-user services and with an aim to stay ahead of the competition – required a turnkey solution.

The international VPN consists of nodes in 14 cities worldwide. The customer's network management centre is monitored and maintained by Ericsson, who also planned and designed the network. The task was to get a voice-services network up and running in six months' time. Simple structure and billing, and efficient customer care were the cornerstones of the customer's offering. The network, which consists of six AXE switches and fourteen BMX nodes, connects major cities in Europe, the USA and Asia.

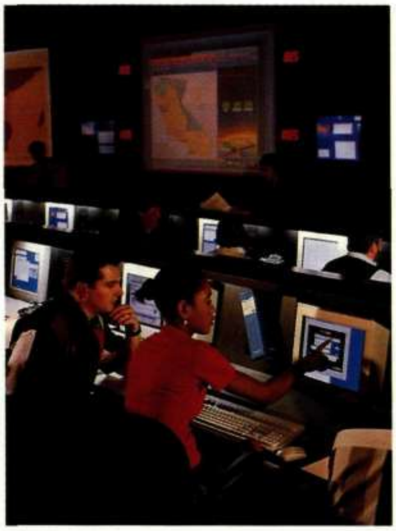


Figure E
A network management centre run by Ericsson.

Box F
Remote network support operations in Japan

Ericsson are building a national network centre to serve several of the wireless operators in Japan. The customer service centre in Yokohama will provide round-the-clock support, which ranges from handling routine queries to overcoming emergency situations.

Thanks to this centralised service, problems can be identified early, and network faults can be solved quickly, which considerably shortens response time and system downtime. In addition, the network centre enables upgrades and new services to be introduced very rapidly.

Through on-line connections from operations support systems (OSS), Ericsson will be able to provide rapid support. Furthermore, remote downloading reduces lead-time for software roll-out.

By means of the new low-traffic operation service offering, the complete network (or parts of it) can be operated from the network centre during off-peak hours, which can greatly reduce operator staffing costs.

The national network centre concept provides an ideal platform for future services, including multi-network offerings, as well as remote services, such as radio network investigation, performance management, and OSS administration.

Figure F
The customer service centre in Japan is a complete support service tool box for handling trouble reports, supervising networks, managing performance, and for remotely upgrading and co-ordinating activities during software roll-out. The centre also offers programmes for competence development and for a technical help desk.

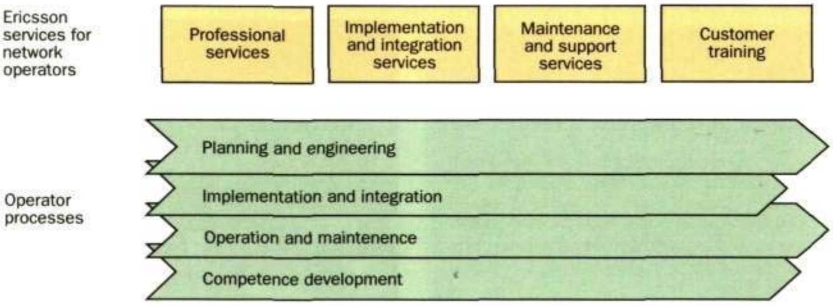

Ericsson services for network operators
Professional services
Implementation and integration services
Maintenance and support services
Customer training
Operator processes
Planning and engineering
Implementation and integration
Operation and maintenence
Competence development

Figure 4
Ericsson services for operators match operator business processes with effective tools for differentiating their service offerings.

petency as fast as their networks expand. In addition, clear training strategies help motivate staff and improve working methods and efficiency.

Courses are held at Ericsson's international training centres as well as on operator premises.

Conclusion

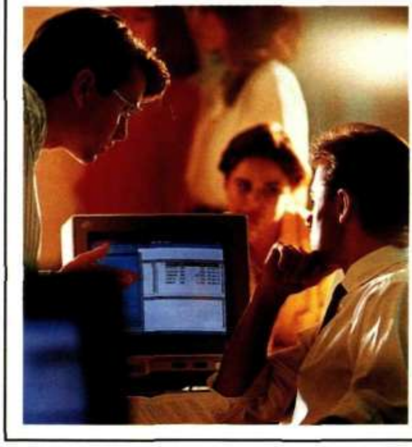
The costs of network maintenance, billing, and customer administration dominate operator budgets. Improvement in these areas strengthens overall operator competitiveness. Through optimising the cost-efficiency and productivity of operations, the Ericsson services for operators play a decisive role for success.

The customer services offering is a concept for integrated service solutions that are tailored to fit the needs of individual operators. Service solutions cover every phase of network development, from the initial stages of planning and design to ongoing operations. More specifically, Ericsson services for operators comprise four areas: professional services, implementation and integration, maintenance and support, and customer training.

Each service area focuses on different aspects of operator organisations: expert guidance for planning new networks or for expanding existing ones; global turnkey network solutions; hands-on network operations; maintenance and support, such as local support with access to global expertise, 24 hours a day, year round; and training at all levels.

The long-term objective of Ericsson's service commitment is to make operators more successful, by reducing time to market, cutting overall cost, improving network performance, and by ensuring more efficient customer care. Today, network operators and service providers are increasingly likely to seek complete business and network solutions from their suppliers. With experience, skill and market presence to give support on a truly global basis, Ericsson's services help operators to stay ahead of the competition.

Figure G
Ericsson operate six international training centres. Each centre is equipped with modern equipment, boasts a professional staff of expert instructors as well as computer-based applications, and includes a guest service centre.



Box G Extensive training programmes

Case 1, Argentina.
In 1996, Ericsson provided training for several hundred people from two operator organisations in Argentina. Tailored-to-fit courses on management, O&M, and engineering prepared staff to run two separate nationwide networks.
On-the-job and instructor-led training were also provided by instructors from the Ericsson international training centre in Mexico.

Case 2, Hungary
Training has been a strategically important part of the Hungarian Telecommunications Company's (HTC) network development plan ever since the company selected the Ericsson AXE 10 switching system for its national network in 1991.
As this was the first time AXE 10 local exchanges were introduced into the Hungarian network, Ericsson provided a complete network implementation training package. The aim of the package was to facilitate the network roll-out, and to enable HTC's staff to operate and maintain the equipment independently. Ericsson worked closely with HTC to determine how many peple needed to be trained as well as at what levels, after which they drew up a comprehensive long-term training plan.
The training programme, which was initiated before installation work began, started with HTC staff who were responsible for system approval. Next, system technicians and engineers were instructed in operating and maintaining the network. Later, more advanced training was provided.
To ensure a smooth transition from the training centre to operational exchanges—after the first AXE 10 exchange had been installed—experts from Ericsson provided on-the-job instruction for seven months.
Another aspect of HTC's plan for self-sufficiency was the decision to set up its own training centre to meet future needs. Ericsson supported this incentive, first by helping train the HTC instructors, and later by supporting these instructors as they taught their courses for the first time.

Head-mounted displays

Hans Brandtberg, Peter Segerhammar and Claes Waldelöf

Head-mounted displays (HMD) – small computer screens attached together with an optical system in front of the user's eyes – are increasingly employed as a means of creating a "virtual reality" in a wide range of professional applications. In training situations, HMD systems are used for computer simulation and for presenting database models of real-world phenomena to create what is called a synthetic environment.

The authors describe Ericsson Saab Avionics high-performance head-mounted display, which is intended for industrial and military applications. Ongoing development of future head-mounted displays for military aircraft is also described.

Head-mounted displays (HMD) are not just novelties to be used for amusement in virtual-reality and cyberspace computer games. They have also proved to be effective in many professional applications. For instance, manufacturers of cars, aircraft, ships and trains find the HMD a valuable development tool that cuts lead time and costs. In the process industry, HMDs will be used for efficient control of processes. Driver and pilot training are other areas where the HMD will play an important role. In defence applications, the HMD will be used for training and for command and control. Employed in telecom systems, an HMD will enable users to receive information hands-free.

HMDs have been introduced in two major types of application: as a virtual-reality display, to create in the user's mind the impression of finding himself in a virtual reality – a computer-generated world; and as an information display that presents information to the user in situations where he cannot reasonably have access to a normal visual display unit (VDU). For obvious reasons, this is the case when he is on the move – walking, driving a vehicle or piloting an aircraft.*

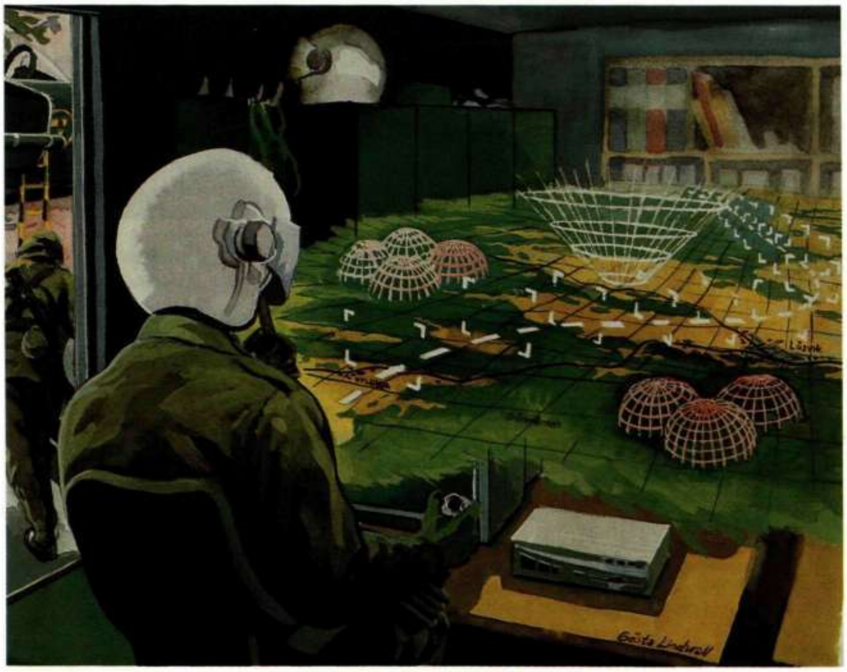


Figure 1
With a virtual-reality system, the user is under the impression that a synthetic world exists in front of him – even "on his desk". He can also move about and act in this world.

An HMD consists of two small VDUs – one for each eye – to be worn by the user, as illustrated in Figure 1. An optical system between the display medium and the eyes provides a suitable viewing distance. Unlike a normal VDU, the HMD "encloses" the user in the displayed picture, which contains the computer-generated information. This effect is produced through a powerful computer with high-performance graphics, a database from which the "virtual reality" is sourced, and a software package for drawing and presenting pictures. Another central part is a head-tracker, which senses the user's head movements. An HMD system is often complemented with audio generators and interactive tools, such as a joystick or a computer glove.

The problem of presenting useful – and often vitally important – information in an efficient and ergonomically acceptable way places exacting demands on the HMD, mainly in terms of optimal field of view and resolution versus light weight and small design.

Ericsson Saab Avionics are working on two main types of HMD products. An HMD for professional virtual-reality applications has been developed and introduced in the marketplace. Different types of special HMDs for aircraft applications are also at an early stage of development.

* Head-mounted displays are used by both men and women. References to users in this article appear in the masculine form solely to simplify sentence structure.

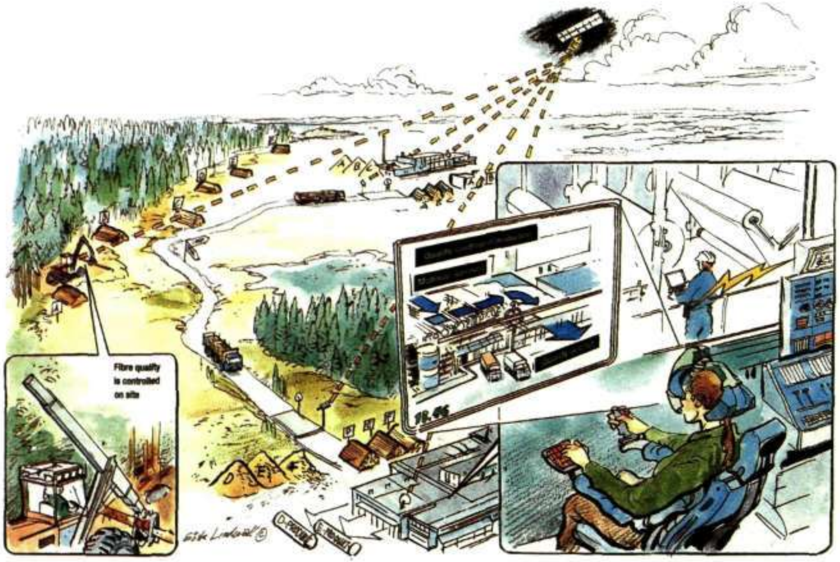


Figure 2
An HMD can be used in process control applications – in the wood-working industry, for example, to control and supervise different processing stages.

Augmented reality

Sometimes reality is not good enough, but has to be "improved". This can be accomplished through the use of powerful, picture-generating computer systems and a head-mounted display worn by the user.

If the user is moving about in or over a sector of terrain – in a vehicle or helicopter – he has to keep track of roads, houses, masts, power lines, etc. If visibility is poor, or if it is dark, an HMD system can help the driver or pilot to identify these features and objects. Selected, computer-generated, man-made objects and natural features can be displayed on a see-through visor in front of the user's eyes and superimposed on "the real world". This is an example of augmented reality.

A highly realistic virtual reality can be generated as a three-dimensional presentation by powerful graphics computers and software packages. This information may have its origin in databases related to the existing environment, such as geographical and topographical data. Data from complementing sources, describing geological or hydrological variations in the scenery, can be superimposed on or mixed with the real image, thus improving the user's view of reality – again an example of augmented reality.

If an imaging sensor is used – for example, a video camera, an infrared camera or a radar device – objects and features from a database can be superimposed on the sensor image and the result visualised on a display. All systems for augmented reality require high accuracy and agreement between reality and the computer-generated information. They must therefore include a tracking and/or navigation system that keeps the computer informed of the user's whereabouts, in order to make it possible for objects and features to be drawn in their correct positions relative to the user.

Another type of augmented reality is used in medical examinations and operations. The doctor can "look into" a patient with the help of images from different diagnosis systems; for instance, X-rays projected onto the body of the patient. The medical field as such presents several worthwhile applications for head-mounted displays and the associated processing and presentation of information.

Virtual reality and synthetic environments

A synthetic environment is a computer-generated copy from a database describing a real environment. Virtual reality could be any imaginary environment or world.

Simulators – when used to train pilots or drivers – create a synthetic environment that enables the user to "fly" an aircraft or "drive" a vehicle while seated in a laboratory under more or less realistic conditions.

Virtual-reality applications also include the area of civil engineering; for example, in planning and constructing roads or bridges. The new infrastructural feature is modelled and presented by the HMD system, and a head-mounted display then makes it possible to "drive" a car on the non-existent roads or bridges. Road safety and even aesthetic considerations – which may influence the decision-making process – can be evaluated in the virtual reality.

In the aviation industry, an HMD system can be used for "testing" the roominess and comfort of a passenger aircraft before it has been developed. A computer-generated presentation of the interior of the aircraft allows a person equipped with a head-mounted display to walk

Ericsson Saab Avionics AB

Ericsson Saab Avionics AB, whose main operations are in the field of military avionics, is a joint-venture company owned by Ericsson and Saab. Several of the company's products are used in the subsystems of the new Swedish fighter aircraft JAS 39 Gripen.

along the aisle, and to look around while seated.

When the architect has designed the house you are planning to buy, you can make a "realistic" inspection of it, with all its mod cons, before you buy it.

Industrial applications

One of the most important industrial applications for an HMD system is in the field of process control and management. All stages of a production process – including the input of raw material, refinement, product manufacturing, stock management, and delivery – can be adequately visualised and efficiently controlled and supervised, Figure 2.

Fleet management of lorries and ships approaching and leaving a harbour is also an important application. Sensors and communication systems, such as vessel or vehicle transponders, radar, and video cameras, are main system components. The environment, with data of objects and their visual representation, is displayed in layers of different information depth, supplying the operator with current, accurate data so that he can manage the movements of fleets.

A third important application is the design and development process for advanced constructions and systems. In the design phase, a product can be modelled and realistically "used" and "tested" for operability and ease of repair before it is developed.

Defence and security applications

On the military scene, highly sophisticated communication and sensor systems exist, but the conditions under which they are to be operated often put tremendous demands on staff, notably in situations that require extremely short reaction time.

This is one of the reasons why HMD systems come in handy in many military applications. In a command and control centre, for example, they lend themselves to arena supervision and unit control. Sensor systems and intelligence reports are used as sources of information. A zone of action is visualised to the command and control operator with the help of an HMD and graphics computers that are connected to communication systems and information databases.

Air reconnaissance

Special operators who control and operate unmanned aerial vehicles (UAV) can also use HMDs to supplement the results of reconnaissance operations. Data obtained from sensors on the UAV can be down-linked to the ground control post, presented in real time and processed. The information – after it has been correlated with detailed data stored in databases – is then visualised by the HMD system. This enables the operator to interpret the information using automatic target-detecting software and image improvement algorithms. The UAV's flight path can also be intermittently altered and optional sections or areas minutely examined.

The sensor equipment on the UAV may include an ordinary TV camera, an infrared camera, or a radar device.

Ground operations

In a military vehicle, such as a main battle tank, the driver and the tank commander can see a visualisation of data obtained from sensors and databases together with tactical information that is superimposed on the "environmental" data. This makes for safe driving at night and in adverse weather conditions. It also facilitates cooperation with mechanised infantry, artillery and air support, Figure 3.

Box A Definitions

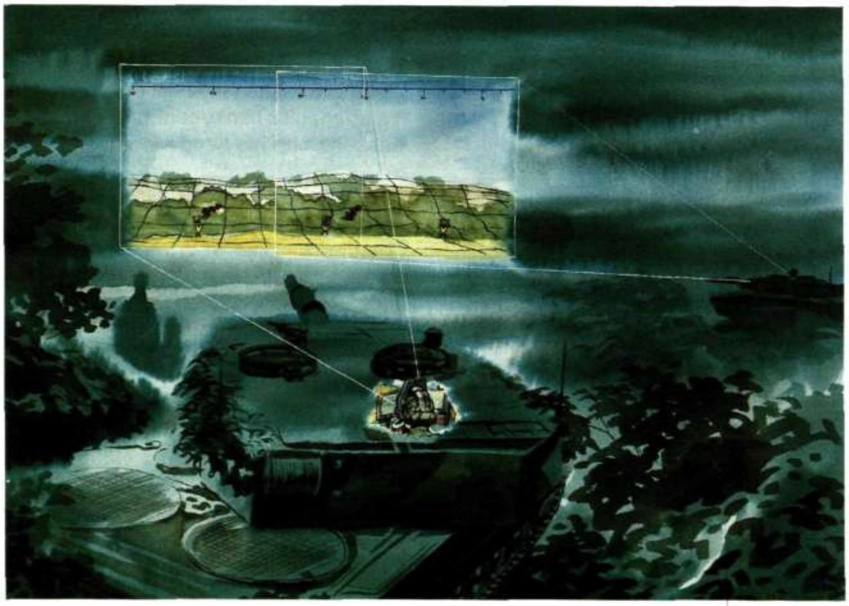
Augmented reality (AR)
"Improved" reality, in which a computer-simulated presentation is superimposed on the real world, or on a sensor image of it, in order to improve the interpretation of pictures of real terrain and real objects.

Head-mounted display (HMD)
Two small computer screens (one for each eye) equipped with an optical system and worn by the user; images are always displayed within his field of view regardless of head movements.

Synthetic environment (SE)
Computer simulation and presentation of a database model of real-world phenomena. SE is a standard feature in flight and driving simulators.

Virtual reality (VR)
Computer simulation and presentation of a three-dimensional surrounding world (either the world in its actual existence or a new, invented one) in which the user finds himself. The position and movements of the user's head are sensed by a head-tracker. Wearing a head-mounted display, with a fixed position relative to his eyes, the user can look around, move about and take action in the displayed "reality".

Figure 3
An HMD in a battle tank application gives the crew in the tank a complete and realistic plan view of the zone of action in which the deployment of friendly and hostile forces can be presented.



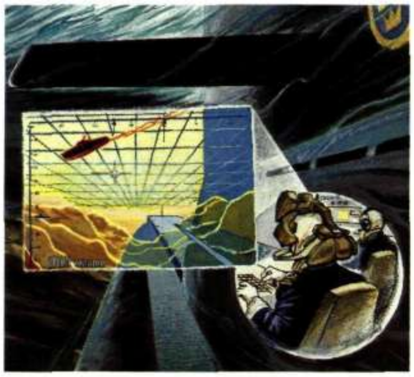


Figure 4
A submarine crew supervises their mission by means of different types of instruments. Using an HMD supplemented with presentations sourced from a bottom topology database, the user can inspect the submarine environment as though he were in a "glass craft" surrounded by clear water.

Submarine operations
So far, submarine crews have had to navigate in the dark depths of the sea without any visual references. Now that HMDs are available, sensor information – adequately mixed with data of the sea bed topography and data from electronic charts – will simplify navigation and increase submarine operators' awareness of their locality, Figure 4. In addition to presenting information about the surroundings, an HMD can display pictures of the interior of the vessel and of ongoing activities.

Air operations
In a helicopter or fighter aircraft, a head-mounted display not only supplies the pilot with relevant positional and situational information in his line of sight, but it also enables him to control sensors and weapons with his head movements.

Training

Operator training (of drivers, pilots, etc) will be highly improved using HMDs. Simply by sitting down, putting the helmet on, and connecting it, the driver is inside a vehicle.

Virtual-reality HMD

Ericsson have developed a virtual-reality (VR) HMD for industrial and military applications, Figure 5. This HMD is intended for the high-end market and features excellent image quality combined with flexibility.

The display is housed in an easy-to-wear headset. It has an image-collimation optical system and a high-resolution, full-colour display with a wide field of view and stereo viewing capability. The HMD fulfils requirements for crisp and user-friendly presentation, which makes it ideal for a wide range of professional applications. It is designed for combined virtual-reality and see-through presentation.

The image presentation part of the HMD is located in the front of the headset. An optical system and a cathode ray tube (CRT) are provided for each eye. The CRT itself is black and white. Colour is obtained by means of a colour shutter and field-sequential technique. A short distance between the CRTs and the high-voltage power supply (HVPS) is important for achieving the best possible image quality. Good balance is another important point. Therefore, to counterbalance the optical systems and the CRTs in the front, part of the electronics – the HVPS, video and deflection electronics – is installed in the rear of the headset, close to the user's neck.

Some special arrangements and controls are needed for the HMD to fit different head sizes, and to keep it firmly in place. The HMD is designed to fit more than 90% of the population, both males and females.

The neck-mounted package is attached to the headset through a metal pin. The "integral frame" in front of the user's chin is not only a protective device, but is also used to squeeze the HMD headset together, keeping it firmly in place on the user's head. This is done by pumping the frame up and down. The HMD is quickly released by pushing the frame upwards.

The optical systems, with the CRTs, must be adapted to different users: distance between the eyes, distance down from the forehead to the eyes, etc. The HMD is therefore equipped with an XYZ table. The optical systems can be moved up or down, in or out, and together or apart. The field of view is also adjustable, which means that the horizontal field of view can be increased, Figure 6.

Control electronics and the low-voltage power supply are housed in a special, easy-to-wear box that can be fastened to the user's belt. A small unit contains brightness and contrast controls and an ON/OFF switch. This unit can be attached to the user's shirt, for example.

Box B

Features of the HMD

- Head-mounted display unit and easy-to-carry electronics box
- Quick and easy to fit, using the front loop
- Easily adjustable to different head sizes
- Well-balanced and easy to wear
- Robust and lightweight carbon-fibre design
- Wide field of view, stereoscopic colour presentation
- High resolution and brightness
- Selectable virtual reality or see-through
- Multi-video standard capability with field-sequential input
- Compatible with magnetic head-trackers
- Low electromagnetic field according to ENV 50166

Technology behind the VR HMD

General

Picture resolution, field of view and size (volume and weight) are three critical parameters in the design of an HMD. The user wants high resolution, a wide field of view, and a headset the size and weight of a pair of sun-glasses. This is impossible with today's technology.

HMD products on the market fall into two categories: low-cost and small-sized displays with poor resolution and medium field of view; and high-cost displays with high resolution and a wide field of view.

Low-cost displays are designed with flat-panel devices for image presentation. The most common element today is a low-resolution liquid crystal display (LCD) of the type used in the viewfinder of many home video camcorders. The resolution of such a system is about 240 TV lines. Professional applications require much higher resolution, which means that the quality of the standard 1024 x 1280 pixel display is well worth aiming at.

So far, the CRT is the only display element that offers a resolution of 1000 TV lines on a small area of approximately 21 x 28 mm, but this CRT is monochrome. The spatial colour-pixel method used in home TV sets gives pixels (red, green and blue) that are much too large and which, unfortunately, cannot be sufficently decreased. Instead, a so-called temporal field-sequential method is used. The CRT is equipped with an electronically switchable colour filter (shutter) that can change colour from red, to green and to blue. The complete colour picture is obtained by first presenting a red image, then a green one and finally a blue one. This method of displaying the primary-colour images sequentially at a high rate makes the user perceive a full-colour picture.

To avoid flicker, the picture frequency of a computer colour screen has to be at least 60 Hz. Dividing the presented picture into three fields – one for each primary colour – will thus require a picture frequency of 180 Hz.

The field-sequential technique applied in the HMD

The HMD is based on a monochrome high-resolution CRT and uses the field-

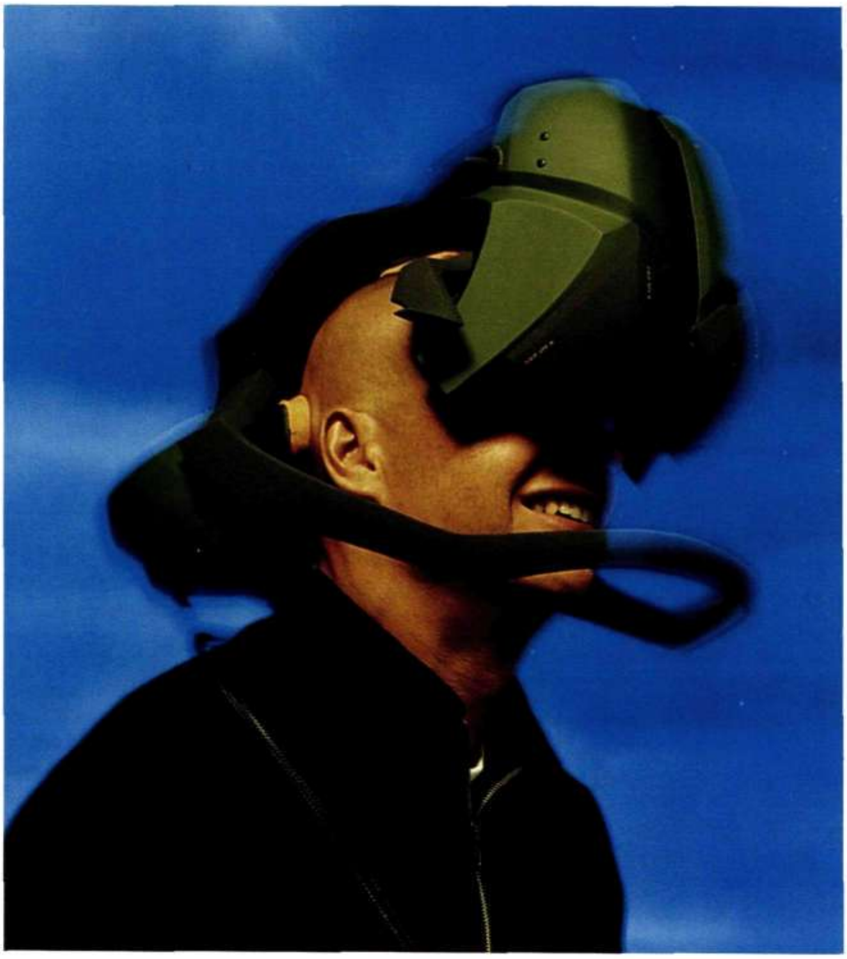


Figure 5
The virtual-reality HMD.

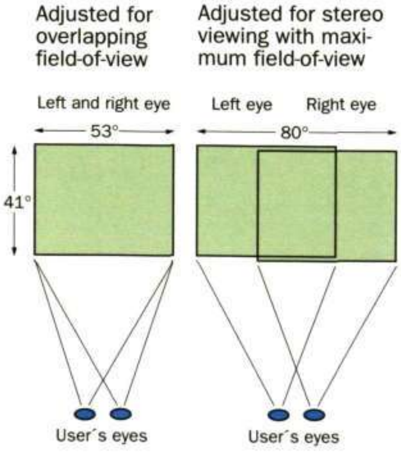

Adjusted for overlapping field-of-view
Adjusted for stereo viewing with maximum field-of-view
Left and right eye
53°
41°
Left eye
Right eye
80°
User´s eyes
User´s eyes

Figure 6
The optical system can be adjusted for different degrees of overlap between the picture seen by the left eye and the picture seen by the right eye.

Box C
Technical data:

Display source	Full colour, from 35 mm B/W CRT with fast ferroelectric liquid crystal (FLC) shutter
Field of view	53° H x 41° V with 100% overlap, adjustable to 80° H with 32% overlap
Interpupillary distance	56 - 76 mm
Eye relief	15 mm (adjustable 15 - 27 mm)
Luminance	25 foot-Lamberts (fL)
Resolution	0.3 MTF at VGA standard (0.001" line width)
Controls	Brightness, contrast and quick ON/OFF Five degrees of freedom in optical adjustment
Video formats	640 x 480 1280 x 960, interlaced 1280 x 492
Video input	Field-sequential RGB Video: 0.7 - 1.0 V, 75 ohms
Power	28 V DC , 50 W (230, 110 V AC adapter optional)
Weight on head	4,2 kg (9.2 lbs)
See-through transmittivity	50%

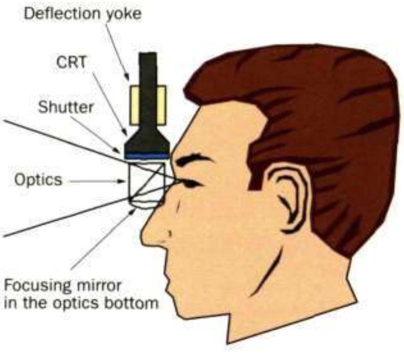


Figure 7
An optical system – with the shutter and CRT on top of it – is attached in front of each eye. The picture passes the shutter, is reflected and focused by the mirroring surface in the lower part of the optical system, and reflected to the user's eye by the semi-transparent flat surface in the middle.

sequential technique together with a ferroelectric colour shutter. The CRT and the shutter are mounted on top of an optical system that focuses and reflects the image to the user's eyes. The user can also see through the optical system and look at the surroundings by pushing up a cover. The principle of the CRT, the shutter and the optical system is shown in Figure 7.

The shutter, which uses the birefringence (double refraction) of a ferroelectric liquid crystal (FLC) material, is built up of three elements:

- A linear polariser—the first of these elements to be passed by the light from the CRT.
- A ferroelectric cell—this cell rotates the polarising angle of the linearly polarised light when an electric field is applied across it.
- Another linear polariser called an analyser.

When an electric field is applied across the ferroelectric cell, the red wavelength of the light rotates to an angle that matches the transmitting direction of the analyser. This causes the shutter to assume the character of a red colour filter. By changing the electric field, the shutter allows the green light to pass, and another change makes the blue light pass. The ferroelectric material is capable of changing colour in less than a millisecond.

Picture generation

The electronics required for generating pictures on the two CRTs include a video amplifier, synchronisation circuitry, a deflection amplifier and deflection yokes, as well as low-voltage and high-voltage power supplies. Control electronics for the shutter are also required, to set the shutter to "red mode" when the red picture is displayed, and so on for green and blue. Figure 8 shows a block diagram of the control electronics.

The input to the HMD is a field-sequential RGB (red – green – blue) video signal. This means that the video signal components are sequential when they interface the HMD. To generate such a signal, the PC or workstation needs a special graphics board (available from different suppliers). The HMD can also interface a standard RGB output when the PC or workstation is set to generate the same picture in three subsequent video frames. It then selects the red picture from the red signal component in the first frame, and the green and blue pictures from the respective components in the second and third frames.

Use of HMDs in fighter aircraft

General requirements

The conditions under which a pilot of a modern fighter operates are contradictory. To fulfil his mission the pilot must have as much information as possible: about his position and situation, the positions and intentions of his wingmen, threats and targets, etc. Yet he must not be distracted by this information while he concentrates on flying his aircraft and searches visually for obstacles and hostile targets. A solution to part of the problem is to project the information onto the visor of the pilot's helmet so that it is presented to him concurrently as he scans the surroundings, Figure 9.

However, the weight and volume added to the head-mounted display by the additional functionality creates another problem, and a major one.

In today's high-performance fighters the pilot will be repeatedly exposed to high g-loads for periods of several seconds (in extreme cases up to 30 seconds). An ejection from the aircraft – an eventuality that must always be taken into account – may produce a thrust of up to 20 g. Needless to say, a heavy HMD system might impair the pilot's performance, even under normal flight conditions; therefore, it is essential that his helmet be as light and well-balanced as possible.

Integrated helmet system, IHS

Air force personnel use the term integrated helmet system (IHS) for a head-mounted display. This term will therefore be used in the following discussion of the design process.

The design of an integrated helmet system for use in an aircraft must incorporate several considerations:

- high g-load during flight
- extremely high g-load, wind blast and possible high pressure and temperature gradients during ejection
- bright ambient light
- fast-moving, non-cooperating targets
- personal protection of the pilot

An integrated helmet system will include many new features, notably a display system that consists of display media, optics and a reflective visor, a head-tracker, and protection against biological, chemical and laser weapons. The technical challenge presented here involves the introduction of new functionality without exceeding the weight and volume of today's type of helmet.

The IHS is part of a greater, complex system that includes the aircraft, its subsystems, and the pilot. In order to achieve a successful design, a number of factors have to be taken into account. Time delays in the system – from the sensing of the position of the pilot's head up to the presentation of a picture – are critical and must be minimised. To fly the aircraft safely, the pilot must have a wide field of view. He must also have what is called a large exit pupil, so that he sees the image even if his helmet slips a bit during manouevres. Satisfying these demands is a delicate balance, which involves adapting the display format to the exacting field-of-view requirements.

The limited size and the complicated design of the cockpit present other difficulties that must be overcome. The head-tracker system is either a magnetic or an optical transmitter/receiver system. If a magnetic system is used, a magnetic-field transmitter has to be placed close to the helmet, preferably above the pilot, and the helmet must be equipped with a receiver. An optical system can use two or three cameras and a pattern of infrared diodes on the helmet. The latter type of system requires an unobstructed line of

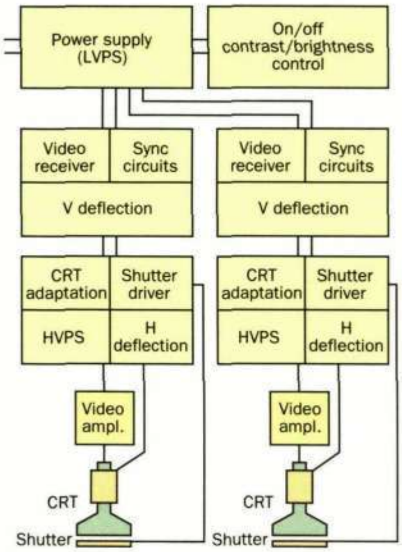


Figure 8
The electronics are designed so as to drive the two CRTs in parallel, equally for the left and right eyes. A common low-voltage power supply (LVPS) is used as well as a control module. HVPS is the high-voltage power supply.

Figure 9
In a fighter aircraft, the pilot uses HMD presentation together with other displays on the instrument panel. Important information is always displayed in front of his eyes. He can also point out targets to the HMD by placing a displayed marker in the respective target positions.

sight from the cameras to the helmet, which can be problematic when the pilot moves his head.

The bright ambient light when flying at high altitudes places special demands on the displays in the cockpit, in terms of brightness and contrast. So far, the only possible solution has been monochrome green CRTs. However, using green CRTs as IHS display media poses a problem. If a high reflection rate is required of the display media in the visor, this will result in a low transmission rate, and the pilot will not be able to see the green head-down displays.

Test and evaluation

The design of an IHS requires an interdisciplinary approach, where experts in avionics, physics, medical engineering and human factors work together in order to reach a solution that is acceptable to the user; in other words, the pilot. Together with the Swedish companies FFV Aerotech AB and Saab AB, Ericsson Saab Avionics AB is preparing a requirements specification for an integrated helmet system adapted to Swedish conditions. The design of a demonstration system is under way.

A test and evaluation programme for an IHS is carried out at Ericsson Saab Avionics, to define the minimum functionality needed to perform a flight mission, and with the overall objective of reducing weight and volume. For this purpose, a simulator (EPSIM) is used with an aircraft cockpit, a set of simulated aircraft functions and a synthetic environment with an "out-of-the-window" view of 135 by 35 degrees. The demonstration equipment includes a CRT-based system – with a monocular or binocular 40x30° field of view – mounted on a lightweight flying helmet. The head-tracker is an off-the-shelf magnetic system.

The future

The VR industry is often referred to as an emerging "zero-billion-dollar industry" (that is, one that might well turn over a billion dollars annually). The entertainment sector will probably see the sharpest expansion. Of course, the HMDs developed by Ericsson Saab Avionics are targeted at the professional sector, which is also expected to show substantial growth.

The size of VR HMDs for industrial and military applications will "approach that of a pair of sun-glasses". The next development step will benefit from new, small, flat-panel colour displays with high resolution. This will lead to a significant decrease in the weight, volume and price of an HMD and very likely widen its use still further.

The integrated helmet system for aircraft will also benefit from these high-resolution flat-panel displays. It will be easier to meet the stringent requirements for weight, volume and wide field of view that apply to this type of application. Integration of aircraft functionality will also be simplified.

The use of HMDs in industry, in command and control centres and in different types of craft, such as aeroplanes and helicopters, will expand and open up new possibilities for different categories of users.

Conclusion

"Virtual-reality" head-mounted displays – small computer screens attached together with an optical system in front of the user's eyes – have proved to be an efficient means of assistance in a wide range of industrial and military applications. The synthetic environment created enables the user to practice advanced skills, such as flying a fighter aircraft; controlling and supervising processes in production and logistics; testing in a realistic manner non-existent roads and bridges, etc. Head-mounted displays are also used in medical examinations and operations.

HMDs will play an important part in the design and development process for advanced constructions and systems. For example, in the design phase, a product can be modelled and realistically tested for operability before it is developed.

HMDs may prove valuable in a field where the importance of a means of assistance is not readily quantified: in the harsh reality of fire-fighting and other rescue operations.

Future HMDs will be based on new high-performance flat-panel displays. Together with fast graphics workstations and software packages these will provide the user with a comfortable and highly efficient tool that is invaluable in tomorrow's information communications world.

Illustrators:
Gösta Lindwall, Lexivision; and
Alf Lannerbäck
Photographer:
Kent Johansson, Loke Bild

Telefonaktiebolaget L M Ericsson
S-126 25 Stockholm, Sweden
Tele: +46 8 7190000
Fax: +46 8 6812710

ISSN 0014-0171
Ljungföretagen, Örebro 1997

Ericsson REVIEW

THE TELECOMMUNICATIONS TECHNOLOGY JOURNAL

NO 2, 1997

ERICSSON

Ericsson REVIEW

The telecommunications technology journal

The purpose of Ericsson Review is to report on the research, development and production achievements made in telecommunications technology at Ericsson. The journal is published in English and Spanish and distributed quarterly to readers in more than 130 countries.

Address:
Telefonaktiebolaget LM Ericsson
S-126 25 Stockholm, Sweden

Tel: +46 8 719 00 00

Fax: +46 8 681 27 10

Internet address:
http://www.ericsson.com

Publisher: Håkan Jansson

Editor: Eric Peterson

Production manager: Eva Karlstein

Layout: Paues Media, Stockholm

Printer: Ljungföretagen, Örebro

ISSN: 0014-0171

Cover: In just two years, engineers at Ericsson have reduced the size of AXE hardware by 90%, producing an AXE switch that fits in a single cabinet.

Contents

Contributors

In this issue

Urban Hägg

Tomas Lundqvist

Bo Stockman

Arne Wallers

Terenzio Paone

Ulf Hansson

Tom Lindström

Magnus Ericsson

Neela Koria

Urban Hägg has held several positions in AXE product management since joining Ericsson in 1978. He is currently responsible for strategic planning of the AXE platform at Ericsson Utvecklings AB. He holds an MSc in physics engineering from the Royal Institute of Technology in Stockholm.

Tomas Lundqvist, who joined Ericsson as a system engineer in 1994, currently works with hardware system design in the AXE Systems Management department at Ericsson Utvecklings AB. He holds an MSc in electrical engineering from Chalmers University of Technology, Göteborg.

Bo Stockman helped define requirements for, and participated in evaluating, the new BYB 501 metric equipment practice. Today, he is a part of AXE Hardware Systems Management at Ericsson Utvecklings AB. He joined Ericsson in 1965 after receiving an MSc in electrical engineering from the Royal Institute of Technology, Stockholm.

Arne Wallers, who joined Ericsson in 1968, is currently a technical expert at Equipment Practices Technology, Core Unit Basic Technology, Ericsson Telecom AB. He has worked with equipment practices and related characteristics since 1990. Before that he worked with analogue electronic and optical-fibre transmission systems. He holds an MSc in electronics from the Royal Institute of Technology in Stockholm.

Terenzio Paone, who is the provisioning manager at Ericsson Telecommunicazioni, Italy, is currently responsible for product development in cooperation with Ericsson Utvecklings AB. He joined Ericsson after having earned an MSc in electronic engineering from the University of Rome in 1975. Until 1986 he assisted in researching and developing cable/optical transmission systems. Later, he worked with network management, customer premises networks, and ATM systems.

Ulf Hansson is a technical coordinator for the group switch subsystem and APT devices at Hardware Development, Ericsson Utvecklings AB. He joined Ericsson in 1981 after graduating with an MSc in electrical engineering from Linköping Institute of Technology.

Tom Lindström works in the area of telecommunications standards and regulations at Telefonaktiebolaget L M Ericsson where, since 1992, he has been in charge of corporate coordination and company representation in management assemblies of international standards organisations. He holds an MSc in physics engineering from the Royal Institute of Technology, Stockholm.

Magnus Ericsson has worked with system management for the adjunct processor at Ericsson Utvecklings AB since 1995. Since joining Ericsson in 1987, he has worked in design, at Ericsson Telecom AB, and in product management, at Ericsson Hewlett-Packard AB. He holds an MSc in physics from Stockholm University.

Neela Koria is a product manager in the Strategic Management Group at Cellular Systems, American Standards, Ericsson Radio Systems AB. She holds a BSc in electronic and electrical engineering, awarded by the University of Surrey, UK.

From the publisher

Håkan Jansson

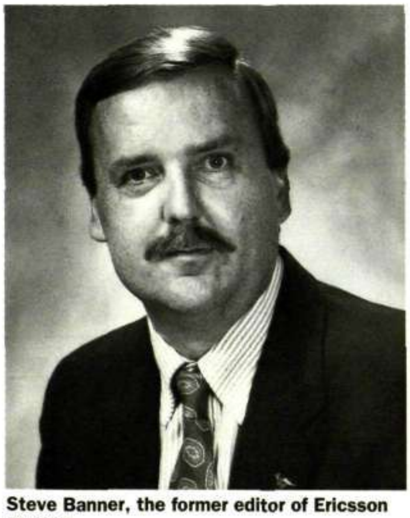
On April 30, 1997, Steve Banner left his position as editor of Ericsson Review for a new post at Corporate Technology, to take up the newly created position of Product Portfolio Information Manager. Succeeding Steve as editor is Eric Peterson.

During his two years as editor, Steve edited over forty articles describing the research, development and production achievements made in telecommunications technology at Ericsson. This has not been an easy task, given the rapid pace at which technology is advancing, and the vast array of technology that makes up the field of telecommunications. Steve also initiated various proposals for improving the journal, laying the foundation for some of the changes that are evident in this issue. For example, the Ericsson Review now clearly states its purpose, however simple, in a mission statement. Also, the journal now incorporates a list of the patents that have been awarded to Ericsson during the past quarter. What is more, the journal is now published in its entirety on the World Wide Web (English version only), ensuring that a much larger circle of readers can access it.

Looking back over his tenure as editor, Steve commented, "It has been a privilege to have been able to present, through the medium of Review, a measure of the outstanding technical expertise that exists in Ericsson's organisation around the world. I am confident that the enthusiastic and very capable hands of Eric Peterson and his editorial staff will ensure that Review continues its long tradition of reflecting the considerable talents of Ericsson's employees. The fast pace of the telecoms industry guarantees that many exciting experiences lie ahead for all concerned!"

Eric has been with Ericsson for four years, during which time he has written, translated, produced and edited technical literature for a variety of media, including Ericsson Review and the World Wide Web. Immediately prior to joining Review, Eric held a position as Web Infomaster, in which role his responsibilities included designing templates and writing publishing policies. He also trained local information owners in the basics of Web authoring and publishing. Eric was born and educated in the United States, and graduated from the University of Oregon with a Bachelor of Arts degree in Japanese language and literature. Having lived and worked in several different countries, Eric speaks Spanish and Swedish, in addition to Japanese and his native English.

Håkan Jansson
Publisher



Steve Banner, the former editor of Ericsson Review.

Eric Peterson, the new editor of Ericsson Review.

AXE hardware evolution

Urban Hägg and Tomas Lundqvist

The AXE system is the most widely deployed switching system in the world. It is used in public telephony-oriented applications of every type, including traditional fixed network applications in local, transit, international and combined networks. AXE is also deployed for all major mobile standards – analogue as well as digital. AXE is very strong in intelligent networks and other real-time database applications. Recent designs also enable data communication capabilities to be added to the system.

From its inception, the AXE system was designed to accommodate continuous change. Throughout the years, new applications have been introduced, its array of functions has grown, and its hardware has been steadily updated.

The authors describe how the latest advances in hardware technology have been brought into the system, thereby dramatically improving such characteristics as floor space, power consumption, system handling, and cost of ownership. As always, backwards compatibility has been maintained to the greatest possible extent, in order to protect previous investments in AXE.

The hardware used in the AXE system has been updated continuously. Initially, all telephony-related hardware in AXE was analogue. Over the years, almost all hardware has been redesigned to take advantage of the formidable advances in electronics. This has been a continuous, ongoing process. Digitalisation was gradually introduced in the early 1980s, followed by application-specific integrated circuits (ASIC) in the mid-1980s. A major breakthrough came in 1986[1]. In the late 1980s and early 1990s, the evolution continued in small steps. A few original products have remained, however. Today, these last remaining products are being replaced. At the same time, almost all other hardware products that make up the basic AXE system are being rationalised.

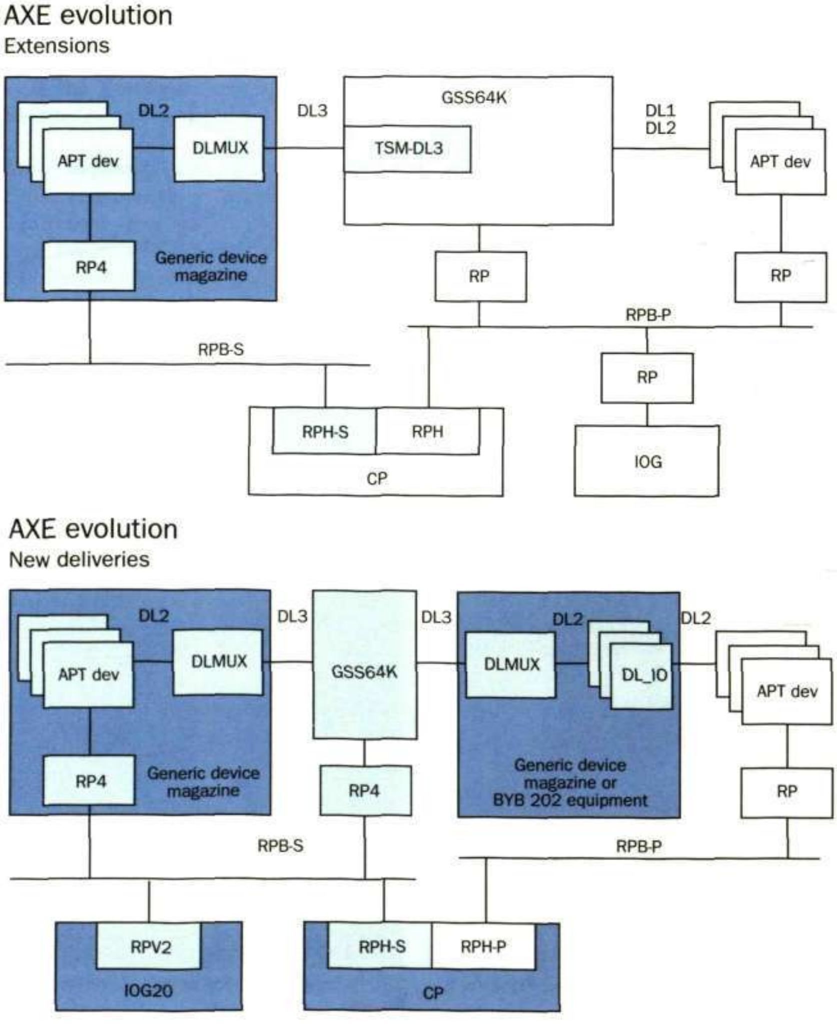


Figure 1
The figures show how the new interfaces are used for extensions and new deliveries.

Architecture

As the AXE system continues to evolve, system designers ensure that the very solid and proven system architecture is maintained. The fundamental principle of a central processor (CP) that controls regional processors (RP), which in turn control hardware services, has proved to be superior. Strict interfaces ensure that different system components can be developed independently. To ensure non-stop operation, all vital traffic and operation and maintenance (O&M) system products are built in duplicated structures.

In order to fully exploit the advantages of modern electronics, some fundamental system hardware interfaces are now being improved and extended. It goes without saying that compatibility is maintained in AXE.

Traditionally, a parallel bus, or a regional processor bus (RPB), has been used for communication between the central and regional processors. Now, however, in order to increase capacity (data transfer rate) and to decrease the need for interface hardware, a serial bus is being introduced alongside the existing RPB (Figure 1). The new RPB permits single-board regional processors to be housed in the same subrack as the devices they control, thus minimising hardware and cable interconnections between hardware devices.

In earlier generations of AXE, an extension module (EM) bus and cables were used to connect regional processors to application hardware (extension modules). In the new hardware design, however, most regional processors are located in the same subrack as the extension modules they control. By locating the regional processors in this way, designers have all but eliminated the EM bus, except in the backplane. The new location makes it much easier for operators to install and extend equipment.

The traditional AXE interface (called the digital link 2, DL2) between the group switch (GS) and its connected devices was at the 2 Mbit/s primary multiplexing pulse code modulation (PCM) level.

Now, a new high-speed interface is being

Box A Abbreviations

ALI	Alarm interface
ANSI	American National Standards Institute
ASIC	Application-specific integrated circuit
AST-DR-V3	Announcement service terminal version 3
ATM	Asynchronous transfer mode
BGA	Ball grid array
BM	Building module (1 BM=40.64 mm)
BSC	Base station controller
CANS	Code answer
CCD	Conference call device
CMOS	Complementary metal-oxide semiconductor
CP	Central processor
CSFSK	Code sender for FSK tones
CSK	Code sender for DTMF tones
CSR	Code sender/receiver
D-AMPS	Digital AMPS
DL2	Digital link interface 2
DL3	Digital link interface 3
DSP	Digital signal processor
DTMF	Dual-tone multifrequency
E0	64 kbit/s digital link
E1	2 Mbit/s digital link
ECP 303	Echo canceller in pool generation 3
ECP 404	Echo canceller in pool generation 4
EM	Extension module
EMB	Extension module bus
EMC	Electromagnetic compatibility
EMI	Electromagnetic interference
ETC5	Exchange terminal circuit generation 5
ETSI	European Telecommunications Standards Institute
FSK	Frequency shift keying
GDM	Generic device magazine (subrack)
GS	Group switch
GSM	Global system for mobile communication
GSS	Group switch subsystem
HLR	Home location register
IN	Intelligent network
I/O	Input/output
IOG11	I/O system 11
IOG20	I/O system 20
IP	Internet protocol
ISDN	Integrated services digital network
ITU-T	International Telecommunication Union - Telecommunications Standardization Sector
IWU	Interworking unit
KRD	Keyset receiver device
LED	Light-emitting diode
LUM	Line unit module
MSC	Mobile switching centre
MTBF	Mean time between failures
MW	Megaword
O&M	Operation and maintenance
PCM	Pulse code modulation
PDC	Pacific digital cellular
PROM	Programmable read-only memory
PSTN	Public switched telephone network
RAM	Random access memory
RMS	Remote measurement subsystem
ROM	Read-only memory
RP	Regional processor
RP4	Regional processor generation 4
RPB	Regional processor bus
RPD	Regional processor device
RPG	Regional processor with group switch interface
RPV	Regional processor connected to VME
SCP	Service control point
SCSI	Small computer system interface
SNT	Switching network terminal
SPM	Space switch module
STC	Signalling terminal central
STM	Synchronous transfer mode
STP	Signalling transfer point
T1	1.5 Mbit/s digital link
TCD	Trunk continuity check device
TSM	Time switch module
TSM-1	155 Mbit/s time switch module
VME	Versa Module Eurocard

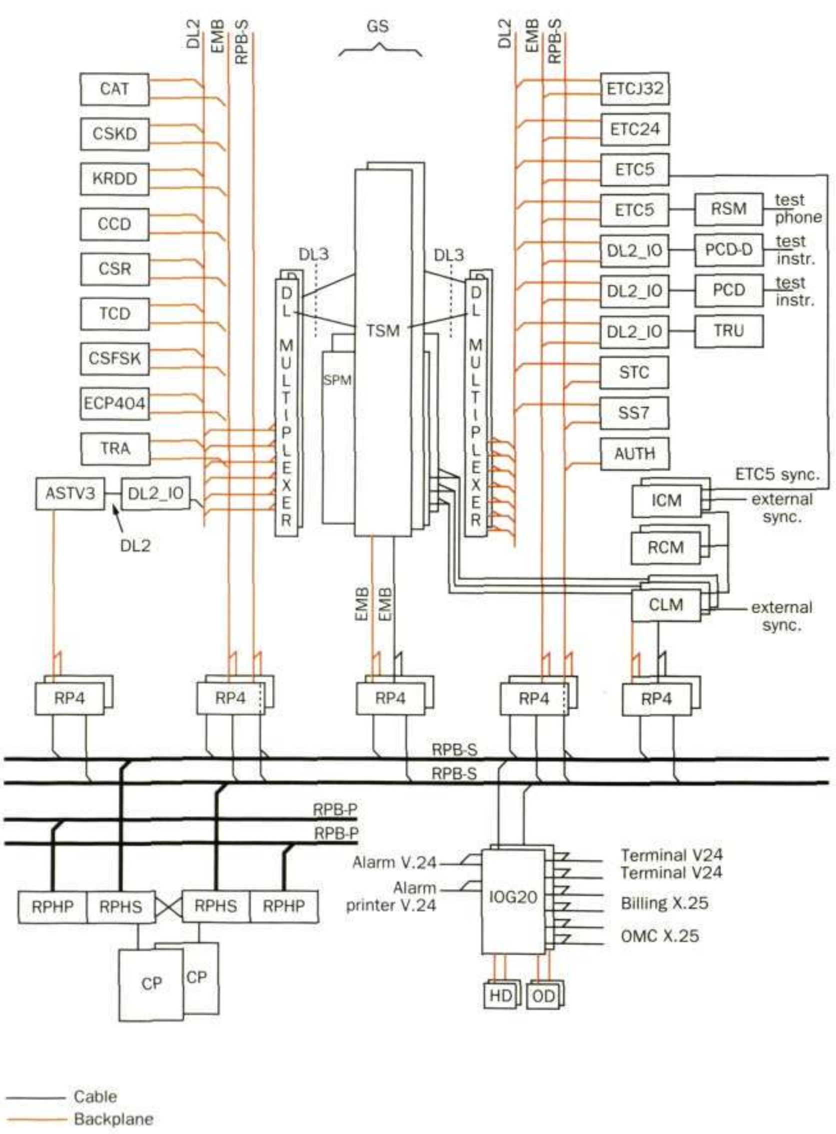


Figure 2
AXE hardware architecture using new hardware.

introduced at the third level in the basic PCM hierarchy. The interface, which is called DL3 (digital link 3), works at the 32 Mbit/s level (overhead excluded).

The introduction of the DL3 interface dramatically decreases group switch and device hardware. Equally important, it removes massive amounts of internal system cabling. The DL2 interface has been retained to ensure compatibility.

Each DL3 interface contains 16 multiplexed DL2 interfaces. In fact, the DL2s run in the backplane of the new device subracks, which means that only one sixteenth of the cabling is needed between the group switch and the devices that are connected to it.

Basic technology

In general, designers taking part in the AXE hardware evolution programme have used ASICs, high-performance microprocessors, digital signal processors (DSP) and faster interfaces to improve AXE hardware. ASICs were chosen where volumes of circuits are very high or where performance is critical. Commercial microprocessors, which are becoming commonplace for more and more applications, have also been integrated into the hardware. These changes allow designers to integrate commercial operating systems and software – especially at the regional processor level.

Also, inasmuch as the processing capacity of regional processors has kept pace with developments in general-purpose processor technology, the new AXE hardware requires fewer processors than were used before. This was another important factor in reducing the size of the exchange.

The most common type of processor in AXE systems today is the digital signal processor. DSPs, which are used in many kinds of application, are flexible platforms that may easily be programmed to provide new functions. Moreover, software at the DSP level may be sourced from other manufacturers, which allows designers to introduce new functionality with shorter time to market.

Today almost all AXE hardware uses a 3.3 V power supply. This change and the use of submicron technology (0.25-0.5 µm) have reduced power consumption to levels far below that of previous hardware generations.

Equipment practice

Owing to the introduction of high-speed interfaces and tougher requirements for electromagnetic compatibility (EMC), AXE hardware designers constructed a new equipment practice, called the BYB 501[2]. The BYB 501 has excellent EMC characteristics and fulfils Class B requirements with good margin. Compared with the BYB 202, whose cabinet shields against electromagnetic interference (EMI), the new equipment practice provides shielding at the subrack level. Note: the standard on which the BYB 501 is based uses the term subrack. However, in AXE terminology, the word magazine is often used.

The equipment practice supports multipoint and single-point earthing. The multi-

point earthing concept will be used in all new AXE deliveries. The equipment practice also supports several different sizes of board and cabinet. However, for use with AXE, two main board sizes are used: 115 x 175 mm, and 265 x 175 mm. The standard dimensions of the cabinet are as follows:
Height: 1800 mm
Width: 600 mm
Depth : 400 mm
Normally, no backplane cabling is needed on the subracks. Consequently, the cabinets may be placed back-to-back, giving the exchange a very small footprint and allowing a flexible cabinet arrangement against walls. The cabinets will also be delivered fully equipped, their hardware tested and cabled at the factory – a feature that greatly reduces installation time and other time-to-customer-related activities.

Group switch

The group switch[3] has been the subject of far-reaching rationalisation. For example, a configuration for 65,536 group switch ports is now contained in two cabinets (Figure 3). What is more, the new group switch consumes 95% less power than its predecessor. Nevertheless, the basic structure of the switch – that is, the time-space-time (T-S-T) switching architecture, the time switch, the space switch, the clock module, and system concepts such as the switching network terminal maintenance (SNT) and DL2 hardware interface – has been maintained, which facilitates hardware and software design and preserves compatibility.

In improving the group switch, designers made the following changes:

- A 32 Mbit/s DL3 interface replaces sixteen 2 Mbit/s DL2 switch interfaces.
- Four time switch module (TSM) functions are grouped onto one board, yielding 2,048 ports per board (Figure 4).
- A space switch module (SPM) function for 16,384 ports now fits on a single board (Figure 5).
- Switching equipment and the RPs that control the equipment are co-located in the same subrack.

These design changes gave rise to a switch subrack that contains eight TSM boards, providing a total of 16,384 switch ports; one SPM board; and four RP boards. Since the switch is duplicated, another plane is located in a second subrack with exactly the same configuration.

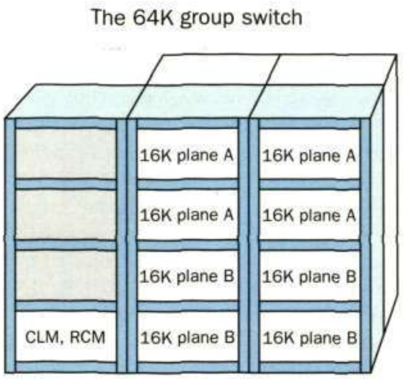


Figure 3
The new group switch in a 64K configuration, including synchronisation equipment.

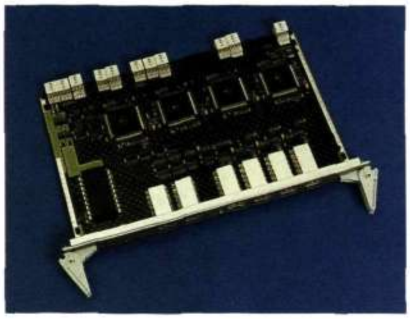


Figure 4
The new time switch module board, which contains 2,048 ports, replaces four BYB 202 subracks.

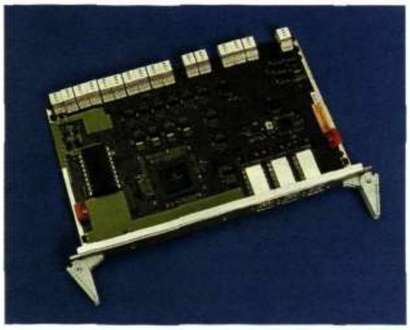


Figure 5
The space switch module board, which handles 16,384 ports, replaces one subrack in the BYB 202.

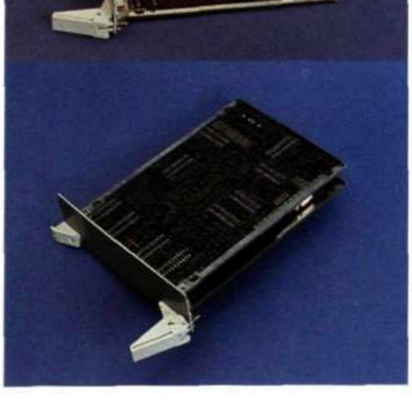
Some mobile systems employ subrate switching to handle bit rates below 64 kbit/s (8 kbit/s; 16 kbit/s; 24 kbit/s ... 64 kbit/s). In its maximum configuration, which has 4,096 ports, the subrate switch is housed in two small subracks: the A-plane is located in one subrack and the B-plane is located in the other.

As in earlier versions of the group switch subsystem (GSS), wideband (n x 64 kbit/s) is supported up to 2 Mbit/s.

The synchronisation equipment, which occupies another two small subracks, consists of:

Figure 6
The APZ 212 25 occupies only half a subrack in the BYB 501.

Figure 7
The RPG consists of two boards mounted together into one plug-in unit. Every interface is to the backplane.



- three clock modules;
- two highly accurate reference clock modules;
- two incoming clock reference boards (for connecting additional clock references);
- regional processors for controlling the synchronisation equipment.

Designers have also constructed a compact switch subrack for switching applications that require less than 4,096 ports. This subrack contains a 4,096-port switch, three clock boards for synchronisation, 1,024 ports for subrate data transfer, and regional processors for controlling the equipment. The two switch planes are co-located in one subrack.

The new group switch was designed to provide backward compatibility. Accordingly, a DL3-to-DL2 converter subrack has been developed. The converter connects hardware that uses a DL2 cable interface to a new switch that uses the DL3 interface. Another way of connecting an old switch to hardware that uses the new DL3 interface is to add an interface board (which supports a DL3 cable interface) in existing TSM64C subracks.

The new design concept also allows the GSS switch to be extended to up to 131,072 ports.

Central processor

Designers of the AXE central processor have always emphasised high processing capacity. This holds true even today. Nonetheless, while developing the next generation high-capacity central processor (APZ 212 30), AXE designers also produced a smaller, power-efficient processor (APZ 212 25) for switching applications that require moderate processing capacity.

The APZ 212 25 has a very small footprint (Figure 6) and consumes only 75 W of power. Designers reduced power consumption by replacing the 5 V supply voltage with 3.3 V, and by using 0.5 μm complementary metal-oxide semiconductor (CMOS) ASIC technology with ball grid array (BGA) packaging. The maximum memory capacity of the APZ 212 25 is 64 Megawords (MW), program store; and 256 MW, data store. Despite its small size, this computer processor is 1.5 to 1.7 times more powerful than its much larger predecessor, the APZ 212 11.

Although it was designed for use in the BYB 501, where it uses the serial RP bus, the APZ 212 25 is fully compatible with the parallel bus used in earlier versions of AXE switching equipment. In a minimum configuration, the APZ 212 25 may connect four of the new serial RP buses, controlling up to 128 regional processors. If more regional processors are required, or if parallel and serial RP buses must be used simultaneously, then extension subracks may be added that allow up to 512 regional processors to be connected.

Regional processors

A new regional processor, called the RPG (regional processor with group switch interface, Figure 7) has been introduced for applications that require high processing capacity. Most applications that previously ran on the regional processor device (RPD – Motorola 68020) have been transferred to the RPG, which has at least four times as much processing capacity as the RPD. The RPG is a single-board processor based on the general-purpose Motorola 68060 running at 50 MHz. On the same board is a communications processor (a Motorola 68360) for handling the switch interface and a 10 Mbit/s Ethernet interface. Although it may be used for any application that requires high processing power, the RPG will initially be used with the following applications:

- signalling system no. 7 – signalling terminal according to ANSI;
- signalling system no. 7 – signalling terminal according to ITU-T;
- signalling terminal central (STC) – signalling terminal for base stations and remote subscriber switches;
- transceiver handler – base station signalling in GSM;
- authentication in all mobile systems;
- integrated services digital network (ISDN) Internet access server (IP routing function).

In AXE, the RPG is the platform for handling packet switched data communication. With respect to traffic handling, these types of regional processor have a more independent role, relative to the CP, than traditional AXE regional processors.

A new version of the traditional regional processor, called RP4 (regional processor generation 4), is used for controlling extension modules. The RP4 is compatible with earlier versions of the regional processor. A prime benefit of the RP4 is that it is co-located in a subrack with the extension modules it controls. This design does away with

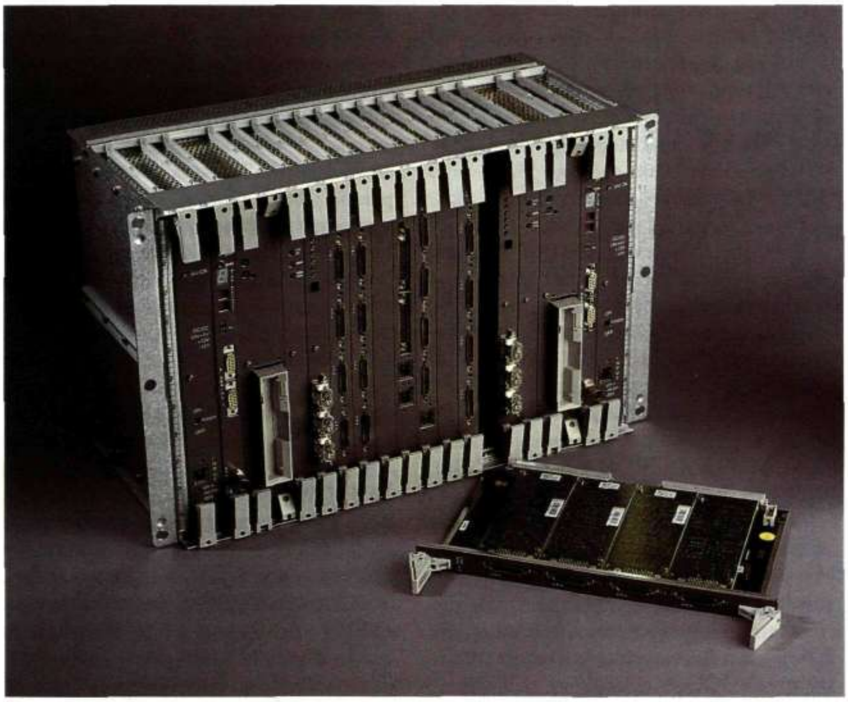


Figure 8
The IOG20C with A- and B-sides in one subrack. With its daughter boards, the line unit module can handle four different interfaces.

a large amount of cable, reduces size, and simplifies equipment handling considerably.

Earlier versions of the central processor may not be connected to new hardware without first modifying their side of the RP bus interface.

The regional processor bus interface VME (RPV) is a conversion product for the connection from the CP to the Versa Module Eurocard-based IOG20, through the RP bus. There are two RPVs: the first, known simply as RPV, is used for the parallel bus connection; the second, called the RPV2, is used for the serial bus connection.

IOG20, the AXE I/O system

A duplicated input/output (I/O) system, known as the IOG20, handles data transport to and from an AXE exchange. Communication to and from the AXE I/O system may be broken down into customer administration and element handling.

The IOG20 is much smaller than the IOG11 – the previous generation I/O system. For example, whereas the IOG11 fills a whole cabinet in the BYB 202 equipment practice, the IOG20 fits into a single subrack. Moreover, the IOG20 outperforms the IOG11 by as much as four to five times but consumes only one third as much power.

The new I/O system, whose design is characterised by modern technology and greater integration, contains relatively few printed circuit board types (seven instead of 25). In taking steps to make the system open to commercially available components, designers used the industry standards Versa Module Eurocard (VME) bus, Ethernet, and small computer system interface (SCSI). Similarly, they implemented Ethernet for connections between nodes and as a line interface. The IOG20 is currently available in three configurations:

- IOG20 – a fully compatible version of the twin-subrack configuration with an interface to a parallel RP bus;
- IOG20B – a twin-subrack version with one node in each subrack (maximum configuration);
- IOG20C – a single-subrack version with two nodes (minimum configuration).

The IOG20B and the IOG20C are designed to operate with the new serial RP bus. The IOG20C is probably the most compact and powerful I/O system ever produced for telecommunications applications (Figure 8).

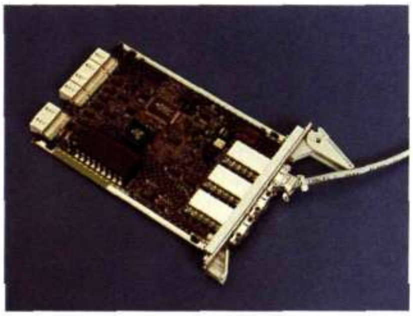


Figure 9
The 32-channel E1 interface is now made on one small board.

In its maximum configuration, the IOG20 stores data on three duplicated 3.5 inch/4 Gbyte hard disks and one duplicated 3.5 inch/640 Mbyte magneto-optical disk. In the compact version, the IOG20 stores data on one duplicated 3.5 inch/4 Gbyte hard disk and one duplicated 3.5 inch/640 Mbyte magneto-optical disk.

To connect data communication interfaces to the I/O system, the twin-subrack version may contain up to four duplicated line unit module (LUM) boards. Likewise, the compact version may contain up to three duplicated LUM boards. A LUM board consists of a main board and as many as four independent line module daughter boards for almost any type of line interface, including V.24, V.28, V.35, V.36, X.21, G.703 E0, G.703 E1, and Ethernet.

An alarm interface (ALI) function consists of two boards: one for supervising fans and external alarm input/output, and another for displaying alarms.

In terms of software and applications, the IOG20 is fully compatible with its predecessor, the IOG11.

Connecting hardware to the group switch

Hardware is connected to the switch either by a trunk (for example, exchange terminals) or by means of pooling (for example, of echo cancellers). In AXE, only exchange terminals and some signalling terminals are connected by trunks. All other equipment is connected in pool, which heightens reliability, flexibility, economy, and maintainability.

Announcement machines

Designers have also developed a new generation of system-integrated announcement machines – AST-DR-V3. The new machines are substantially smaller than their predecessors, but have more capacity for speech storage and for a larger number of dual-tone multifrequency (DTMF) receivers. The announcement machines are available in different sizes (configurations). The largest machine has capacity for 256 DTMF receivers, up to eight hours of stored speech, and provisions for backing up speech on the hard disk.

The smallest configuration has capacity for 32 DTMF receivers and two hours of stored speech. Depending on how often stored phrases are changed, speech may be stored in either random access memory (RAM) or in read-only memory (ROM) on memory boards that support up to one hour of speech per board. The high-capacity announcement machine occupies one and a half subracks: one subrack for the control subrack that contains the DTMF receivers, and half a subrack for the memory boards and hard disk. The smallest machine occupies only half a subrack. As many as 20 systems may be run in parallel, providing a total of 5,120 ports. The systems may also be used in large intelligent network (IN) nodes or in other service-providing functions. The AST-DR-V3 forms a powerful voice-response system that may be used as a base product for the future development of such applications as voice or fax mail, cashless calling, and the virtual telephone.

Exchange terminals

AXE supports every kind of trunk interface that has been incorporated into the new equipment practice. By integrating all functionality into one ASIC, designers were able to fit the 32-channel E1 (2 Mbit/s digital link) interface onto one small board (Figure 9). New versions of the 24-channel T1 and the Japanese 32-channel interface have also been designed.

In time, STM-1 (155 Mbit/s synchronous transfer mode) terminations will be designed in AXE for each relevant standard. Once they become available, the terminations will greatly reduce (possibly completely eliminating) operator requirements for transmission equipment and generally simplify system handling.

DSP platform

To date, much of the telephony devices in AXE – conference call device (CCD); trunk continuity check device (TCD); code sender/receiver (CSR) for R1, R2, and no. 5 code; code sender for DTMF (CSK); code sender for FSK tones (CSFSK); code answer (CANS); keyset receiver device (KRD); and several maintenance functions – is delivered from separate subracks that range in size from 3 to 12 building modules (BM) in the BYB 202 equipment practice. Nonetheless, designers have developed a new digital signal processor platform board that can be programmed to provide the functionality of any one of these applications. Initially the boards will be programmed at the factory. In a second step, operators will be able to change the onboard software from the AXE system, giving them tremendous flexibility and excellent means with which to handle redundancy and spare parts. Should a fault occur on a board that provides the functions described above, then an operator can remotely activate an unprogrammed standby board by command, taking it into operation. This feature will simplify maintenance and reduce operating costs.

Echo cancellers

The ECP 303[4] has been replaced with a new echo canceller, called the ECP 404. The ECP 404 has a capacity of 512 channels per subrack, which is twice the capacity of the ECP 303. As with its predecessor, the ECP 404 is connected to the group switch by means of pooling.

Transcoders

Transcoders, which are included in all digital mobile telephony systems, are used for speech compression – from 64 kbit/s to bit rates below 16 kbit/s in the downlink direction, and from bit rates below 16 kbit/s to 64 kbit/s in the uplink direction. Limited bandwidth in the air interface, which is a major challenge of mobile telephony systems, requires that speech be compressed before it can be sent over the interface.

As with all other devices, the transcoders are connected to the switch and supervised by AXE. The capacity of each board differs depending on the mobile standard for which it has been deployed (for example, D-AMPS, GSM or PDC). Each standard uses unique algorithms that require different processing capacity.

By employing the latest techniques in digital signal processing, designers have been able to more than double the capacity of the transcoder boards. This represents a significant achievement since the transcoders make up a large part of mobile exchanges.

Data transmission interworking unit

Interworking functions are needed to provide the digital transmission of data services within mobile networks as well as between them and other networks. This is because protocols for the standard network use analogue tones, which are not suitable for transmission over the radio interface to mobile terminals.

An interworking unit (IWU) extracts analogue information received from a public switched telephone network (PSTN) modem and sends it to the mobile terminal by means of a digital protocol. The opposite function is performed for signals from the mobile terminal to the PSTN modem. The interworking function is implemented in a 7.5 BM subrack, which can handle up to 32 simultaneous data or fax calls. The subrack fits in the new equipment practice.

This function was previously used as a stand-alone product in the Ericsson GSM system, but it will now be integrated into the system and supervised by AXE.

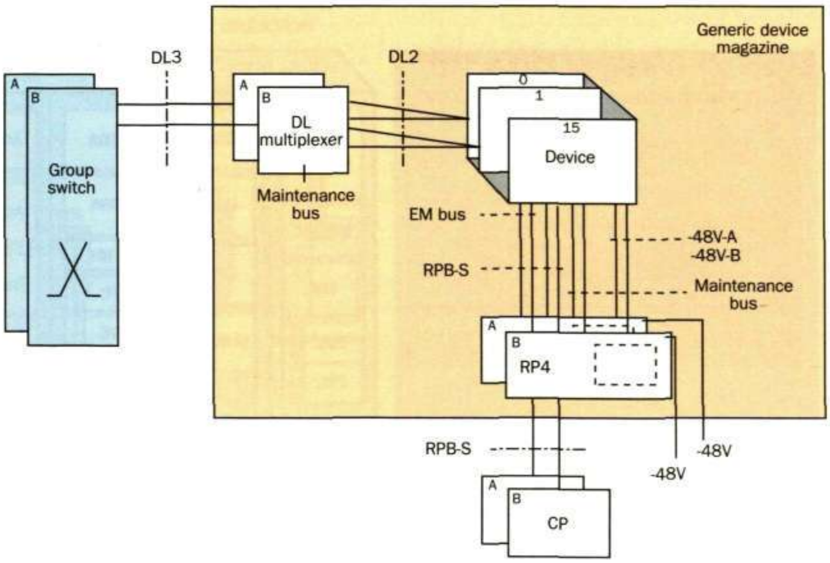


Figure 10
Hardware architecture of the generic device magazine (subrack).

Remote measurement subsystem

The AXE remote measurement subsystem (RMS) measures characteristics and transmission quality between telephony exchanges. It performs digital, analogue and signalling tests. To date, this function – which occupies one subrack in the BYB 202 equipment practice – has been used solely in transit exchanges. In the new rationalised version of the RMS, the function will be constructed from powerful DSPs on a single board.

Subracks for switch-connected hardware

Nearly all telephony devices in AXE are now single-board applications, giving rise to the development of a new concept for generic device magazines (subracks). The concept is based on a subrack with 16 slots for device boards. From the backplane, the boards are connected to a duplicated group-switch interface, a duplicated RP bus interface, a duplicated EM bus interface, and a maintenance bus. Moreover, each board is given an EM bus address and supplied with duplicated –48 V (Figure 10).

Besides the 16 device boards, two multiplexers on the front of the board are connect-

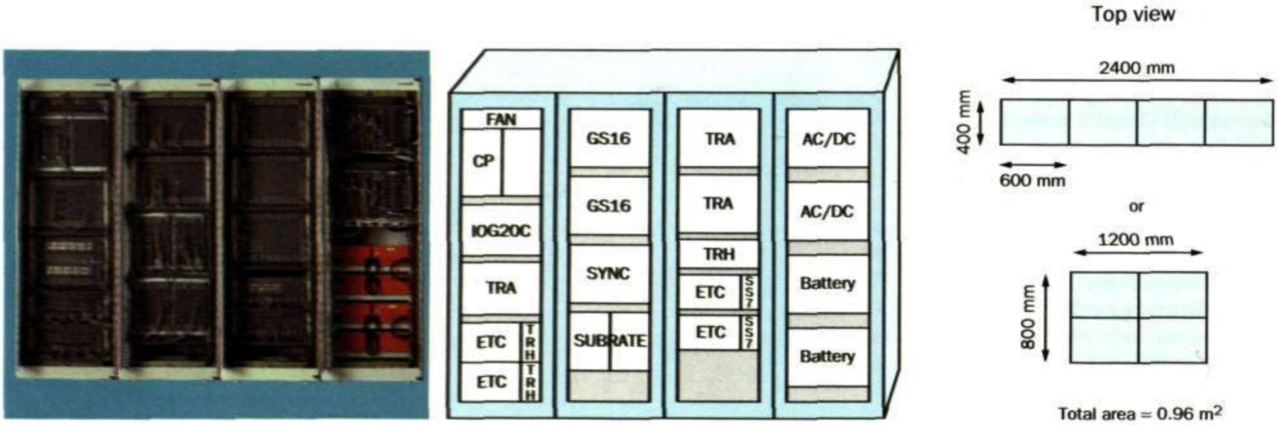


Figure 11
A powerful BSC with capacity for more than 300 transceivers including power supply and batteries. The configuration is similar to a small local exchange (subscriber stage excluded) or to a small MSC.

ed to the switch by means of a DL3 interface. The multiplexers split the DL3 interface into 16 DL2 connections in the backplane, one connection per board. The other interfaces are connected to a pair of regional processors, one at each end of the subrack. Moreover, since the RP bus is also distributed in the backplane, it is possible to mix – in the same subrack – boards that use the EM bus with boards that use the RP bus. Because some applications require a large board size while others require a small one, two versions of the generic subrack have been constructed.

Product identification

A new function has been introduced for checking the hardware of an AXE exchange. Each board contains a small programmable read-only memory (PROM) that stores the unique serial number, product number, revision state and manufacturing date of that board. Operators may fetch and read this information by command (on site or remotely), which enables them to check:

- hardware when replacing faulty units;
- revision states when upgrading hardware;
- for compatibility when introducing new software.

Visual indication

Most boards in the new AXE system contain a light-emitting diode (LED) on their front. The LEDs help operators in various maintenance situations; for example, when locating boards that need to be removed for repair or for upgrade.

The indicator does not necessarily indicate that a board is faulty. Instead, it indicates whether or not a board may be removed without disturbing traffic.

Power supply

An optional battery backup and modular power supply are offered for exchanges whose power consumption is below 6 kW. The batteries and rectifiers are housed together with the switching equipment. A

Figure 12
A complete AXE exchange in one cabinet. This configuration can be used for any of the following applications: HLR, STP, SCP, or BSC (more than 120 transceivers).

single cabinet with battery backup can provide a 3 kW power supply for nearly two hours. A 6 kW power supply is sufficient to operate approximately 15 AXE cabinets, which more or less corresponds to a high-end mobile switching centre (MSC) in a mobile telephony system.

Most hardware in the BYB 501 equipment practice is fed with a redundant power supply to each subrack through two branches of –48 V. Each branch of power is filtered and distributed to the subrack backplane, from which each board is supplied through a double-diode configuration. This arrangement increases reliability, since the subrack continues to work even if one branch of the power supply is lost.

The power distribution system also allows boards to be inserted into a subrack that is in service, which greatly simplifies procedures when boards in the subracks must be replaced.

Results

At the system level, recent developments in the AXE hardware evolution programme have reduced the number of board types used in AXE and made them smaller and much more power-efficient. For example, it was possible to reduce the size of a base station controller (BSC) for a GSM configuration that supports approximately 300 transceivers by nearly 90% – including power supply, battery backup (Figure 11) and transcoders. Today, power consumption for a complete base station controller of this type is less than 1500 W. Moreover, when the BSC is delivered for installation, very little additional work is required, since the cabinets are equipped with subracks and internal system cables at the factory.

For the first time, a complete AXE exchange fits into a single cabinet (Figure 12). This configuration can be used for a home location register (HLR), signalling transfer point (STP), service control point (SCP), or for a base station controller application.

Conclusion

The AXE hardware evolution programme has successfully reduced the size of hardware by between 70% and 90%; cabling in the exchange has been reduced by 90%, and power has been reduced by 75%. Therefore, operators can expect that the time and resources needed for installing the hardware will also decrease by between 70% and 90%. The delivery of fully equipped and tested exchanges will further simplify installation.

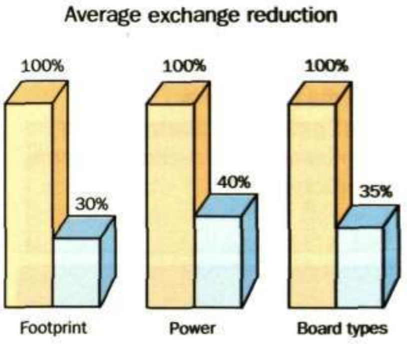


Figure 13
Average reduction in footprint, power, and board type for an AXE exchange.

The following aspects contribute towards reducing operator costs for running the new exchanges:

- smaller footprints require less floor space (reduced overhead);
- costs of power (batteries, rectifiers and kW) and cooling are reduced (reduced overhead);
- fewer spare parts are needed (smaller facilities, smaller stores);
- operations have been simplified (less staff, less training);
- less hardware implies that the mean time between failures (MTBF) increases, while the repair time decreases – in that way, the total down time, due to hardware failures, will decrease;
- pooled devices;
- programmable platforms.

The hardware evolution described in this article represents only a first step in Ericsson's AXE hardware evolution programme. In subsequent phases of the programme, AXE will be migrated towards an open hardware architecture that supports datacom functionality, asynchronous transfer mode (ATM) switching, high-speed interfaces and multiprocessor configurations.

References

1 Hägg, U., Persson, K.: New hardware in AXE 10. Ericsson Review 63(1986):2, pp 86-92.
2 Stockman, B. and Wallers, A.: BYB 501 metric equipment practice. Ericsson Review 74(97):2, pp. 62-67.
3 Hansson, U., Paone, T.: The group switch subsystem – an enhanced competitive group switch. Ericsson Review 74(1997):2, pp 68-73.
4 Eriksson, A., Eriksson, G., Karlsen, J., Roxström, A., Vallon Hulth, T.: Ericsson echo cancellors – a key to improved speech quality. Ericsson Review 73(1996):1, pp 25-33.

The BYB 501 metric equipment practice

Bo Stockman and Arne Wallers

Ericsson have developed a new cabinet-based equipment practice – called the BYB 501 – that is intended for use at public telecom sites and at customer sites in private networks. Applications for the BYB 501 range from small access units to large switches and powerful processors.

The authors describe the BYB 501 and its most prominent features compared with existing equipment practices: compliance with metric standards from the IEC and ETSI, and excellent EMC characteristics.

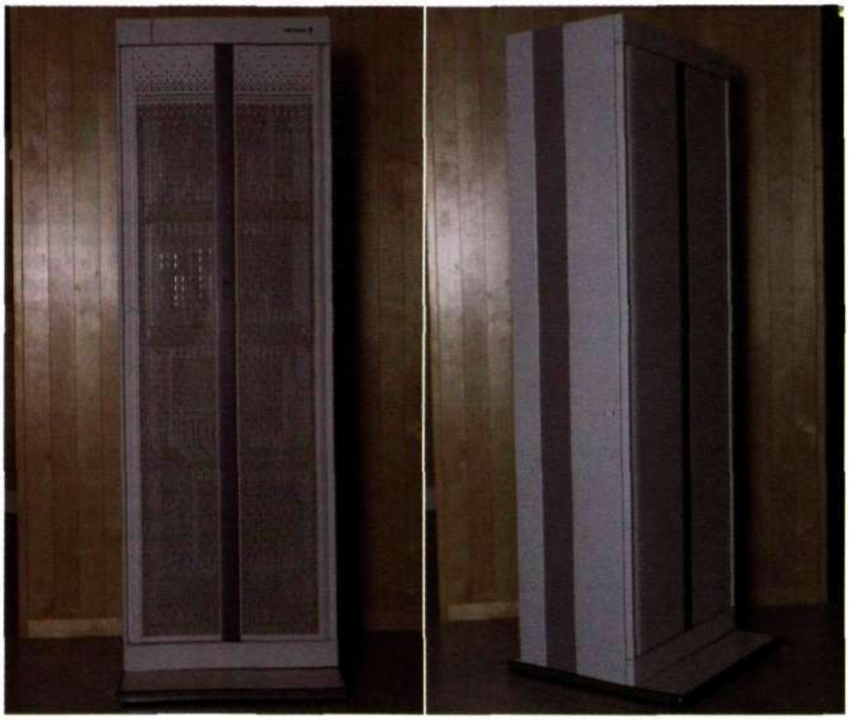


Figure 1
The BYB 501 equipment practice fulfils the metric IEC and ETSI standards and has excellent EMC characteristics.

Not only does an equipment practice make up the literal framework of a telecommunications system, but it provides the framework and basis for the physical properties of the system as well, including such obvious features as dimensions, weight, and cooling performance. Less obvious, perhaps, are the environmental properties of an equipment practice that are associated with production, installation, operation, and scrapping. These properties (electromagnetic, seismic, chemical, etc) are normally specified in international, regional or national standards.

The equipment practice is a synthesis of these external requirements and of requirements that were derived from the system design process. Mechanical interfaces, for example, must be harmonised with corresponding system interfaces.

Main characteristics

Standards

The BYB 501 equipment practice complies with metric standards according to the International Electrotechnical Commission (IEC) and the European Telecommunications Standards Institute (ETSI), Box B. The Bellcore standards have also been considered.

Compliance with internationally recognised standards is the foundation of the open mechanical interfaces of the BYB 501. The use of an open mechanical interface offers important advantages:

- different subsystems, including 19-inch subsystems, may be mixed and matched;
- short time to customer;
- sourced products, based on standards such as Versa Module Eurocard (VME) or CompactPCI, are easily accommodated;
- Ericsson subracks may be mounted in existing 19-inch racks or cabinets at the customer's site with little or no adaptation.

Details of the BYB 501 cabinets, subracks and plug-in units (PIU) are provided at the end of this article.

Electromagnetic compatibility

One of the main objectives in developing the BYB 501 was to create a system with excellent electromagnetic compatibility (EMC).

EMC, which is an area of ever-increasing importance, is defined as the ability of equipment to function satisfactorily in its electromagnetic environment without

Box A
Abbreviations

EMC	Electromagnetic compatibility
EMI	Electromagnetic interference
ESD	Electrostatic discharge
ETSI	European Telecommunications Standards Institute
IEC	International Electrotechnical Commission
PIU	Plug-in unit
TS-HOD	Two-step high-ohmic distribution
VME	Versa Module Eurocard

introducing intolerable electromagnetic disturbances in that environment. EMC-related requirements and legislation will have a great influence on current and future hardware designs.

A fundamental principle of hardware design calls for interference to be neutralised as close to its source as possible. By adhering to this principle, electrical designers are able to avoid most EMC-related problems. By means of shielding subracks, cables, plug-in units and components, the BYB 501 equipment practice further improves a system's EMC characteristics.

The connection of cable shields at the front of a plug-in unit is extremely important to the system's EMC performance. To keep the cable from acting as an antenna for radiation to and from the system, maximum impedance is less than 1 ohm at 30 MHz (or 5 nH inductance). Similar requirements apply to the earthing of filters for unshielded cables.

No requirements for shielding have been imposed on cabinets. Instead, all shielding functions have been allocated to the subrack level. There are two main reasons for doing so:

- subracks installed in customer cabinets meet requirements for EMC without the need of further protection;
- cabinet doors can be opened during maintenance work without causing the stipulated emission limits to be exceeded.

The shielding concept is illustrated in Figure 2.

The subrack shielding efficiency is at least 20 dB for frequencies up to 5 GHz. Gaskets that block electromagnetic interference (EMI) are placed between the board front panels, thereby maintaining a closed subrack shield. Shielding covers are provided for individual components at the board level.

Verifying measurements taken on delivered products show that the design fulfils, with ample margins, EMC requirements according to Class B.

Electrostatic discharge (ESD), which is the transfer of electric charge, represents an EMC-related problem of special concern. Unless this phenomenon is properly dealt with, it may cause considerable maintenance problems. Thus, the following measures were taken or have been prescribed for working with the BYB 501:

- as basic protection, all mechanical parts are low-inductively connected to earth;
- electrostatic discharges are diverted through high-ohmic ESD wrist-straps, which are worn by maintenance personnel and connected to the mechanical structure of the cabinets.

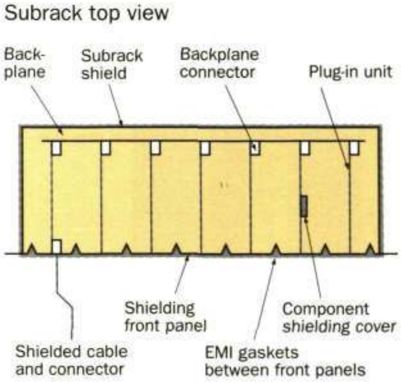


Figure 2
The shielding concept. Shielding functions are allocated to the subrack level.

Signal transmission

Multilayer-printed-board technology for high-speed communication and high signal density has been applied. The printed boards minimise supply-voltage ripple and make for low radiated emission.

Shielded front cable connectors were developed for power, twisted pair and coaxial cables. Each of these connectors has been verified in terms of EMC performance.

For signal transmission, 2 mm grid metric connectors are used between plug-in units and the backplane. The connectors are available as standard low-cost models or as high-performance shielded models with controlled characteristic impedance.

Box B
The BYB 501 equipment practice complies with the following international standards

IEC 917-2-1
Interface coordination dimensions for the 25 mm equipment practice. Detail specification of cabinets and racks.
IEC 917-2-2
Detail specification of the 25 mm equipment practice. Dimensions of subracks, chassis, backplanes, front panels and plug-in units.
ETSI ETS 300 119-2
Engineering requirements for racks and cabinets.

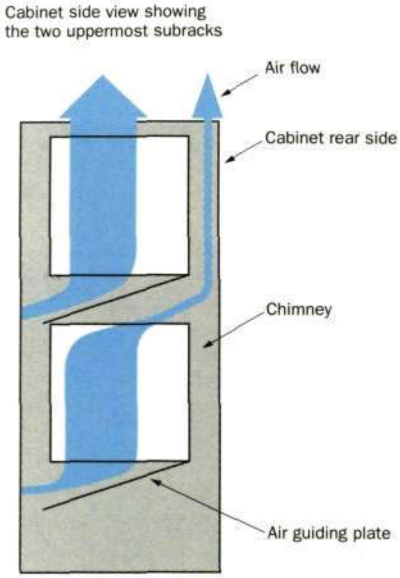


Figure 3
The principle of parallel cooling. Each subrack has a separate inlet for air at room temperature.

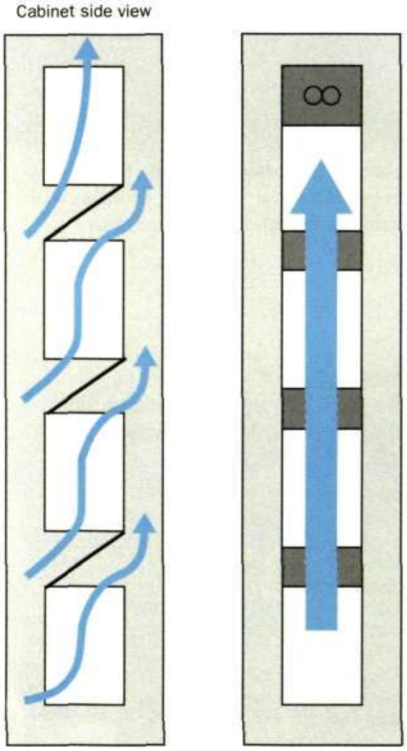


Figure 4
Examples of cabinet cooling. Left: parallel, natural-convection cooling in a single-depth cabinet. Right: serial fan cooling.

Cooling

Ordinarily, the cooling capacity at a customer's premises determines how much power may be dissipated in a cabinet. Some common maximum values are 500 W for a single-depth cabinet and 1000 W for a double-depth cabinet. However, because higher levels are accepted in some applications, engineers designed the BYB 501 equipment practice to accommodate a maximum cooling capacity that is considerably higher than these values.

Natural convection cooling is the preferred choice for all applications. Forced-convection cooling (fan cooling) is allowed only when the technical solutions required to provide natural convection are unreasonably complicated.

Either parallel or serial cooling may be used in the BYB 501. The principle of parallel cooling has certain advantages, Figure 3. For example, since each subrack has a separate inlet for air at room temperature, the parallel cooling method:

- facilitates a higher plug-in unit heat load;
- permits each subrack in a cabinet to be configured independently;
- greatly improves fire resistance, by preventing fire from spreading from one subrack to another.

Parallel cooling is the preferred cooling method to be used for natural convection. However, for fan cooling, the serial cooling method is preferred.

Figure 4 shows examples of cooling techniques used in single- and double-depth cabinets. Several configurations are possible, including those that combine natural convection and fan cooling in the same cabinet.

Power distribution and earthing

For power distribution, the BYB 501 normally employs a technique known as two-step high-ohmic distribution (TS-HOD), which minimises the effects of short-duration voltage transients (spikes) produced by short-circuit currents.

Cabinets and subracks are designed to sustain external fault currents of up to 3000 A. The difference in earth potential between subracks is less than 3.0 V; within subracks the difference is less than 0.2 V.

Environmental management and production

In recent years, environmental management – with emphasis on scrapping, recycling and environmental load – has become an important issue at Ericsson. Thus, when selecting the materials and assembly methods that went into making the BYB 501, designers were careful to ensure that they fulfilled existing as well as anticipated requirements in this field. What is more, the designers specifically chose materials and production methods that could be obtained and used to produce the mechanical structure in practically any part of the world.

Handling and installation

One of the driving forces behind the development of the BYB 501 was the need to shorten installation time, which is part of the overall requirement for short time to customer. This requirement was met in two ways:

- Mechanical endurance characteristics of the BYB 501 permit fully equipped and tested cabinets to be delivered to the installation site. Earthquake resistance is maintained without requiring additional strengthening elements.
- The time needed to install cables was reduced substantially. Thanks to the design of the BYB 501, all external cabling may be installed in the cable distribution system before the cabinets are delivered to the site. Prefabricated cable sets are readily plugged into the cabinets. The cabinet design allows external cables to be installed without feeding them through any holes in the rack.

Cable distribution system

The Ericsson cable distribution system is mounted either above the equipment in cable trays, or below the equipment in raised floors. Moreover, different cable types may be separated:

- Optical fibre cables, which are usually sensitive to mechanical stress, may require separate routes.
- AC power cables may be separated from signalling cables in order to meet customer safety requirements.

Existing customer cable distribution systems may be used instead of the one supplied by Ericsson.

Cabling in cabinets

Figure 5 shows cabling routes in and between adjacent cabinets.

Cables between cabinets are usually run in vertical cable ducts via the mechanical cabling structure above or below the cabinets. For high-performance systems, short, straight cable paths may be arranged between adjacent side-by-side or back-to-back cabinets.

Ordinarily, cables are connected directly to the front panel of plug-in units. However, for system parts with extensive cabling, two additional methods are used to connect cables from the back. One method, which mounts the connectors directly on the backplane, is intended mainly for internal cabling between subracks within the same subsystem (Figure 6, left). The second method, which uses the connection plug-in units that are located at the back of the main subrack (Figure 6, right), is best suited for interfaces that require different types of connector.

Other methods apply when the cabinets and other systems/subsystems are electrically or optically interconnected.

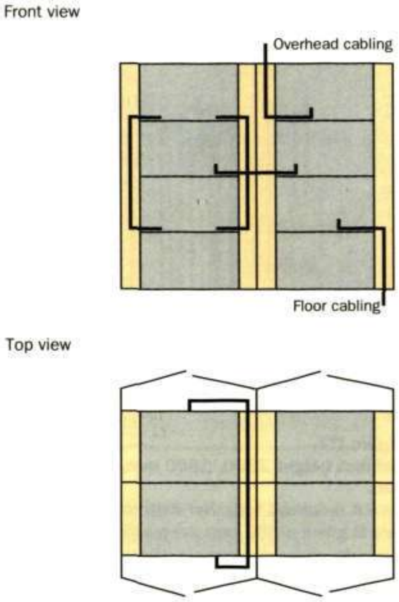


Figure 5
Cable pathways in the cabinets: vertical cable ducts for access to the cable distribution system; short, straight cable paths between adjacent cabinets.

🟨 Vertical cable ducts

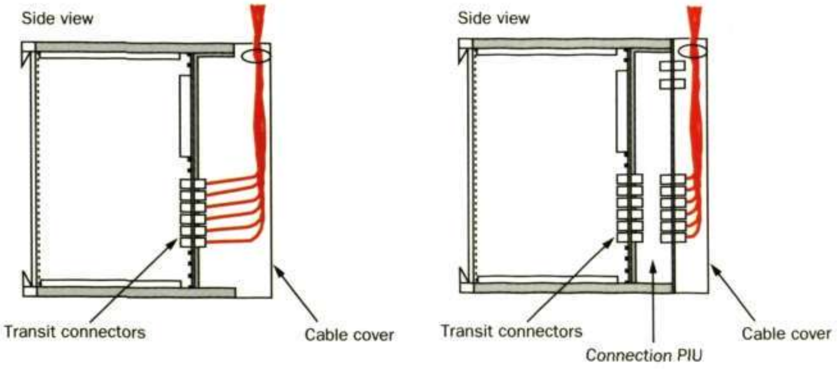


Figure 6
Left: subrack with direct backplane connections. Right: subrack with connection plug-in units.

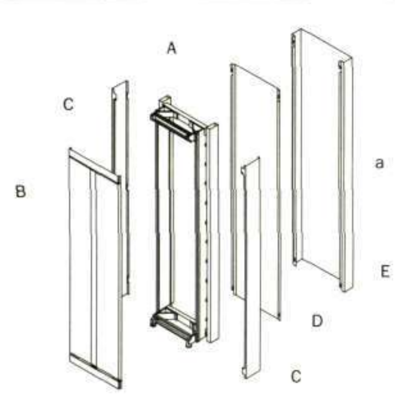


Figure C1
Cabinet height 2200/1800 mm, width 600 mm.
With D the depth is 300 mm; with E, 400 mm.

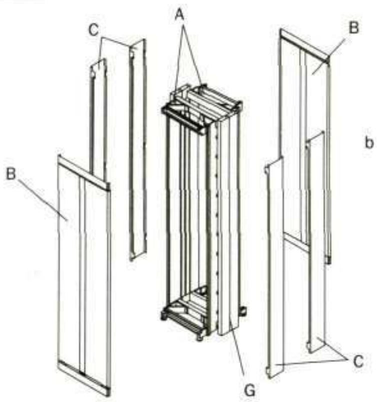


Figure C2
Cabinet height 2200/1800 mm, width 600 mm.
Two A mounted together with four C and two B gives a 600 mm deep cabinet.

Box C
Cabinet

Basic parts in cabinet make-up

A	Cabinet frame
B	Door set
C	Side cover
D	Back cover 300 mm
E	Back cover 400 mm
F	Side cover 800 mm
G	Back-to-back mounting set

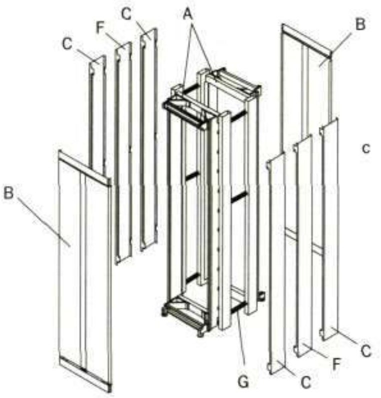


Figure C3
Cabinet height 2200/1800 mm, width 600 mm.
Two A mounted together with four C, two B, two F and one G, gives an 800 mm deep cabinet.

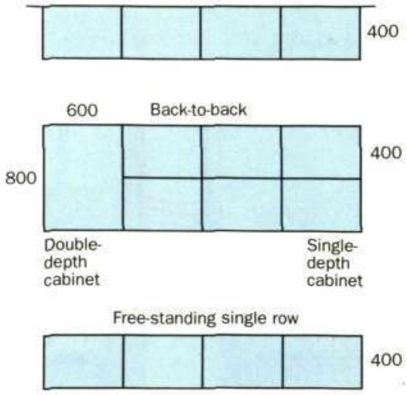


Figure 7
Floor plan of a telecom centre showing possible arrangements of cabinets.

The BYB 501 was designed to make operation and maintenance procedures as simple and self-explanatory as possible. Every unit that may have to be disconnected and replaced is labelled with a product identity that includes product type and number. The label information is presented in both alphanumeric and bar-code form.

Dimensions

Cabinets

Box C shows different ways of combining the elements of the cabinet. Only the outermost cabinets of a multi-cabinet system are equipped with side covers. Back covers are used only for single in-line configurations.

The normal aperture width is 450 mm. This dimension, which gives ample space for vertical cabling, is the metric standard equivalent of the 19-inch standard. The mounting plates of the cabinet may also be mounted to form an aperture width of 500 mm, which complies with European telecommunications standards (ETS).

Figure 7 shows a mixed configuration of single-depth back-to-back and double-depth cabinets.

Subracks

A range of subracks is available for different system needs. The standard subrack, which is 450 mm wide, houses up to 21 plug-in units with 20 mm spacing. For small subsystems, such as the new APZ 212 25 compact central processor in AXE[6], a special half-width subrack provides complete redundancy without wasting valuable space.

The subracks have a height modularity of 150 mm; that is, they are 150, 300, or 450 mm high. Different kinds of plug-in unit may be mixed in the same subrack.

In accordance with the IEC 917-2-2 standard, the plug-in unit widths are in steps of 5 mm. All subracks are shielded to meet requirements for EMC.

Besides the normal subracks for the housing of plug-in units, fan-unit subracks, air guides and other accessories are also available. Note: the term subrack is used in international standardisation; however, AXE documentation often uses the term magazine.

Plug-in units

Different types of plug-in unit (PIU) may be inserted into the subracks. The most common type consists of a single printed board assembly (ROJ 207+). Several types of composite plug-in unit have also been defined (BFB 301+) for combinations of two or more boards in a single unit. The same size applies to both ROJ and BFB plug-in units, Box D.

Conclusion

Ericsson's new cabinet-based equipment practice – the BYB 501 – which complies with the metric standards from the IEC and ETSI, offers important advantages compared with other equipment practices. For example, the BYB 501 easily accommodates other, standardised products; Ericsson subracks may be mounted in existing cabinets at the customer's premises; its excellent EMC characteristics provide superior protection of subracks, cables, plug-in units, and components.

Moreover, thanks to its modern, well-planned design, the BYB 501 enables operators to cut installation time substantially. All external cabling may be installed in the cable distribution system before the fully equipped and tested cabinets arrive. When the cabinets are delivered, prefabricated cable sets may readily be plugged into them.

The materials and assembly methods used in the BYB 501 were carefully selected with an eye to fulfilling existing as well as anticipated environmental requirements.

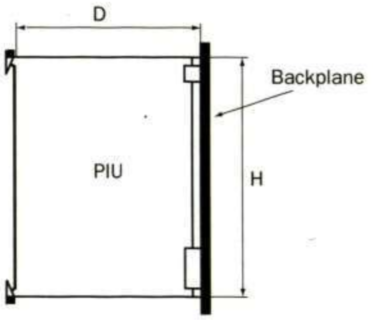
Box D

Plug-in unit

The following PIU dimensions have been defined for BYB 501 (in mm):

ROJ 204	
H=115	D=175 mm
ROJ 207	
H=265	D=175 mm
ROJ 208	
H=265	D=225 mm
ROJ 212	
H=265	D=300 mm
ROJ 237	
H=415	D=175 mm

PBA dimensions



References

1 IEC 917-2-1: Interface coordination dimensions for the 25 mm equipment practice. Detail specification of cabinets and racks.
2 IEC 917-2-2: Detail specification of the 25 mm equipment practice. Dimensions of subracks, chassis, backplanes, front panels and plug-in units.
3 ETSI ETS 300 119-1: Introduction and terminology.
4 ETSI ETS 300 119-2: Engineering requirements for racks and cabinets.
5 ETSI ETS 300 119-3: Engineering requirements for miscellaneous racks.
6 Lundqvist, T., Hägg, U., AXE hardware evolution. Ericsson Review 74 (1997):2, pp 52-61.

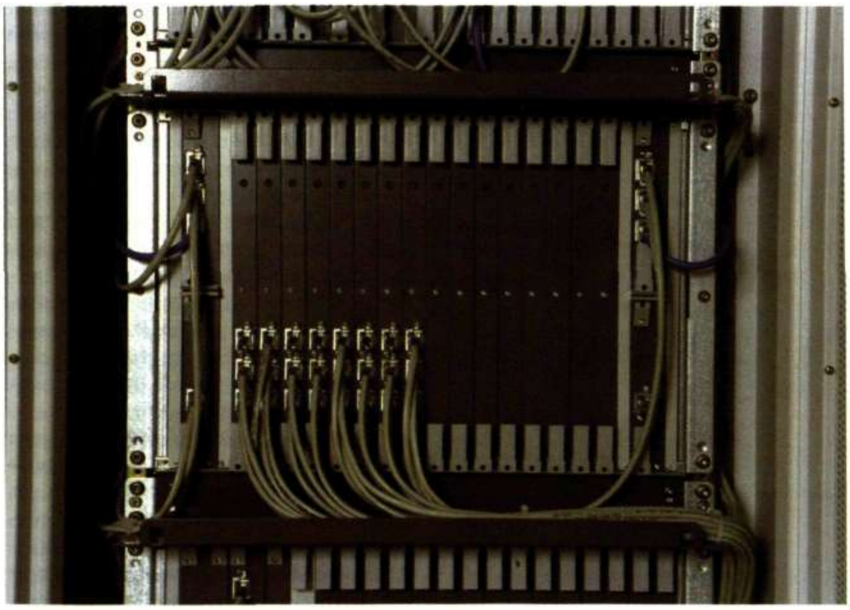


Figure 8
BYB 501 subrack with a mix of full-size and half-size plug-in units.

New hardware in AXE – The group switch

Ulf Hansson and Terenzio Paone

Within the framework of the most recent AXE hardware development programme, products that relate to the group switch subsystem (GSS) have been rationalised extensively. The main objectives of this work – which began in the spring of 1996 and is scheduled to end summer 1997 – were: to radically decrease the footprint of the group switch subsystem; to radically decrease power dissipation; to reduce the number of hardware board types in the group switch; and to provide a more efficient hardware interface to devices connected to the group switch.

The overall solution made extensive use of breakthroughs in available technology, including 0.5 μm ASICs and a new set of compact interfaces within the core blocks of the switch.

The authors describe various aspects of the new group switch hardware in AXE, highlighting the key features and advantages of the new products. As always in AXE development, special attention has been paid to compatibility, which is especially vital to a central system product such as the group switch.

The group switch holds a central position in AXE telephony applications, since practically all application hardware is connected to it. The group switch determines many vital system characteristics. In small-to-medium-sized applications, group switch hardware has represented only a minor fraction of the total hardware needed. In larger applications, however, the group switch has represented a somewhat voluminous hardware layout, consuming considerable amounts of power. The installation of a large switch has also required extensive internal cabling, which demanded a great deal of time and labour to install. Over the years, the AXE digital group switch hardware has gradually been modernised. Nonetheless, the potential for new compact hardware has existed for some time, given that some architectural changes were introduced to take advantage of recent electronics technology. Today, the time has come to exploit this potential.

Box A Abbreviations

ASIC	Application-specific integrated circuit
BM	Building module (1 BM = 40.64 mm)
BSC	Base station control
CCM	Caesium-based clock module
CLM	Clock module
DL2; DL3	Digital link 2; digital link 3
DLHB	Digital link multiplexer (half-size) board
DLIC	Digital link integrated circuit
DLMUX	Digital link multiplexer
EM	Extension module
EMB	Extension module bus
EMC	Electromagnetic compatibility
ESD	Electrostatic discharge
ETSI	European Telecommunications Standards Institute
GDM	Generic device magazine (subrack)
GS	Group switch
GSS	Group switch subsystem
ICM	Incoming clock conversion module
IEC	International Electrotechnical Commission
LED	Light-emitting diode
MSC	Mobile switching centre
MUP	Multiple position (time slot)
PBA	Printed board assembly
PCM	Pulse code modulation
RCM	Reference clock module
RP	Regional processor
SNT	Switching network terminal
SPDB	Space switch diagonal board
SPIB	Space switch interconnecting board
SPIC	Space switch integrated circuit
SPM	Space switch module
SRS	Subrate switch
TPI	TSM parallel interface
TSFI	Time slot frame integrity
TSIC	Time switch integrated circuit
TSM	Time switch module
TSSI	Time slot sequence integrity
TST	Time-space-time architecture

Customer demands for a compact, low-power, easy-to-install product as well as technological breakthroughs in the area of highly integrated circuits (ASIC) are the prime factors behind the latest improvements in the group switch hardware, which comprise the switch matrix and the control and synchronisation systems.

For reasons of compatibility, most concepts in the existing group switch have been maintained. These have, moreover, proved to be highly successful.

Compared with other switching structures, the time-space-time structure has several advantages. Thus, designers retained the key structural principles in the group switch subsystem (GSS), leaving it virtually unchanged. To connect the channels of two digital pulse code modulation (PCM) systems, the digital group switch first selects a path – which is defined by reading a time slot position (incoming time switch module); it then closes a crosspoint in a space switch module; and finally it writes a time slot position (outgoing time switch module).

Speech samples are managed in three steps by time switch modules (TSM) and space switch modules (SPM). In terms of modularity and traffic capacity, this structure also offers appropriate performance in conditions when traffic is high.

The group switch is terminated on switching network terminals (SNT), which are designed to operate with either 24-channel or 32-channel PCM systems. Timing is controlled by a synchronisation system that is implemented using three clock units and appropriate software algorithms. A clock module (CLM) provides STRATUM 3E stability performance (local exchange equivalent). Applications that demand greater synchronisation accuracy may connect a reference clock module (RCM) or a Caesium-based clock (CCM) to the group switch.

To guarantee reliability, the entire switching architecture is duplicated with two identical, fully synchronised networks.

In the past, the group switch interconnected the time-space-time stage with external switching network terminals using a DL2 (digital link 2) interface, which carries

32 time slots at 4 Mbit/s through a double four-pair cable. A fully equipped 64 K switch using the DL2 interface contained 2,048 such cables. Incoming data from 16 switching network terminals was multiplexed and stored in the incoming speech store of the time switch module. The data was then read and addressed to the space switch module by the control store. All reading or writing procedures were executed by direct control of the regional processor (RP).

The enhanced GSS

The newly enhanced group switch subsystem requires only half as many board types as its predecessor; its footprint is between 80% and 95% smaller in large configurations, and power has been reduced by similar amounts. Four main factors have made these achievements possible:

- 3.3 V application-specific integrated circuit technology for the basic components that make up multiplexing, time switching and space switching hardware;
- a new high-speed internal interface (DL3), which drastically reduces the volume of internal cabling;
- a suitable, advanced equipment practice;
- use of subrack-integrated regional processors.

Architectural aspects

To minimise the impact that the new hardware would have on software, designers preserved the basic principles of the switching structure. Thus, only minor architectural changes have been made within the new group switch subsystem. These include a new multiplexing stage and a faster switch port interface. In addition, the regional processor control, associated control interfaces, and the clock distribution have been optimised.

When introducing these and other changes – which include additional or improved functionality, such as switching capacity up to 128 K multiple-positions, subrate (= n x 8 kbit/s) switching facilities, and advanced wideband (= n x 64 kbit/s) switching capabilities – designers took care to ensure that the enhanced group switch subsystem would remain compatible with previous generation hardware. Figure 1 shows the architecture of an enhanced 64 K group switch.

Outlined below are the key concepts on which the architectural solutions are based.

- Functional blocks remain unchanged – the switching functions continue to perform in three steps: time-space-time, where time switch modules handle 512 multiple positions, and space switch modules handle 32 time switch modules. A new TSM-SPM interface operates at 48 MHz.
- A 16 K switch, complete with regional processors (excluding timing modules), fits in a single subrack for each plane.
- A 64 K switch fits in only four subracks for each plane.
- A full 4 K switch solution, complete with timing and regional processor modules, fits in a single subrack for two planes.
- The switch is extended to 128 K by redesigning the subrack slightly.

The new hardware solution gave rise to a number of structural modifications.

- Four time switch modules have been co-located on the same board (TS4B) – with only a minor impact on the blocking and maintenance software and on the TSM connection.
- One space switch module now fits on a single hardware board (SPIB or SPDB).
- A new digital link interface (DL3) has been introduced.
- Clock modules are connected to the diagonal SPM board (SPDB) within the 16 K subrack only – synchronisation and clock pulses are then distributed through the backplane to all non-diagonal SPM (SPIB)

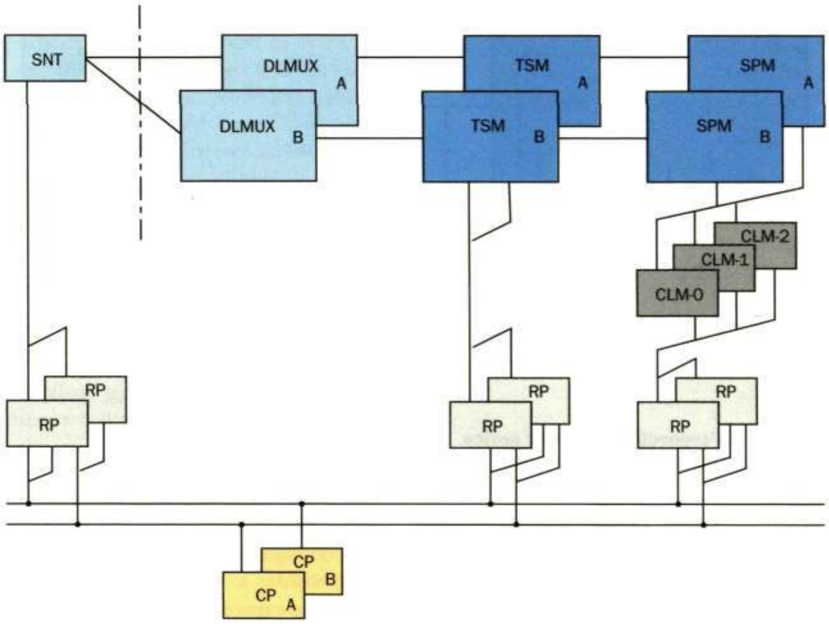


Figure 1
Architecture of an enhanced 64 K group switch.

Box B
Main characteristics of the new group switch subsystem

Size reduced by more than 90%.
Power consumption decreased by more than 90%.
Number of boards reduced by 50%.
Drastic reduction of internal cabling.
Improved fault-localisation performance.
Subrate switching facility.
Non-contiguous wideband functionality.
Electromagnetic compatibility (EMC) compliance, Class B or better.
Group switch disturbance statistics.
Idle pattern, programmable choice.
Low-voltage high-integration technology.
Equipment practice, BYB 501.

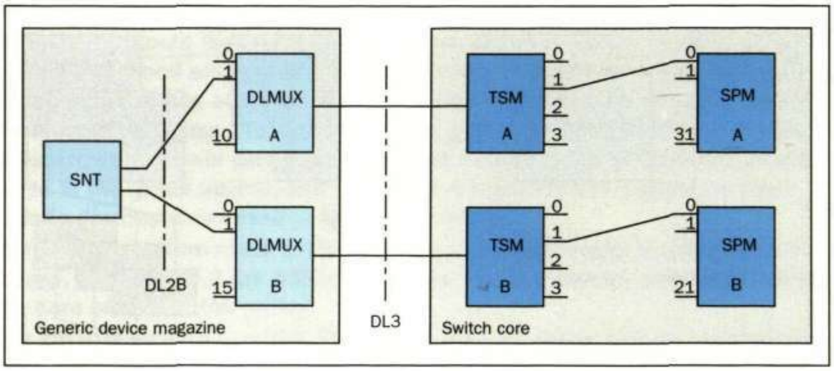


Figure 2
DL3 interface connecting a generic device magazine (subrack) to the switch core.

and time switch modules in the same subrack; thus, the 16 K group depends on, and follows the maintenance state of, one SPM board.

Since a single space switch module can handle 32 time switch modules, only one SPM function is needed for a 16 K switch. Above this threshold, the number of space switch modules quadruples. That is, four space switch modules are needed to handle a 32 K switch, nine are needed to handle a 48 K switch, and 16 are needed for a 64 K switch. Each time switch module sends its data to the space switch modules in the TSM's rows through four (or eight) separate ports, called horizontal highways. Data is received from the space switch modules in the TSM's columns through four (or eight) separate ports, called vertical highways. Note: four ports apply to 64 K solutions; eight ports apply to 128 K solutions.

Extension to 128 K multiple positions

A 16 K group switch subrack may contain up to eight space switch modules. Consequently, because the ASIC that implements the TSM function is designed to send data to (and receive data from) eight space switch modules, the switching capability may be extended to 128 K multiple positions.

A 128 K group switch is made up of eight 16 K group switch subracks per plane. Each plane is housed in two cabinets. Due to the large volume of cables, additional empty cabinets are required at the back for interconnecting the 16 K group switch modules. In a second development step, which will take place in 1997, the switch is to be extended to 128 K.

Subrate switching facility

A subrate switch (SRS) module, which is interconnected as an add-on to the group switch, enables switching functions to operate more effectively at subrate levels (8 kbit granularity). This function is primarily used in the digital mobile GSM application. The subrate module lowers the switching rate from 64 kbit/s (normal) to 8 kbit/s. Accordingly, up to eight times as many cellular calls may be handled during the same time span.

The subrate function makes efficient use of transcoder equipment in pool as well as transmission resources between the base station controller (BSC) and the mobile switching centre (MSC).

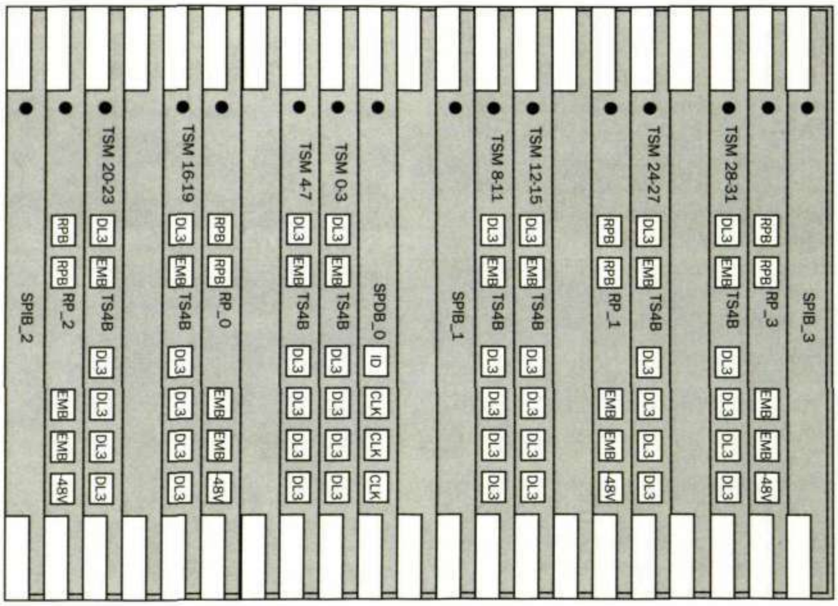
The subrate switch module is available for existing and enhanced group switch subsystems. The initial release, in the BYB 202 equipment practice, uses a TSM parallel interface (TPI) to the time switch module. Subsequent releases, which are being implemented in the BYB 501 equipment practice, use the new digital link interface (DL3).

The subrate module is duplicated. Its two planes are terminated at either connection side: the time switch module terminates the subrate switch duplication, and the subrate switch terminates the group switch duplication.

Other new or improved functionality

Worth mentioning among new or improved

Figure 3
The 16 K group switch subrack (GS16M).


SPIB_2
RPB RPB RP_2 EMB EMB -48V
TSM 20-23 DL3 EMB TS4B DL3 DL3 DL3 DL3
TSM 16-19 DL3 EMB TS4B DL3 DL3 DL3 DL3
RPB RPB RP_0 EMB EMB -48V
TSM 4-7 DL3 EMB TS4B DL3 DL3 DL3 DL3
TSM 0-3 DL3 EMB TS4B DL3 DL3 DL3 DL3
SPDB_0 ID CLK CLK CLK
SPIB_1
TSM 8-11 DL3 EMB TS4B DL3 DL3 DL3 DL3
TSM 12-15 DL3 EMB TS4B DL3 DL3 DL3 DL3
RPB RPB RP_1 EMB EMB -48V
TSM 24-27 DL3 EMB TS4B DL3 DL3 DL3 DL3
TSM 28-31 DL3 EMB TS4B DL3 DL3 DL3 DL3
RPB RPB RP_3 EMB EMB -48V
SPIB_3

functionality are advanced wideband characteristics, group switch disturbance statistics, and the free-alternative selection for the idle pattern.

Time slot sequence integrity (TSSI) and time slot frame integrity (TSFI) may be used to switch n x 64 kbit/s channels. TSSI and TSFI provide support for either contiguous or non-contiguous channel allocation ($2 \leq n \leq 31$). The blocking probability of an n x 64 kbit/s connection is the same as for connecting n individual 64 kbit/s channels. The traffic capacity may be increased, depending on the value of n.

The idle pattern may be set in the hardware solution according to the μ-law or A-law. The pattern is written automatically by the new time switch module under the supervision of a regional processor.

The DL3 interface

The DL3 interface – a new high-speed serial interface between the digital link multiplexers (DLMUX) in the generic device magazines (subracks) and time switch module stages in the switch core – operates at 49 Mbit/s, carrying 512 time slots per frame. Existing switching network terminals with a DL2 interface may be connected to the switch core by means of a DLMUX. Digital link multiplexers can either multiplex 16 DL2s to one DL3 or demultiplex one DL3 to 16 DL2s. Each DLMUX is supervised by the regional processor that controls the time switch module. Also, in accordance with requirements for reliability each DLMUX is duplicated. The DLMUX is transparent to equipment connected on the DL2 side of the interface, which is a key requirement for backward compatibility.

System modularity, hardware implementation and the equipment practice

The time-space-time structure is a multistage architectural solution that allows modular hardware to be installed and optimised according to requirements for switching capacity.

In terms of implementation, the 16 K group switch subrack is the basic building block (Figure 3). Each subrack contains:

- up to eight boards with four TSM functions per board (TS4B);
- one SPM board – which forms the 16 K group switch with time switch modules;
- up to three SPM boards – for interfacing time switch modules that belong to other subracks (they are needed to extend the switch capability above 16 K);
- up to four regional processors interconnected by a serial RP bus – each regional processor uses one extension module (EM) bus in the front and another in the backplane; thus, one regional processor controls eight extension modules per plane.

The space switch matrix is implemented with a single ASIC, called the space switch integrated circuit (SPIC). This circuit comprises a complete 16 K group switch matrix with an approximate complexity of 40,000 equivalent gates. There are two types of space switching board:

- the space switch diagonal board (SPDB), which contains the common clock functions;
- the space switch interconnecting board (SPIB), which contains transmitting and receiving drivers for the cable interconnection.

The time switch function, called a time switch integrated circuit (TSIC), has also been implemented as an ASIC. The circuit, which does not include the link multiplexing function, has an approximate complexity of 90,000 equivalent gates. Four TSICs are co-located on the same board, making it possible to increase modularity in steps of 2 K instead of 512.

Each board contains a light-emitting diode (LED) indicator and is equipped with an E2PROM that stores inventory-related information, in accordance with the new AXE hardware principles described in AXE hardware evolution[1].

Timing functions are implemented in a separate subrack, which contains the clock module, the reference clock module, and the incoming clock conversion module (ICM). The Caesium high-stability clock is separate.

Thanks to the building block design of the 16 K switch, the number of subracks

Box C

Time-space-time versus monostage matrix switching architectures

The three-stage time-space-time (TST) architecture offers a number of advantages compared with monostage architectures.

For example, the TST architecture essentially allows the hardware required for a given switching capacity to be optimised and combined in the most favourable way – something that other architectural solutions cannot normally do.

On the other hand, monostage structures offer better performance in terms of

- switch blocking probability characteristics (for example, strictly non-blocking);
- optimal efficiency in specific capacity ranges.

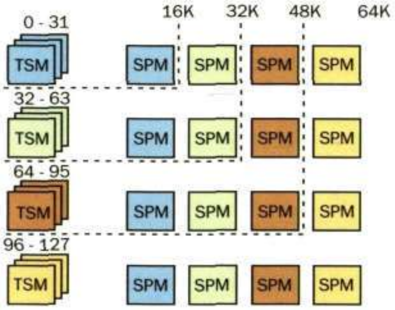


Figure 4
Logical view of the 64 K group switch (GS64K): one pattern is represented by one 16 K group switch (GS16K) subrack.

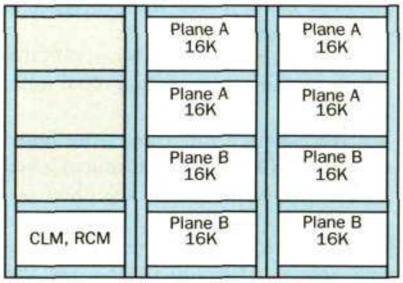


Figure 5
Cabinet assembly for the 64 K switch.

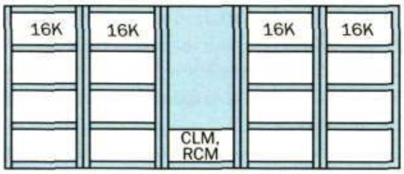
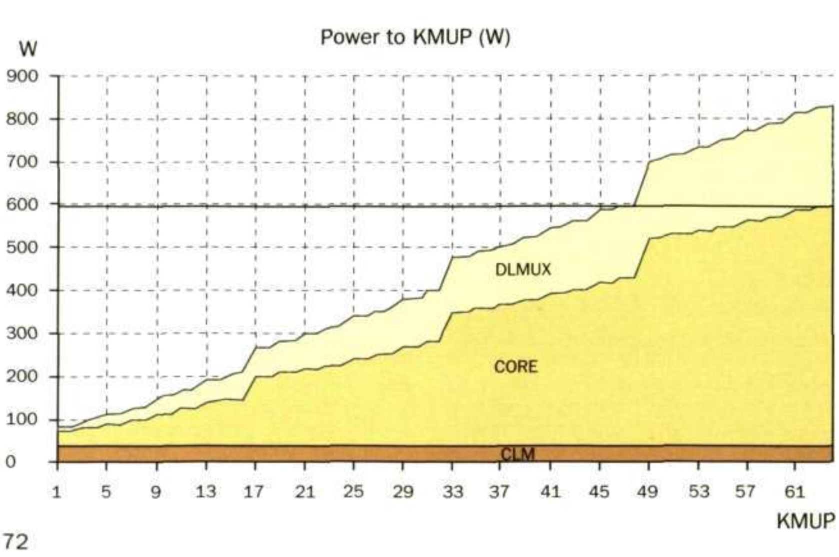


Figure 6
Cabinet assembly for a 128 K switch.

Figure 7
Relative power consumption of the main switch components.



may grow linearly in accordance with requirements for switching capacity.

Figure 4 shows a logical view of the 64 K group switch.

The new equipment practice, BYB 501[2], not only represents the basis for efficient physical dimensioning and cooling of the system; it also satisfies properties – electromagnetic, seismic, electrostatic discharge, installation and handling – normally specified in international environments. The BYB 501 complies with standards issued by the International Electrotechnical Commission (IEC) and the European Telecommunications Standards Institute (ETSI). Therefore, it may readily be adapted now and in the future to accommodate sourced products.

A relevant application of the new equipment practice is the generic device magazine, which is a subrack that houses a pair of digital link multiplexer (half-size) boards (DLHB), a pair of regional processors, and 16 switching network terminals. Connections between the boards are made in the backplane. Thus, the DL2 and extension module bus interfaces are no longer implemented using cable. The DLHB multiplexes 16 DL2 connections into a single DL3: the DL2-to-DL3 interface is implemented as another ASIC called the digital link integrated circuit (DLIC).

Another application of the new equipment practice is the 16 K group switch (GS16K) subrack. This single-depth subrack houses a 16 K group switch. Each additional increment of 16 K (up to the maximum capacity of 128 K, Figure 6) requires a double-depth subrack per plane.

Double-depth subracks in 800 mm cabinets are needed to accommodate the large volume of cabling in the backplane. Despite significant reductions in cable volumes, cabling still limits the size of this type of group switch. This is because nearly every board is connected to one or more cables. Likewise, cabling is required for regional processors and power supply.

Technical characteristics of the enhanced GSS

The hardware evolution project for the group switch subsystem has achieved truly amazing results. Two key factors of the project's success are the extensive use of low-voltage technology (ASIC) in the new hardware design and the adoption of the new equipment practice, BYB 501.

In terms of power consumption, a complete configuration of the new 64 K switch dissipates approximately 600 W. Compared with the more than 19,000 W for the equivalent previous-generation switch, power consumption has been reduced by more than 95%.

Figure 7 shows the relative power consumption of the main switch components (core switch, digital link multiplexer, and clock module).

In terms of footprint, the new switch core houses time switch modules and space switch modules within a volume of only 96 building modules (BM), compared with 2,400 BM in earlier configurations. Again, this is a reduction of more than 95%.

Figure 8, which relates to the mechanical solution as a function of the switching capacity, compares the number of subracks needed in the new group switch with the number of subracks needed in its predecessor.

Figure 9 shows a comparison of the volumes of wires needed in the cabling of the internal TSM-SPM interconnections.

Conclusion

The enhanced group switch subsystem represents a significant contribution to the new AXE hardware. The implemented solution is an efficient enhancement of a central subsystem that, while based on reliable, proven switching principles, manifests vastly reduced footprint and power consumption.

The positive consequences of this evolution are reduced manufacturing time; faster, less complicated installation; and reduced cost of ownership.

To the network provider, the design steps achieved thus far are very evident; they also have the greatest impact on operator costs. Nevertheless, the subsystem – as well as most of the external devices connected to it – will continue to evolve, particularly in terms of added functionality and in incorporating relevant breakthroughs in technology.

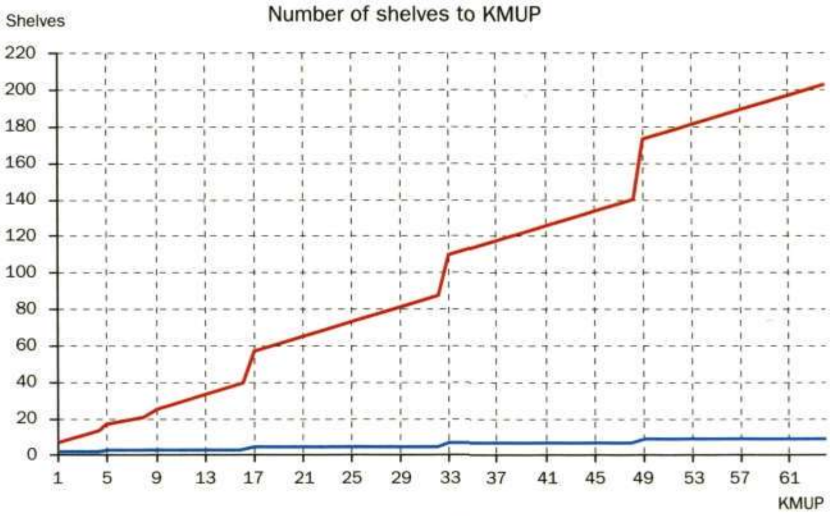
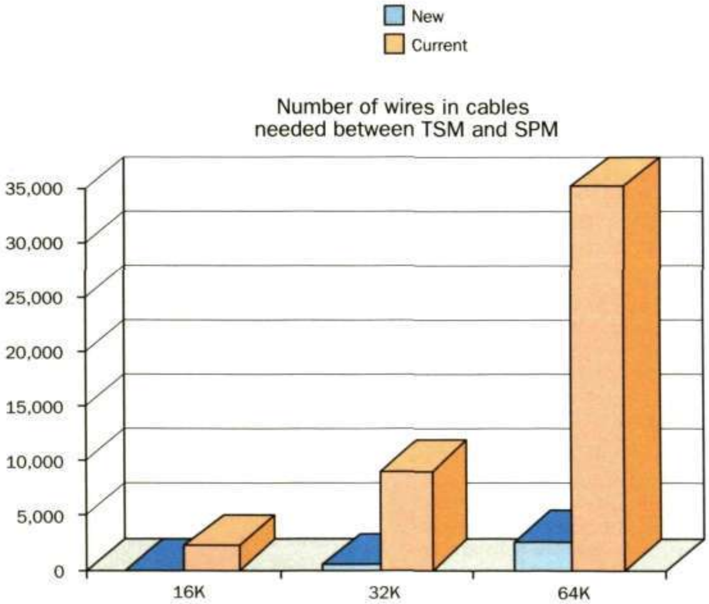


Figure 8
Number of subracks needed in the new group switch versus the number of subracks in its predecessor.

Figure 9
Comparison of the volume of wires needed in the cabling of the internal TSM-SPM interconnections.



References

1 Lundqvist, T., Hägg, U.: AXE hardware evolution. Ericsson Review 74 (1997):2, pp. 52-61.

2 Stockman, B., Wallers, A.: BYB 501 metric equipment practice. Ericsson Review 74(1997):2, pp. 62-67.

3 Ericson, B. and Roos, S.: Digital Group Selector in the AXE 10 System. Ericsson Review 55(1978):4 pp. 140-149.

4 Braugenhardt, S. and Nordin, J.-E.: Ericsson Review 55(1978):4 pp 150-163.

Standardisation in the world of information and communications technology

Tom Lindström

Standardisation in the area of telecommunications and information technology is undergoing rapid development. As relates to telecommunications, this phenomenon was triggered by the wave of deregulation that began sweeping the globe in the early 1980s.

Old standards-setting organisations are adapting their processes and agendas to the requirements of a new, more global arena. Their cherished processes, which involve carefully selecting and specifying a "best technical solution", must now give way to a standardisation arena where product vendors and service providers actively position their solutions for future markets.

As several industries converge, and as information and communications technologies become more global, a very complex mechanism is emerging to deliver the standards that are required for supporting the envisaged information infrastructures and societies of the future.

The author describes the forces behind developments in standardisation that pertains to information and communications technology. He concludes that standardisation activities must be driven by leading business-oriented competitors who ensure that the outcome meets the needs of their customers and the end-user.

Box A
World Trade Organization definition of standard

Standard (extract)
For the term "Standard" the following definition shall apply:
Document approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for products or related processes and production methods, with which compliance is not mandatory.

Explanatory note:
Standards as defined by ISO/IEC Guide 2 may be mandatory or voluntary. For the purpose of this Agreement standards are defined as voluntary and technical regulations as mandatory documents. Standards prepared by the international standardization community are based on consensus. This Agreement covers also documents that are not based on consensus.

Source:
World Trade Agreement, The Final Act and Agreement Establishing the World Trade Organization. (including GATT 1994), Uruguay Round. Marrakech, 15 April 1994. Agreement on technical barriers to trade, Annex 1: Terms and their definitions for the purpose of this agreement.

Industries and markets are merging. Today's hottest markets are found where telecommunications industries and industries that create or process information intersect. At the same time, an information society is being ushered in by national and regional authorities who do not want to be left behind in the race to generate or keep business and employment. New national, regional and global information infrastructures are being established, while information and communications technologies (ICT) cater to the needs of the infotainment/edutainment industry.

To support these many developments, and to keep information moving freely, all sorts of ICT-related standards are needed as facilitators at all levels of technical systems. Further, the general view is that these standards need to be developed more rapidly than ever before. Without them, a jungle of incompatibility will arise that leaves targeted users disenchanted, thereby spoiling valuable business opportunities. Standards play a crucial role in creating and – at least temporarily – stabilising market segments that are caught up in rapid technological change.

Global players in the area of ICT must use relevant standards when designing products. Moreover, to use standardisation as an effective tool for business development they must be deeply involved in the process of setting standards, as well as in forming the structures and processes that are used to develop, approve, market, and finally win user acceptance of the standards.

Trends in standardisation

The role of standards

The role of ICT-related standards is growing in importance. Recognition of this fact is also growing. Commercial enterprises who keep pace with, or stay ahead of, and adapt their behaviour to the changing environment will emerge as tomorrow's winners in the ICT market.

In the past, international telecommunications standards were needed to ensure that international services could be delivered and charged successfully. All other aspects of operation were handled at national or local levels. Today, however, telecommunications and information-processing systems need to interconnect and interoperate at a variety of levels and connection points. Even so, the global communications web is not particularly well-structured, but an expanding maze of alternative networks.

Time to market, as a result of rapidly changing technologies and of a globally competitive market, has become a determining factor for the business success of service providers and product vendors. Achieving short time to market not only requires that standards be developed quickly, but that they fulfil genuine user needs and can be implemented profitably.

Open standards

Today it is generally understood that open standards benefit the area of information and communications technology as a whole. Indeed, few if any commercial enterprises are fully capable of creating an entire market on their own. It has also been generally accepted that if a standard is to be widely adopted, then the intellectual property rights (IPR) that are associated with it must be licensed under fair and reasonable conditions.

Proliferation of standardisation groups

In order to meet the demands for new standards, established standards-setting organisations such as ITU-T and ETSI are currently in the process of adapting their structures and working procedures. As a basis for this change, they analyse changes within their environments, determining what the market expects of them. In part, they must overcome conservatism and ineffective processes. Similarly, they must try to recruit new members who have no tradition within the organisation. Rather than joining forces with what they perceive to be a slow, unwieldy giant (many committees, and lots of liaison), these potential members frequently bypass established standardisation organisations. Instead, they search out other partners with similar business interests with whom they can form new fora/consortia to draft the specifications that are needed to create a new market.

Although the officially recognised standardisation bodies have enjoyed an influx of new members in recent years, special interest groups continue to flourish for several technical areas – particularly for those areas that boarder on telecommunications and data communication, and in the area of "middleware". These special interest groups are frequently modelled on examples from the computer and data communication industries, where shorter product life cycles have required manufacturers to agree very quickly on the basic common specifications that are needed for creating a market.

Equipment suppliers often take the initiative to set up special interest groups for new areas of technology that cut across several standardisation bodies. Today numerous groups of this kind are active in a broad variety of areas. Examples of their activities include:

Box B ITU-T study groups – general areas of study

Study Group 2	Network and service operation
Study Group 3	Tariff and accounting principles including related telecommunications economic and policy issues
Study Group 4	TMN and network maintenance
Study Group 5	Protection against electromagnetic environment effects
Study Group 6	Outside plant
Study Group 7	Data networks and open system communications
Study Group 8	Characteristics of telematic systems
Study Group 9	Television and sound transmission
Study Group 10	Languages and general software aspects for telecommunication systems
Study Group 11	Signalling requirements and protocols
Study Group 12	End-to-end transmission performance of networks and terminals
Study Group 13	General network aspects
Study Group 15	Transport networks, systems and equipment
Study Group 16	Multimedia services and systems

Box C ETSI technical bodies

ETSI projects
- Analogue terminals and access
- Broadband radio access networks
- Corporate networks
- Cordless terminal mobility
- Digital enhanced cordless telecommunication
- Digital terminals and access
- Multimedia terminals and applications
- Pay terminals and systems
- Special mobile group
- Terrestrial trunked radio
- Tiphon, Telecommunications and Internet Protocol Harmonization Over Networks

Technical committees
- Communication, networks and systems interconnection
- EBU/CENELEC/ETSI Joint technical committee
- EMC and radio spectrum matters
- Equipment engineering
- Human factors
- Integrated circuit cards
- Methods for testing and specification
- Network aspects
- Radio equipment and systems
- Satellite earth stations and systems
- Security
- Signalling, protocols and switching
- Speech processing, transmission and quality aspects
- Transmission and multiplexing

Ericsson standardisation guidelines

- market liberalisation
- minimum technical regulations
- international standards and global aspects
- open organisations and direct membership
- minimum political involvement
- minimum national levels in development and approval stages
- maximise dissemination by minimising publications price
- adequate compensation for IPR use
- fair licensing to support standards development

The importance of standardisation

To technical units
- helps monitor new technology
- affects product strategies
- is increasingly important to follow
- can be influenced together with others
- is a good investment

To business units
- creates new markets
- provides market information
- illustrates trends in the industry
- promotes company image if done well
- is a good investment

Figure 1
The internal supply of standards-related information and documentation is assisted by an intranet Web site. Standards organisations deliver their output on-line or on CD-ROM.

Box D
Working groups of the ATM Forum

- Broadband inter-carrier interface (BICI)
- LAN emulation (LANE)
- Multiprotocol over ATM (MPOA)
- Network management (NM)
- Physical layer (PHY)
- Private network-to-network interface (PNNI)
- Residential broadband (RBB)
- Security (SEC)
- Service aspects and applications (SAA)
- Signalling (SIG)
- Testing (TEST)
- Traffic management (TM)
- Voice and telephony over ATM (VTOA)
- Wireless ATM (WATM)

Box E
Working groups of the Internet Engineering Task Force

- Applications
- General
- Internet
- Network management
- Operational requirements
- Routing
- Security
- Transport
- User services

- defining implementation agreements based on existing standards;
- complementing existing standards;
- creating specifications for interoperability;
- developing specifications for the testing of conformance.

Another important goal of these groups is the marketing of specific solutions.

Additionally, operator and user initiatives often take the form of memoranda of understanding (MoU), which are used to promote specific standards and to formulate requirements for standards that are being developed.

Being set up to cater to a particular subject, special interest groups can contribute significantly to the international standardisation process. Their members – many of whom are also active in recognised standardisation bodies – can apply the groups' results to the work of these organisations. Moreover, in cases where the subject is not covered in depth by a standardisation body, a special interest group may act as a subcontractor for the organisation.

Many large players of the ICT market are obliged to participate in most new initiatives. Therefore, in the interest of time and resources, their challenge is to limit the number of potential parallel or overlapping initiatives without suppressing the timely development of needed specifications.

Different standardisation cultures

There are two very different approaches to organising the standardisation process. To prosper in the different markets that emerge, global companies must adapt their operations to both approaches.

According to the first approach, which is exemplified by segments of the standardisation system in the US, the organisations form a competitive standardisation market. Ultimately, the market determines – by the number and kind of products and services sold – a "best" standard. To achieve rapid time to market for their standards, these organisations prefer a simple majority-vote procedure. The advantage of this approach is that it permits standards to be developed more quickly than in a planned process, since not every diverging interest needs to be reconciled by achieving consensus (Box G).

The second approach which, for example, is practised in the standardisation system of Europe, involves a planned and strictly controlled standards-setting process that is

Box F Organisations that set standards for information and communications technology

ADSL Forum	Asymmetric Digital Subscriber Line Forum
ANSI (US)	American National Standards Institute
ARIB (Japan)	Association of Radio Industries and Broadcasting
ATM Forum	Asynchronous Transfer Mode Forum
CEN (Europe)	Comitté Européen de Normalisation
CENELEC (Europe)	Comitté Européen de Normalisation Electrotechnique
CITEL (Americas)	Comison Interamericana de Telecomunicaciones
CTIA (US)	Cellular Telecommunications Industry Association
CV Forum	Charging Vendors Forum
DAVIC	Digital Audio Visual Council
DVB	Digital Video Broadcasting Project
ECMA	A Europe-based international industrial organisation for standardising information and communication systems
ECTEL	European Telecommunications and Professional Electronics Industry
ECTF	Enterprise Computer Telephony Forum
EFTI3	European Forum for Telecom Industry Information Interchange
EITIRT (Europe)	European IT Industry Round Table
ERT (Europe)	European Round Table
ETNF (Europe)	European Telecom Numbering Forum
ETSI (Europe)	European Telecommunications Standards Institute
Eurobit (Europe)	European Association of Manufacturers of Business Machines and Informatiön Technology Industry
FIPA	Federation for Intelligent Physical Agents
FRF	Frame Relay Forum
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IITC (US)	Internetwork Interoperability Testing Council
IMTC	International Multimedia Teleconferencing Consortium
IN Forum	Intelligent Network Forum
IISP (of ANSI) (US)	Information Infrastructure Standards Panel
IPNS Forum	ISDN PABX Networking Specification Forum
ISO	International Organization for Standardization
ITS (Sweden)	Informationstekniska standardiseringen
ITU-D	International Telecommunication Union - Telecommunication Development Sector
ITU-R	International Telecommunication Union - Radiocommunication Sector
ITU-T	International Telecommunication Union - Telecommunications Standardization Sector
JTC1 (of ISO-IEC)	Joint Technical Committee 1 (IT)
MMCF	Multimedia Communications Forum
NIUF (US)	North American ISDN User Forum
NMF	Network Management Forum
NOF (US)	Network Operations Forum
NRIC (US)	Network Reliability and Interoperability Council
OMG	Object Management Group
PCIA (US)	Personal Communications Industry Association
SEK (Sweden)	Svenska Elektrotekniska Kommissionen
SGML Open	Standard Generalized Markup Language Open
T1 (US)	ANSI Committee T1, Tele-communications
TIA (US)	Telecommunications Industry Association
TTC (Japan)	Telecommunications Technology Council
W3C	World Wide Web Consortium
VoIP	Voice Over IP Forum
UMTS Forum	Universal Mobile Telecommunications System Forum
TINA-C	Telecommunications Information Networking Architecture Consortium
XIWT (US)	Cross-industry Working Team
X3, IT (US)	ANSI Committee X3, Information Technology
X12, EDI (US)	ANSI Committee X12, Electronic Data Interchange

characterised by the objective (sometimes perceived as being mandatory) of having only one standard per application or product area. This approach is based on the premise that a single specification gives users the greatest benefit by eliminating incompatible solutions from the outset; competition in the market is reduced to the provision of different implementations of the same specification. The existence of multiple (competitive) standards would indicate that the process has failed. A drawback to this approach is that consensus is often difficult to obtain.

Many in industry believe that competition among standards will provide the best solution. In the short term, this approach permits several solutions to be introduced into the market quickly, thereby allowing free competition. In the long term, the competitive approach contributes towards making the market more dynamic, a place where new ideas may be tested freely.

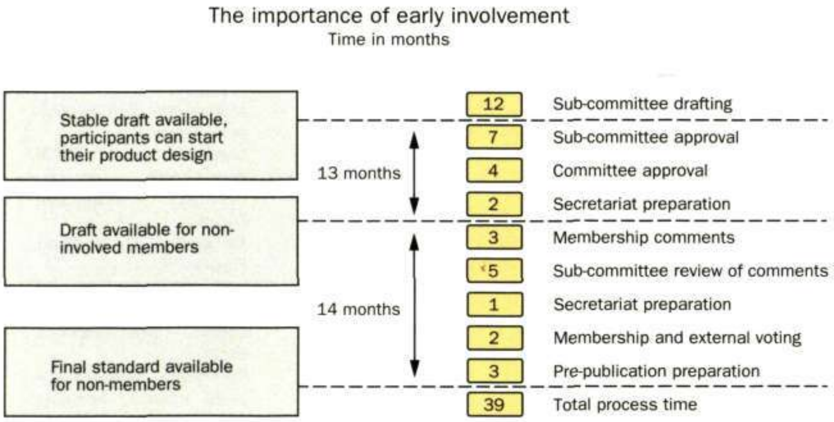
If standards are to be developed for the ICT market – which is in the midst of dynamic development and whose technologies are in a state of flux – then it may be advantageous to pursue the competitive approach. This way, the market does not have to take a premature stand on any particular solution. Nevertheless, as relates to certain areas of telecommunications, standardisation is needed to ensure that incompatibility is not introduced into large existing installations.

For business enterprises, the prime concern is that an open specification be created and adopted by the market. How this is

Box G
ISO/IEC definitions of consensus and standardisation

Consensus
General agreement, characterized by the absence of sustained opposition to substantial issues by any important part of the concerned interests and by a process that involves seeking to take into account the views of all parties concerned and to reconcile any conflicting arguments. Note: Consensus need not imply unanimity (ISO/IEC Guide 2, 1.7).

Standardisation
Activity of establishing, with regard to actual or potential problems, provisions for common and repeated use, aimed at the achievement of optimum degree of order in a given context (ISO/IEC Guide 2, 1.1).


The importance of early involvement
Time in months
Stable draft available, participants can start their product design
Draft available for non-involved members
Final standard available for non-members
13 months
14 months
12
7
4
2
3
5
1
2
3
39
Sub-committee drafting
Sub-committee approval
Committee approval
Secretariat preparation
Membership comments
Sub-committee review of comments
Secretariat preparation
Membership and external voting
Pre-publication preparation
Total process time

Note: this is a ficticious example of lead time

Figure 2
Involvement in a standards committee gives early access to the direction of future standards, and facilitates early market introduction of new products.

achieved is less important. The ideal standard would stabilise the market and allow for products to be sold in large volumes. It would also give users a variety of interoperable products that cost less. Obviously, timing is paramount: if a standard arrives on the market too late – even when it is technically superior – it cannot compete with specifications that have already been established.

Restructuring the old organisations

The challenge to the standardisation community is to combine the working methods and results of each approach in order to produce appropriate and timely standards for information and communications technologies. All wasteful duplication of work must be avoided without stifling healthy competition. The focus and drive of special interest groups must be combined with the public consultation that traditional organisations use to reach a broad consensus.

Technology-driven versus business-driven standardisation

Competition for expert resources

Because some important parts of ICT-related principles and product architectures are created in standardisation organisations, the commercial enterprises who want to influence the standardisation process send (at different stages in the process) business developers and technical experts to work on standardisation committees. Competition for these resources is great. Moreover, technical experts cannot be assigned on a full-time basis to external tasks without eventually losing contact with their internal projects.

Commercial enterprises in an increasingly competitive market cannot afford to allocate expert resources to standardisation processes unless these contributions can be justified from a business point of view. That is, decisions to spend resources on standardisation must be based on business criteria with an aim to creating markets for future products. Because practically every product in today's ICT market is affected by standards, product design and input to standardisation must be carried out as parallel activities. The technical experts involved must carefully coordinate their activities with business and product strategies alike.

The business drive

In addition to technical criteria, there is a driving force to introduce business criteria into ICT-related standardisation. This is because the proposed technical solutions are becoming more and more equivalent in scope. If products are to reach the market early enough to reap a significant market share, then product development must begin at a very early stage of the standardisation process. When setting a standard, the choice in a committee of one technical solution before another gives competitors with a lead in the chosen area a distinct advantage over other competitors.

Information and communication technologies are becoming increasingly complex and the standardisation process therefore takes more time to complete. Thus all interested parties need to become involved in related elaborations at an earlier stage. Otherwise it may be difficult to achieve an implementation that fulfils customer needs in the market that the standard is meant to create.

To some extent, the standardisation organisations set the agenda for the development departments of the ICT industry. To minimise the negative effects of this influence, industry needs to take a more active, business-oriented role in the standards-setting process.

International versus regional or national standards – a contest for territory

Open global markets are best served by globally approved and adopted standards. Nevertheless, regional and even national standards continue to exist, along with a national level in the approval process. This is due in part to the sluggish manner in which the standardisation community is being restructured. Likewise, local standards may be used as tools for creating trade barriers. Separate national standards are hardly required for information and communications technologies.

A national organisation, however, could represent small businesses and function as a central information bureau. Similarly, regional standardisation organisations might serve as consensus-builders, as part of the global standardisation process. They might also facilitate a coordinated introduction of standards at the national level.

To maximise the speed at which global standards are developed, approved, and adopted by the market, the commercial enterprises who sponsor expert resources need to be given direct representation and equal status (that is, decision-making power) in standards-setting organisations.

Intellectual property rights (IPR)

It is generally accepted that there is a need for policies on intellectual property rights – policies that prescribe fair licensing agreements to any party who wishes to implement an approved standard. To date, however, there is no solution to the problem of essential (standards-blocking) IPR (Box H) that come up late in the standardisation process. Moreover, this problem worsens as business and market positions increasingly influence the scope of ICT-related standards and how they are initiated.

Standardisation as a source of information

Although serious involvement in the standardisation process is costly and time-consuming, it can yield significant commercial benefits. For example, as participants from each sector of the ICT market become involved in the standards-setting process, they gain access to valuable information on customer needs, market demands, and supplier and technological capabilities.

Their participation also helps them to maintain their levels of proficiency, since standardisation activities are a driving force behind technical competence.

Making room for competition

Competitive environments require rapid procedures. That is, they require that the basic aspects needed for creating a market be drafted and standardised as early as possible. Later, in subsequent stages of development, additional functionality may be standardised – without covering too much detail, however, since product differentiation must be allowed in order to keep the market competitive. This approach marks a shift in attitude on the part of manufacturers and operators, whose current focus is on creating foundations on which to base future systems.

Today manufacturers no longer see the role of standardisation as one of merely providing a technological fix. Instead it functions as an important mechanism for helping industry to understand market needs. More importantly perhaps, the standardisation process helps increase the rate at which markets are developed. For manufacturers, this is the true value of standardisation, since the worst mistake a manufacturer can make is to address a market that does not exist.

The consequences of globalisation

An increasingly competitive global environment effectively precludes solutions that favour one particular country or market segment. Attempts by governments to mandate specific directions that are at odds with the global marketplace will in all likelihood only achieve two things. They will limit the quality and performance of available products and services to users within their jurisdiction and limit the scale of the market that is available to vendors.

Consensus among competitors

There is a growing realisation that, in order to create and encourage new markets, members of industry must collaborate in producing standards. Large commercial enterprises are no longer able to impose proprietary solutions. Today private service providers, ICT-related equipment manufacturers, and user organisations are joining with administrations and public operators in established standards organisations. The specific intention of these new members is to cooperate in creating useful standards.

Rapidly changing markets and converging areas of technology require that more standards be developed more quickly than ever before. This, in turn, implies that technical and organisational coordination must improve. Recent attempts to establish or to increase cooperation between different standardisation bodies are encouraging, but

Box H
Essential intellectual property rights

"Essential" as applied to IPR means that it is not possible on technical (but not commercial) grounds, taking into account normal technical practice and the state of the art generally available at the time of standardization, to make, sell, lease, otherwise dispose of, repair, use or operate equipment or methods which comply with a standard without infringing that IPR. For the avoidance of doubt in exceptional cases where a standard can only be implemented by technical solutions, all of which are infringements of IPRs, all such IPRs shall be considered essential.

Source:
ETSI Rules of Procedure, Annex 6, ETSI Interim Intellectual Property Rights Policy.

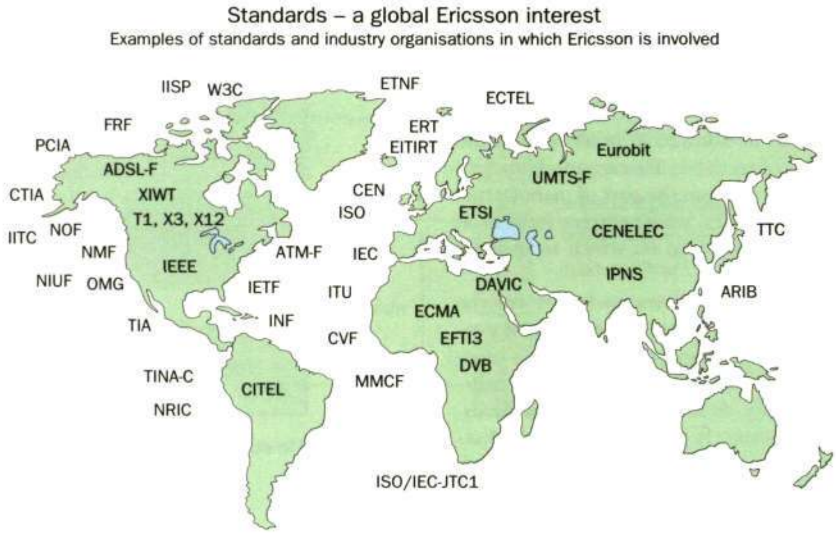
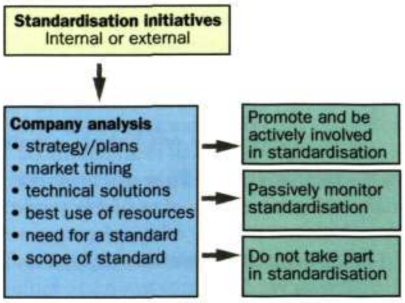


Figure 3
Ericsson have a global interest in standardisation. Many standards and industry organisations must be covered by appropriate strategists and experts.



Standardisation initiatives
Internal or external

Company analysis
- strategy/plans
- market timing
- technical solutions
- best use of resources
- need for a standard
- scope of standard

Promote and be actively involved in standardisation

Passively monitor standardisation

Do not take part in standardisation

Figure 4
New standardisation initiatives must always be analysed in the context of existing business strategies and product plans. The outcome of this analysis does not always lead to involvement.

also imply that an extra, unwanted layer of administration might compound the process.

Conflict or complement?

The different organisational structures of traditional standardisation bodies and new fora/consortia give onlookers the impression that these standardisers are competing with one another. Old standardisation bodies do not always have a direct membership option for the commercial enterprises who provide expert resources. Moreover, their working methods imply that they create committees for all eventualities and then try to fill them with work. New fora, on the other hand, are nearly always formed to work in a limited area.

Not all fora attempt to set standards, however. Many define their task as one of promoting a particular technology or solution. In doing so, they refine market requirements and specify architectures. This may be perceived as leaving to the established standardisation bodies – now without first-hand knowledge of new market opportunities – the sole task of delivering a technical specification.

Besides procedures, old and new organisations differ greatly in their ways of financing the standarisation process and of distributing standards. For example, the official organisations generally finance significant parts of their secretarial operations – centrally as well as at existing national levels – through state contributions and by selling copies of their documents. Firm in their conviction that documents are valuable products that should be sold at "market value", they have carried this practice over into the new era of standardisation – despite new electronic communications tools and distribution methods that considerably reduce costs.

On the other hand, many new organisations are required by their members to make their documents available in electronic format at no charge. Their aim is to disseminate their results to the widest possible audience hoping that their work will gain acceptance and be adopted by the market. Moreover, these new organisations conduct their work primarily in the electronic domain.

Another layer of consensus

The world of standardisation is affected by the burden of increasing needs for inter- and intraorganisational coordination. Industry is no exception. As commercial organisations evolve to become true global operators, their internal priorities diverge between product and geographical market areas. When this happens, they are faced with the challenge of not compounding the external complexity with that of their internal coordination.

Ericsson's support of global standardisation efforts

In order to serve their customers well, Ericsson engage in all standardisation activities that pertain to their different markets and products. Although a single coherent standardisation system is encouraged

Standardisation coordination networks

- Subject area network (for example, IP communication, network management) – covers all organisations active in the area.
- Standards organisation network (for example, ITU-T, IETF) – covers all committees of a specific organisation.
- Geographical network (for example, USA, Europe) – covers all standardisation organisations in a country or region.

Although the geographical networks pay special attention to regulatory groups within their areas, each network integrates coverage of relevant product-related regulatory aspects.

and supported, business needs often call for involvement that is independent of structures or processes.

Delegated responsibility

Ericsson have established a set of guidelines – based on overall corporate policies – for their activities in the global standardisation arena. A basic principle emphasised in these policies is that each Business Unit at Ericsson takes full responsibility for its individual standardisation needs. Even so, the Business Units are required to coordinate their activities with other parties in the corporation who have related activities.

Furthermore, each Business Unit coordinates its own standardisation involvement. Corporate engagement is kept to a minimum and exists only when initiated and approved by the Business Units.

Global involvement

Knowing that manufacturer support is vital to the efficient operation of the standardisation process, Ericsson are deeply committed to the standardisation of ICT-related products and services at the international level. Today several hundred company experts are involved in a variety of standardisation organisations.

Ericsson's philosophy, which is in harmony with current trends in standardisation, is to standardise the foundation on which marketable products can be based, rather than to standardise the details of each technology. This approach allows players to establish new markets, and supports innovation in the use and supply of ICT-related networks, products and services.

Ericsson maintain a set of internal coordinating networks for coping with the maze of organisations that cover relevant aspects of information and communications technology. These networks are made up of representatives of business and product lines, technical experts, market experts and standardisation process experts. There are several kinds of network:

- networks that approach subject fields from a global perspective, addressing every organisation that is active in each field;
- networks that focus on a specific organisation, coordinating Ericsson's engagement in that organisation's committees;
- networks that focus on a geographical market, covering every standards organisation and relevant regulatory bodies in that market.

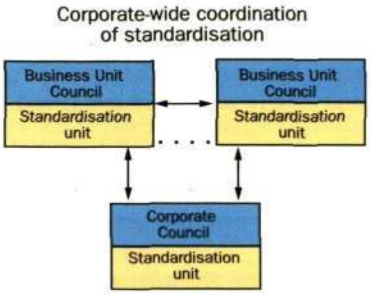


Figure 5
Business units are responsible for their own standardisation policies and needs. Business units cooperate one-to-one, and as collective bodies through a corporate council.

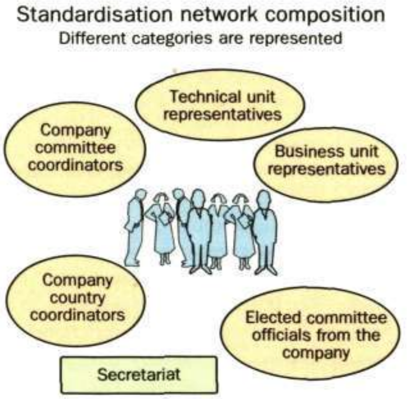

Standardisation network composition
Different categories are represented
Technical unit representatives
Company committee coordinators
Business unit representatives
Company country coordinators
Elected committee officials from the company
Secretariat

Figure 6
To succeed, internal standardisation coordination networks must comprise representatives from all related disciplines.

Conclusion

Good standardisation results require serious commitment on the part of operators and manufacturers. Their involvement, however, must not be limited solely to technical aspects, but must also take on a pragmatic business approach.

Today the number of parties outside of industry who have the knowledge and resources that are necessary to support essential standardisation efforts is on the decline. Therefore, industry must play a leading role in fulfilling the market-related needs of their customers: operators, businesses, and private users.

Appropriate coordination is needed to avoid increases in total product costs, as well as to avoid tardy market introduction of products and services. Coordination is needed between standardisation organisations; between parties who are active in these organisations; and internally, among the various interests of these parties.

Ultimately, if those who do the work and finance the processes feel that a proposed standard is worthwhile, then that standard will be developed in a timely manner.

Generic objectives of the standardisation coordination network

- Forum to coordinate total company efforts in subject area, organisation or geographical area.
- Ensure internal coordination of positions and review of contributions.
- Facilitate the flow of information to involved and interested units.
- Coordinate to optimise use of expert resources.
- Promote own areas internally.

References

The Structure of IT Standardisation. Oksala, Steven, Rutkowski, Anthony, Spring, Michael, O'Donnell, Jon, ACM StandardView, Vol.4, No. 1 (March 1996)

Today's Co-operative Competitive Standards Environment For Open Information and Telecommunication Networks and the Internet Standards-Making Model, A Rutkowski. Standards development and information infrastructure workshop, June 1994, John F Kennedy School of Government and NIST

Standards Practices, Ericsson Connexion, March 1994

Adjunct processor – A new AXE-integrated open-standard computer system for call-related data processing

Magnus Ericsson and Neela Koria

The adjunct processor is a new addition to the APZ processor family for AXE exchanges. Implemented on open-standard computers and integrated as a subsystem of AXE, the adjunct processor is a platform for developing new service-related applications in fixed and mobile telecommunications networks.

The first commercial application for the adjunct processor is a charging information system, known as the formatting and output subsystem, for the digital advanced mobile phone service standard for wireless network services.

The authors describe the new adjunct processor, how the adjunct processor **concept further strengthens AXE, the first application to run on the adjunct processor, and how the adjunct processor concept may be applied in future implementations. The adjunct processor will play an important role in opening up the AXE architecture.**

Network operators are finding that, thanks to the brisk pace of development in fixed and wireless telecommunications, they have many new opportunities to increase their revenues by adding new services to their networks. The faster these new services can be brought to market, the sooner operators can reap new revenues, thereby strengthening their competitive positions. Due to growing competition, network operators are also under pressure to increase the range and flexibility of their services while reducing their operating costs.

The net result of these trends is that, in addition to pure switching tasks, the nodes that switch traffic in a telecommunications network are becoming burdened with an ever-increasing scope of service and business-support functions. For instance, network operators are taking more innovative approaches to charging, which place heavy demands on the availability of call-related data in near real-time. To support new charging schemes, the service- and business-related data from switching nodes is usually transferred to external data-processing systems.

Demands for call-related data tend to be greater, and are growing even faster, in mobile networks than in equivalent fixed networks. The main reasons for this are:

- associated functions that deal with mobility management;
- the number of network elements from which charging data is collected;
- the greater range of services and statistics that are offered by a mobile network.

The data-processing load from these and other factors in mobile networks is currently growing at a rate of around 25% a year. And since new services are constantly being developed and enhanced, this trend is likely to continue for several years. Given these circumstances, it clearly makes sense to off-load capacity-demanding tasks that are not related to traffic.

Against this background, Ericsson have taken steps to integrate a new processor subsystem into AXE. The adjunct processor (AP), is an application platform with an open-standards interface that provides the specific data-processing, storage, and transmission capacity that new business and service management applications require.

In keeping with the growing trend towards open-standard computing platforms, the adjunct processor is implemented on industry-standard computers. These may range from fault-tolerant computers for critical, real-time or near-real-time applications, such as charging applications, to personal computers and plug-in processor modules for less demanding, less time-critical applications, such as statistical analysis.

The first commercial application introduced on the adjunct processor is a charging information system for wireless communication services based on the digital advanced mobile phone service (D-AMPS) standard (IS-136/TDMA). Wireless networks that use this standard are being deployed in North and South America, Asia Pacific and Europe.

Just as the concept of the adjunct processor is an important part of the ongoing evolution of AXE, other applications will follow that further increase operators' ability to post-process revenue-earning data quickly and to streamline access for operation and maintenance.

Open systems and interfaces

The decision to base the AXE adjunct processor on an industry-standard computer reflects changing market requirements. Data-processing loads are increasing rapidly. But more than this, how data is being used is changing. For instance, the introduction of new, innovative subscriber services requires more complex charging schemes. Charging data must be processed and made accessible in real-time or near-real-time.

Increasingly, call-related data is being transmitted to external computer systems for subsequent processing. These external systems are generally open-standard (such as

TCP/IP), commercially available computer systems that comprise anything from a PC to a mainframe computer. In many cases, even the applications that run on these systems are provided by the computer industry and make extensive use of standard software packages.

Given the decision to distribute capacity-demanding data-processing tasks away from the AXE central processor, and to assign them instead to a dedicated computer, it makes sense that the dedicated computer should consist of commercially available systems. This approach permits network operators to use an external data interface that they are familiar with, thereby reducing their cost of ownership. In addition, the approach enables network operators to take full advantage of the economies of scale in the computer industry. This freedom to source computer hardware and some software from alternative suppliers helps cut initial and life-cycle costs to competitive levels.

Operators may also benefit from steady increases in processing power as well as from favourable cost trends and reductions in the physical size of computer equipment. In addition, new features can be introduced with shorter lead time, shortly after they are implemented and become available in the computer industry marketplace.

Another benefit of using industry-standard platforms is that networking engineers gain a standardised external data interface. Data from a network node is generally transferred to business data systems in a billing centre or in an operation and maintenance (O&M) centre by means of a wide area network (WAN) or by some other long-distance data communication system. A standardised interface in these mixed environments is a welcome feature.

System overview

Principles

The adjunct processor is integrated into the APZ. It is open, easy to use, and can be scaled for cost-effective solutions. The adjunct processor platform provides real-time services and processing power for a variety of applications, of which the formatting and output subsystem (FOS) is the first.

The three main principles that guided the first implementation of the adjunct processor platform are:

- connectivity – the adjunct processor must have flexible and open means for connecting to external systems;

Box A Terminology

AP	Adjunct processor. Part of the APZ.
API	Application program interface.
ASN.1/BER	Abstract syntax notation number 1/basic encoding rules (CCITT X.208/X.209). A language for describing an abstract information structure complemented with a standardised set of rules for encoding a representation of the abstract structure.
BSS	Business support system.
Corba	Common object request broker architecture. A protocol mechanism and naming or directory service for managing resources in a network by accessing and manipulating the representations of resources known as objects. These objects are data structures, commonly created using an abstract definition language; for example, interface definition language (IDL) specifications.
CP	Central processor. Part of the APZ. Control system in AXE.
D-AMPS	Digital advanced mobile phone service (same as American digital cellular, ADC). Name of digital wireless service using IS-136 time division multiple access (TDMA) technology.
DCE	Data ciruit-terminating equipment.
DMH	Data message handler according to Interim Standard 124 (TIA/EIA/IS-124).
DTE	Data terminal equipment.
FOS	Formatting and output subsystem. Application in the APT.
FTAM	File transfer, access and management. ISO/ITU OSI standard for file transfer.
FTP	File transfer protocol. De facto Internet standard for file transfer.
I/O	Input/output. A set of functions, of which the core are the basic I/O functions man-machine communication, data communication, file management, alarms, and central processor loading and backup.
IOG	Input-output group.
IOG11	I/O system 11.
IOG20	I/O system 20.
ISO code	International standardised interchange code where digits and characters are represented by the IA5 character set.
ITU-T	International Telecommunication Union - Telecommunications Standardization Sector.
JTP	Job transfer protocol. Bidirectional proprietary protocol for commands, alarms, and heartbeats between the central processor and the adjunct processor.
LAN	Local area network.
LBB	Large building block. Integrated computer system, including hardware and software, supplied by a vendor as a unit treatable as an Ericsson product.
MML	Man-machine language.
MTAP	Message transfer protocol AP. Proprietary protocol for high capacity transfer of data from the central processor to the adjunct processor.
O&M	Operation and maintenance.
ORB	Object-request broker.
OSI	Open systems interconnect.
OSS	Operations support system.
Packed ISO code	Interchange code where IA5 character set is used to encode characters, and BCD is used to encode digits.
RP	Regional processor. Part of the APZ.
RPC	Remote procedure call.
SP	Support processor. Part of the APZ.
SPIRIT	Service provider's integrated requirements for information technology, a team of international telecommunication service providers, vendors, and ISVs. SPIRIT blueprint: ISBN 1 85912 059 8 "Spirit Platform Blueprint" (SPIRIT Issue 2.0, Volume 1)
SQL	Structured query language. Used as a machine-machine interface to a relational database.
STOC	Signalling terminal open communication. RP bus-to-Ethernet converter.
TCP/IP	Transport control protocol/Internet protocol. The standard Internet protocol suite that provides network addressing and secure transport of data over a LAN-type network.
TDMA	Time-division multiple access.
Telnet	Internet standard application protocol on top of TCP/IP for terminal access.
TMOS	Telecommunications management and operations support.
TT	Toll-ticketing.
WAN	Wide area network.
X.25	X.25 is the ITU-T recommendation for the interface between data terminal equipment (DTE) and data circuit-terminating equipment (DCE) for terminals operating in the packet mode and connected to public data networks by dedicated circuit.
XPG4	X/Open Portability Guide 4. Specifies an operating system interface, commands, and APIs.

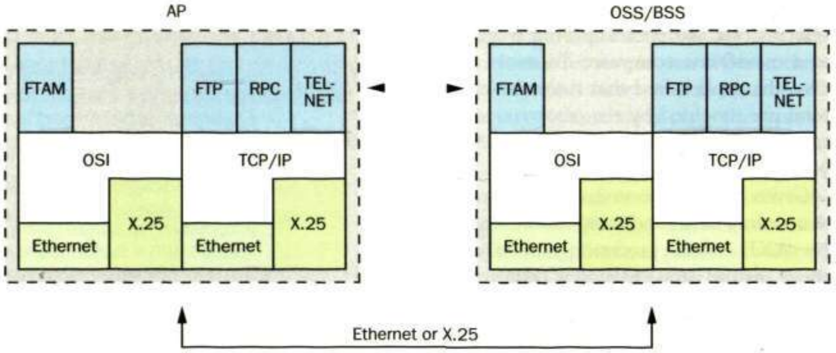


Figure 1
Overview of the AP to external system communication protocols. Note: this listing does not show all possible combinations of protocols.

- openness and system integration – the adjunct processor must be based on an open-standard computer system that can be integrated into the AXE system;
- external provisioning – the computer system must be supplied through an external source, integrated at a high level.

These main principles were supplemented with two additional principles:

- The adjunct processor must fulfil certain application-specific requirements (for the first application);
- The adjunct processor must embody an efficient internal development environment.

Connectivity

The main reason for arranging a connection to the adjunct processor is to enable network operators to extract data from their network nodes. A characteristic example is charging data. Connections may also be required for sending commands to applications that run on the adjunct processor platform, or to the adjunct processor itself, typically for operation and maintenance purposes (Figure 1).

File transfer

A main benefit of the adjunct processor in its first release is that it gives operators a secure method of remotely handling charging data files. This feature reduces the need for handling and storing tapes at the local level.

Data from the switching node's central processor may be collected, processed, and stored on disk in files that are ready to be sent to an external system using either of two protocols.

- The file transfer protocol (FTP) – the Internet standard.
- The file transfer, access, and management (FTAM) protocol – the open systems interconnection (OSI) standard.

Message transfer

In some situations, instead of collecting several messages in a file, an urgent, short data message must be sent to a post-processing system. When this is the case, the remote procedure call (RPC) protocol is used.

Depending on the design and the needs of an application, the adjunct processor is able to output data in response to operator-initiated requests. It can also automatically transmit files or other types of data to a billing centre or to some other location, according to a pre-determined schedule.

Operator access

Operator terminal access (alphanumeric terminals) from a remote system is supported using Telnet protocols. An important benefit of using industry-standard platforms is that operators who use Telnet to connect to the adjunct processor are greeted with a familiar environment; for example, the standard UNIX file system.

Data transmission

At lower protocol levels, data is transmitted using familiar local area network (LAN) or wide area network protocols:

- Transfer control protocol/Internet protocol (TCP/IP) and Ethernet enable LAN-type communication in local and large networks.
- X.25 is used for communicating in a WAN – X.25 is the ITU-T recommendation for the interface between data terminal equipment (DTE) and data circuit-

terminating equipment (DCE) for terminals operating in the packet mode and connected to public data networks by dedicated circuit.

Openness

To obtain the maximum benefit of using industry-standard platforms, and to protect investments made in designing applications, Ericsson selected for its adjunct processor platform a base of providers who offer the widest support of non-proprietary computer hardware and software. Consequently, the adjunct processor operating environment is based on:

- the open-standard XPG4 (X/Open Portability Guide 4, version 2), and the UNIX Implementation System V, Release 4 (SVR4);
- the SPIRIT blueprint (specification from service provider's integrated requirements for information technology).

The introduction of the adjunct processor in AXE is an example of how AXE is gradually evolving to become a more open system.

Internal development environment

A development environment, which consists of well-documented C and C++ application program interfaces (API) and a set of design rules, enables Ericsson application developers to provision new functions rapidly. The platform design is further supported by an array of open-standard, commercially available tools. The design rules, the APIs, and the tools assist designers in developing quality software that is largely system-independent; that is, the software may be ported to other systems if necessary.

AXE system integration

Obviously there is a trade-off between having an entirely independent, open-standard computer system and having a completely integrated system that cannot be distinguished from a proprietary one.

The purpose of the adjunct processor system is to enable external systems to connect to AXE without:

- allowing the AXE internal structure to limit the open characteristics of that connection;
- introducing adverse effects into switching tasks or into operation and maintenance.

To meet these requirements, the level at which the adjunct processor is integrated into other components in AXE has been set very carefully. In particular, the integration involved the following areas:

- communication channels for high-speed and high-volume data transfer from the central processor to the adjunct processor;
- bidirectional communication for commands, alarms, and supervision between the central processor and the adjunct processor;
- existing access methods for local and remote operation and maintenance;
- alarm handling, software upgrades, and process supervision (to provide the characteristics expected of a telecommunications network element).

Communication channels

The adjunct processor is connected by means of the regional processor bus (RP bus) to the central processor in AXE. Therefore, the adjunct processor coexists with other devices on the RP bus – in particular, with the input-output group (IOG) system (Figure 2). The connection of a standard computer system to the RP bus is implemented using a signalling terminal for open connections (STOC – hereafter called a signalling terminal), which translates RP bus communication to the Ethernet protocol.

The signalling terminal provides a high-capacity communication channel between the central processor and the adjunct processor. Several signalling terminals may be used to distribute load as well as to provide redundancy. The central-to-adjunct processor communication channel uses an Ericsson

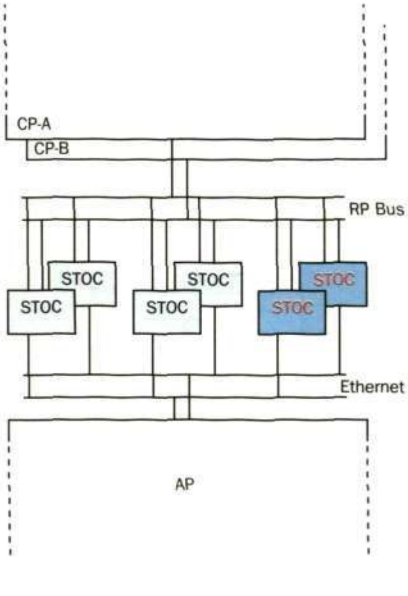


Figure 2
Outline of the hardware connection of the AP to the CP. The STOCs provide protocols conversion between Ethernet and the RP bus. This channel, which is used by the MTAP and JTP proprietary protocols over the standard TCP/IP protocol, provides the connection between the AP and the CP. Up to three pairs of STOCs can be used, depending on demands for capacity.

proprietary protocol (message transfer protocol adjunct processor, MTAP) that implements TCP/IP communication between the central processor and the adjunct processor through the signalling terminal.

In the current adjunct processor, the communication channel contains two Ethernet networks (for redundancy); thus, the signalling terminals are paired, one terminal for each Ethernet network. The number of paired signalling terminals (maximum three) is chosen to suit requirements for throughput (Figure 2).

Messages that are sent to the adjunct processor from the central processor are stored in a disk-based message store. The central processor keeps the messages in buffers until the adjunct processor acknowledges that it has received them. A message must be safely stored before an acknowledgement signal can be sent to the central processor, which then releases the message buffers. Once it has been stored, data in the message store is available to external applications.

Operation and maintenance

Integrating the adjunct processor into the AXE is not an end in itself. However, in terms of operation and maintenance, there are compelling reasons for doing so. For instance, integration enables operators to manage the adjunct processor from existing operations support systems (OSS). Also, an integrated computer system meets the operational requirement for a very minimum of local operation and maintenance.

Connections to the adjunct processor from an operations support centre are implemented through existing communication links to the IOG11/IOG20 (Figures 3 and 4). For local access, a terminal emulator may be used; for example, a PC-based local user interface to AXE. Access security to the adjunct processor is implemented using standard XPG4, UNIX System V, Release 4, security.

Alarm handling

As with all other equipment in an AXE installation, the adjunct processor system must be able to alert operators to important changes in its state, particularly if its hardware or software malfunctions. The adjunct processor system and associated applications contain functions that monitor themselves and generate event reports. Some functions generate alarms that are sent to the central processor using the proprietary

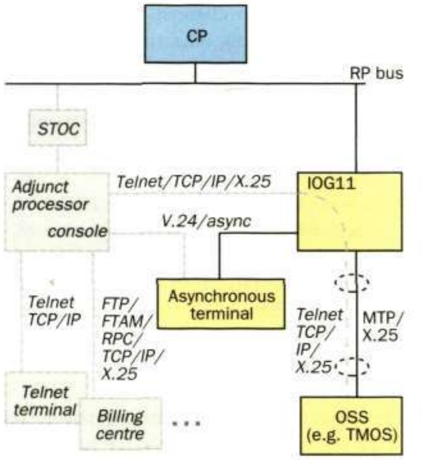


Figure 3
Operation and maintenance model for AXE man-machine language (MML) commands in a configuration with the AP and an IOG11 or IOG20. The highlighted parts are used with no changes from the presence of the AP (greyed-out parts).

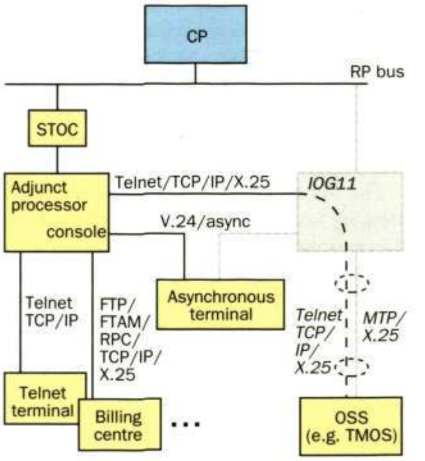


Figure 4
Operation and maintenance model for AP commands (UNIX commands). Access for operation and maintenance is provided transparently through the IOG11 or IOG20 from the same OSS. Other means of access to the AP are also indicated along with some protocols.

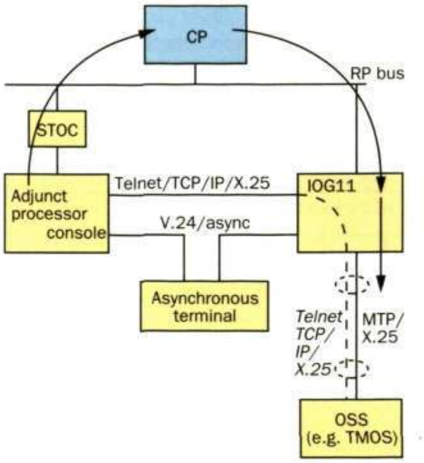


Figure 5
Arrows indicate the flow of alarms from the AP via the CP through the IOG11 or IOG20. There is no direct communication between the AP and the IOG. This ensures that AP alarms are treated as any other AXE alarm.

job transfer protocol (JTP) over TCP/IP and the signalling terminal. This adaptation allows adjunct processor-related alarms to be treated as normal AXE alarms and sent by means of the IOG11/IOG20 (Figure 5). Adjunct processor-related alarms are listed with all other AXE alarms. Each alarm is time-stamped by a clock that is synchronised with the clock in the central processor.

Process supervision

The system is designed to function with little or no preventive maintenance or other local management. Therefore, every process in the adjunct processor is supervised. A supervised unit, called a process group, may consist of one or more UNIX processes whose interdependence is specified during start-up. Should a supervised process die, then the process supervisor restarts the group to which the process belongs. If several attempts to restart a process group fail, then the adjunct processor system is rebooted. Rebooting the adjunct processor system does not have any adverse effects on traffic being switched.

Besides processes, log files are also maintained. Normally, the log files in a UNIX system must be removed by preventive maintenance. However, because preventive maintenance is not part of the adjunct processor operating environment, a log maintenance function has been included which ensures that the log files in the system are routinely maintained and deleted. The adjunct processor is designed to run with a minimum of operator intervention.

Software upgrades

The adjunct processor has been designed to operate with near-continuous availability during a software upgrade. The software upgrade facility ensures that no data is lost, and that traffic-executing tasks in the central processor are not disturbed.

A further feature of the software upgrade facility is that it enables the system administrator to supervise the operation of a system that has been upgraded. If the new system does not perform satisfactorily, the system administrator may revert to the software that was in use before the upgrade.

External technology provisioning

Most of the desired functions and characteristics of the adjunct processor, such as storage capacity, processing power, connectivity, and openness, are obtained through technology from the computer market. Nearly all externally provisioned products, which include the computer system and third-party products, are integrated into a large building block (LBB) before they are delivered to Ericsson (fully integrated to work with AXE). This allows the provisioning to be maintained at a very high level.

Ericsson and their external providers have worked out inter-company processes that facilitate:

- routines for ordering products;
- product maintenance and support throughout the lifetime of the product;
- the management of product enhancements and substitution.

Crucial characteristics are gained by integrating the adjunct processor into the AXE computer system; however, the greater part of the software is already integrated when delivered in the LBB.

One should remember that although the adjunct processor system opens AXE to external systems, the adjunct processor is not intended to be used as a general-purpose office system. The number of options (add-ons, plug-ins) has been kept to a minimum, in order to simplify site planning, ordering, support and maintenance.

Another benefit of using an externally provisioned, industry-standard system is that it is generally flexible and can be adapted more readily to subsequent (future) advances in computer system technology.

Specific first-application characteristics

The adjunct processor concept enables designers to use the commercially available processor that is most appropriate for any given application. To some extent, the initial implementation of the adjunct processor run-time platform was influenced by the first application to run on it, whose required characteristics include:

- near-continuous availability to the central processor;
- high throughput to the billing centre;
- seven-day retention storage capacity;
- safe collection and storage of call records.

Each of these characteristics is a high-end requirement that the adjunct processor concept was able to meet from the start. Of course, different applications will have different configurations, each of which takes into account important requirements for cost and footprint.

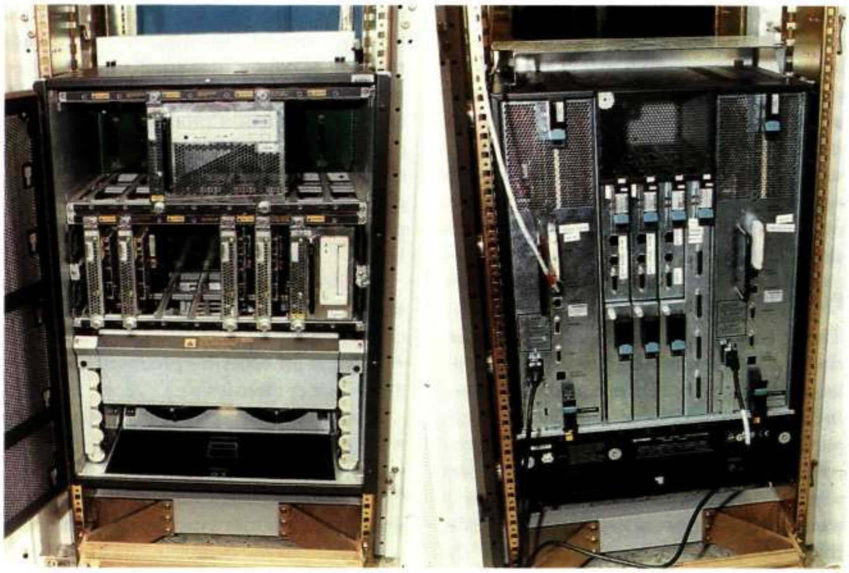


Figure 6
Front and back views of the AP hardware from Tandem Computers, suitable for such applications as the formatting and output subsystem (FOS).
Photo © Tandem Computers 1997.

First implementation, the formatting and output subsystem

The first commercial implementation of the adjunct processor is the formatting and output subsystem, which is a near-real-time charging application that requires the computing platform to be available on a near-continuous basis. The compliance requirements for a system used in central office applications necessitated that this application be a high-end system. Therefore, the adjunct processor system was first implemented on fault-tolerant hardware from Tandem Computers (Figure 6).

As its name implies, the formatting and output subsystem is a post-processing system that formats data so that it can easily be handled by other post-processing systems, such as a billing centre.

Before the formatting and output subsystem was introduced, the central processor of the AXE switch was tasked with formatting and outputting all charging data. After the data had been formatted, it was sent through the IOG11/IOG20 input/output system, which is based on the Ericsson proprietary fault-tolerant support processor (SP) in AXE. However, because the charging data that is required from mobile switches is increasing in volume, the load on the central processor is also increasing. Further, there is currently a growing demand for larger storage capacity and higher throughput from the AXE switch to billing centres. These characteristics are similar to those of a commercial charging system.

In the present implementation, raw charging data is collected by the formatting and output subsystem of the adjunct processor, where the data is safely stored in the adjunct processor message store (Figure 7). The input format of the call records is abstract syntax notation number 1 (ASN.1) using basic encoding rules (BER) encoding. The formatting and output subsystem extracts call records from the message store and decodes them. Depending on input values, data types are selected, formatted, and multicast to post-processing systems; that is, to business support systems such as a billing centre. The output encoding format may be toll-ticketing (TT) – as is currently received in the IOG – or some other format of toll-ticketing and call-event records. Supported formats include ISO code, Packed ISO code, ASN.1/BER, and DMH. ISO code and Packed ISO code are commonly used for toll-ticketing data. The encoding formats ASN.1/BER and DMH are more structured, making them more efficient.

Data may be distributed to multiple destinations using the FTP and FTAM protocols, or by means of the message-based remote procedure call protocol. The transfer of files is invoked either by the formatting and output application or by the post-processing system.

Hot billing is implemented by using the remote procedure call to output call records. The delay of records through the system for hot billing is less than 10 seconds.

As new formats are requested and introduced, the formatting and output subsystem will be adapted to make changing the format a simple matter of configuration. This will enable changes to reach the market more quickly.

Adjunct processor of the future

The adjunct processor concept is so open-ended and flexible that it is difficult to envisage any limits to its full potential. How it is implemented, however, will be determined by customer needs and market requirements. Nonetheless, even today several suitable areas for development have been identified.

Scaleability

New applications as well as those that can suitably be transferred from the central processor to the adjunct processor will require system characteristics that differ from the first implementation. One of the strengths of the adjunct processor concept is that it can be adapted to meet these different needs.

Many kinds of computer system may be employed as an adjunct processor. They may vary in size from a small, built-in or single-board PC or client-server system to a full-fledged, high-end system similar to the one used for the formatting and output application. The cost and footprint of a particular adjunct processor application depends on several requirements, including:

- system availability;
- central office compliance;
- disk storage capacity;
- safe storage through mirroring;
- throughput;
- the number of communication facilities specified.

Distribution

The first adjunct processor is a single-node computer system. This design enables network operators to protect the operation of critical revenue-earning applications by running less critical applications on a different host. Should a software error occur, then it will be limited to the host on which the software is running.

The design also allows operators to add applications gradually – matching the level of investment to growth in the network – rather than requiring them to make a large up-front investment in equipment whose capacity they do not need.

In today's world of public telecommunications, return on investment is a high-priority consideration. Although operators

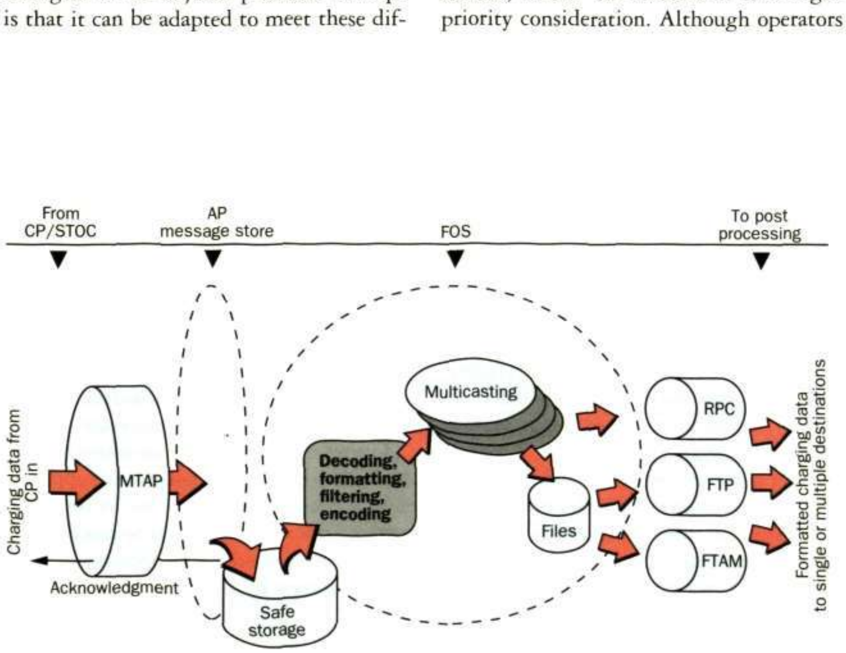
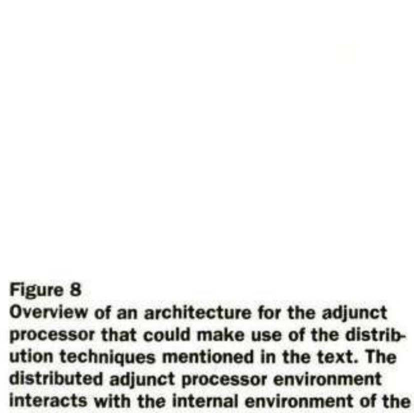


Figure 7
The flow of charging data through the AP for the formatting and output subsystem (FOS). The data, which is received in the AP from the CP via the high-capacity MTAP protocol, is stored safely, and acknowledged to the CP. The FOS accesses the stored data, processes it as indicated, and prepares it for dispatch to a post-processing system through any or several of the communication channels provided.

Figure 8
Overview of an architecture for the adjunct processor that could make use of the distribution techniques mentioned in the text. The distributed adjunct processor environment interacts with the internal environment of the switch, offering a service-oriented distributed machine-machine interface from the AXE network element. This interface is intended for local support systems *and* remote external systems. Two objectives for introducing such an interface are to facilitate the design efforts required for an application, and to facilitate the location of an application.


AXE network element
AXE/CP
Network element internal bus
Distributed environment for adjunct processor
Service-oriented distributed machine-machine interface
Possible support applications
Dedicated connection system-to-system
Operations support systems/ post-processing systems

must be able to add equipment to their computer systems, different applications put different requirements on their system resources; therefore the size and type of equipment they add vary. Nevertheless, the applications that run on each device need to share common resources and data. Finally, network operators want a unified view of the network element. They do not want to have to connect separately to a multitude of different nodes. In short, network operators want a heterogeneous, distributed system.

Open-standard versus integrated systems

The views on what data or applications should reside in a network element are changing. Some applications (new or existing) should reside on hosts that are only loosely coupled to an AXE network element. In other cases, the original data should not be in a network element at all, but should instead reside at a network level.

The architecture for distributing the system must take into account that access to network data needs to be transparent relative to the physical location of the data. Therefore, new and established mechanisms for distributing access to data are being considered. Established mechanisms include a database for persistent storage or for storage of semipermanent data that can be accessed through a structured query language (SQL) interface. By contrast, new methods of distributing access might address data using object-oriented techniques; for instance, using an object-request broker (ORB) such as Corba.

A management interface that uses ORB techniques could provide designers of management applications with an object-oriented view of the network element – with this perspective, designers must not possess in-depth knowledge of the inner structure of the network element in order to build their applications. A machine-machine communication mechanism of this kind represents a modern and efficient way of accessing input/output (I/O) functions either from a business support system (BSS) or from an operations support system in the operator's network (Figure 8).

A very attractive way of providing operator access is through Web technology. Web browsers give operators platform-independent access to data, allow them to connect to object-request brokers, and provide a full suite of functionality through applications that are built using Java techniques.

The adjunct processor exemplifies how AXE has been developed to evolve and embrace proven, standard, commercially available products – always with an eye to adopting maturing technologies that can quickly provide needed benefits. The foundation on which the adjunct processor concept was developed has shown such strength that, besides input/output, it may be applied in other parts of AXE as well.

Conclusion

The first release of the adjunct processor along with its first application – the formatting and output subsystem – has shown that the adjunct processor concept gives to AXE the welcome benefits of an open-standard computing platform.

The adjunct processor concept represents just one example of how AXE is taking on more and more open characteristics as it evolves to meet the needs of rapidly growing networks, new services, and the needs of future markets.

The drive to develop applications for the adjunct processor will come from operators who want to add new services to their networks in order to generate new revenues or from those who are being pressured to reduce their cost of ownership.

The first natural candidates for an adjunct processor application are basic I/O functions for man-machine communication, data communication, file management, and alarms, as well as loading and backup functions for the central processor.

A full array of I/O functions on the adjunct processor, together with functions for computing and for gathering statistics, fulfils the basic requirements for openness and distribution to an operations support system.

Near-real-time data may be provided for fraud prevention, hot billing, subscriber analysis, or subscriber crediting.

Applications that require near-real-time data processing – to protect the flow of revenue, or to enable telecommunications networks to process information more rapidly – are candidates for a full or partial implementation on the adjunct processor.

Thanks to more open characteristics in AXE, decisions to implement such applications as fraud prevention or hot billing on a particular host will be easier to reach.

The optimum host of an application may be determined by the total network solution.

New patents within Ericsson

INTEGRATED MULTIPLICATOR
Evald Koisalu
Patent number 504371

ERROR CONCEALMENT ALGORITHM
Karim Jamal, Fredrik Jansson
Patent number 9403386-7

DISTURBANCE 2
Håkan Andersson, Magnus Madfors, Bengt Persson, Håkan Eriksson, Krister Raith
Patent number 5594949

MDF CONNECTION BLOCK
Sture Roos
Patent number 504425

GENERATOR FOR CLOCK AND DATA
Tord Haulin
Patent number 9501608-5

SW STRUCTURE FOR PATH OBJECT
Per Israelsson
Patent number 504050

MODULO-3 OPTICAL CABLE
Odd Steijer
Hans -Christer Moll
Bengt Lindström
Patent number 504426

DELTA QUANTIZATION
Ylva Timner
Patent number 9501640-8

PEAK CIRCUIT FOR LED
Gunnar Forsberg
Patent number 9501821-4

POINT-TO-MULTIPOINT CONNECTION
Lars Novak
Staffan Andersson
Torgny Lindberg
Erik Bogren
Patent number 9501073-2

AUTOMATIC CONF.FEATURE
Vladimir Alperovich
Patent number 5559876

WIRELESS EARPIECE
Nils Rydbeck
Patent number 5590417

CALL FORWARD/LOOK AHEAD
Karen Cook-Hellberg, Amelia Noriega, Kathleen Angerer
Lorie Presto-Railey, Prafulla Shintri, Susanna Adam, Suzy Vasa, Vladimir Alperovich
Patent number 5530931

C C TIMING DETECTION
Thomas Brown
Patent number 5594761

INTEGRATED ANTENNA/MICROPHONE
Seung Kim
Patent number 5555449

ENCRYPTING CODEC
Daniel Schwed
Patent number 5592556

FAST A G C
Paul Dent
Patent number 5568518

PSUEDO CLOCK IN CELLULAR PHONE
Michael Fehnel
Patent number 5590092

EXTEND BATTERY LIFE
Raymond Henry
Patent number 5590396

ANTENNA APPARATUS
Nils Rydbeck
Patent number 5590416

TRUNKED RADIO REPEATER SYSTEM
Jeffrey Childress, Marc Dissosway, Gerald Copper, Houston HughesIII
Patent number 5574788

AM-FM POWER AMPLIFIER
Paul Dent
Patent number 5570062

EFFICIENT PAGING SYSTEM
Paul Dent
Patent number 5594776

MULTIPLE ACCESS CODES 2
Gregory Bottomley, Paul Dent
Patent number 5550809

POWER EFFICIENCY
Phillipe Charas, Paul Dent
Patent number 5548813

KEY TRANSFORMS DISCRIMINATE
Paul Dent, Krister A Raith
Patent number 5594795

IMPROVED FREQUENCY RE-USE
Paul Dent
Patent number 5594941

SELECTIVE RE-SYNCHRONIZATION
Krister Raith, Paul Dent
Patent number 5546464

WASTE ENERGY RECOVERY
Paul Dent, Ross Lampe
Patent number 5574967

SELF CAUITY PARKING
Daniel Dulong, Richard Brunner
Patent number 5530921

PCS COMMUNICATIONS MODEL
Yves Lemieux
Patent number 5594739

HLR FOR MANUAL VISITORS
Viet Nguyen
Patent number 5564068

CHANNEL-INDEPENDENT EQUALIZER
Paul Dent
Patent number 5557645

TDMA/FDMA/CDMA HYBRID ACCESS METHODS
Paul Dent
Patent number 5539730

WASTE ENERGY RECOVERY
Paul Dent, Ross Lampe
Patent number 5574967

NTH BEST DECODER
Paul Dent
Patent number 5577053

DOWNLINK CDMA-SIGNALS
Paul Dent, Gregory Bottomley
Patent number 5572552

DIRECT UPDATE EQUALIZER
Gregory Bottomley, Sandeep Chennakeshu, Paul Dent, David Koilpillai
Patent number 5577068

SUPERFRAME STRUCTURE
Håkan Andersson, Bengt Persson, John Diachina, Krister Raith, Anthony Sammarco, Francois Sawyer
Patent number 5604744

IS-7X PATENT PROGRAM
Krister Raith, Bengt Persson, Anthony Sammarco, Anders Hoff , John diachina, Joseph Turcotte, Håkan Andersson, Francois Sawyer, Patrice Marsolais, Roland Bodin
Patent number 5603081

AUTOMATIC FOLLOW-ME
Johan Falk, Björn Ahlberg, Anders Mölne
Patent number 5600704

AUTOMATIC ANSWER
Björn Ahlberg, Anders Mölne
Patent number 5570413

DELAYED DISCONNECTION
Björn Ahlberg, Anders Mölne, Johan Falk
Patent number 5574774

MULTIPROSESSOR RAM SHARING
Paul Dent, Alf Larsson
Patent number 5598575

DATA DRIVEN CONTROL
Ingemar Gard, Stefan Larsen, Göran Eneroth, Tord Nilsson
Patent number 5513127

TEST TRANSCEIVER
Ulf Hagström, Magnus Isaksson
Patent number 5613217

SMART SCAN
Sven Ryding, Hans Blackman
Patent number 5613208

PCB AIR COOLING
Lars Nygren, Jan Wennerberg, Lars Bertilsson, Uno Dahl
Patent number 504430

DIFFERENTIAL DEMULTIPLEXOR
Mats Bladh
Patent number 504533

RIBBON FIBRE FANOUT
Anders Sjöberg, Kristian Engberg
Patent number 504587

DIFFERENTIAL MULTIPLEXOR
Mats Bladh
Patent number 504521

CHANNEL RESERVATION
Per Beming, Dalipor Turina
Patent number 504577

DIFFRACTION LIGHT
Sasan Esmaeili
Patent number 504588

INDUCTIVE COMPONENT
Arne Lindqvist
Patent number 504592

SWITCH WITH INDUCTOR FUNCTION
Arne Lindqvist
Patent number 504591

REINFORCING CAPSULES
Karl-Erik Leeb
Patent number 504623

RECEIVER DEVIATION CORRECTION
Dan Weinholt
Patent number 504341

Contents

Previous issues

No. 2, 1996

No. 3, 1996

No. 4, 1996

No. 1, 1997

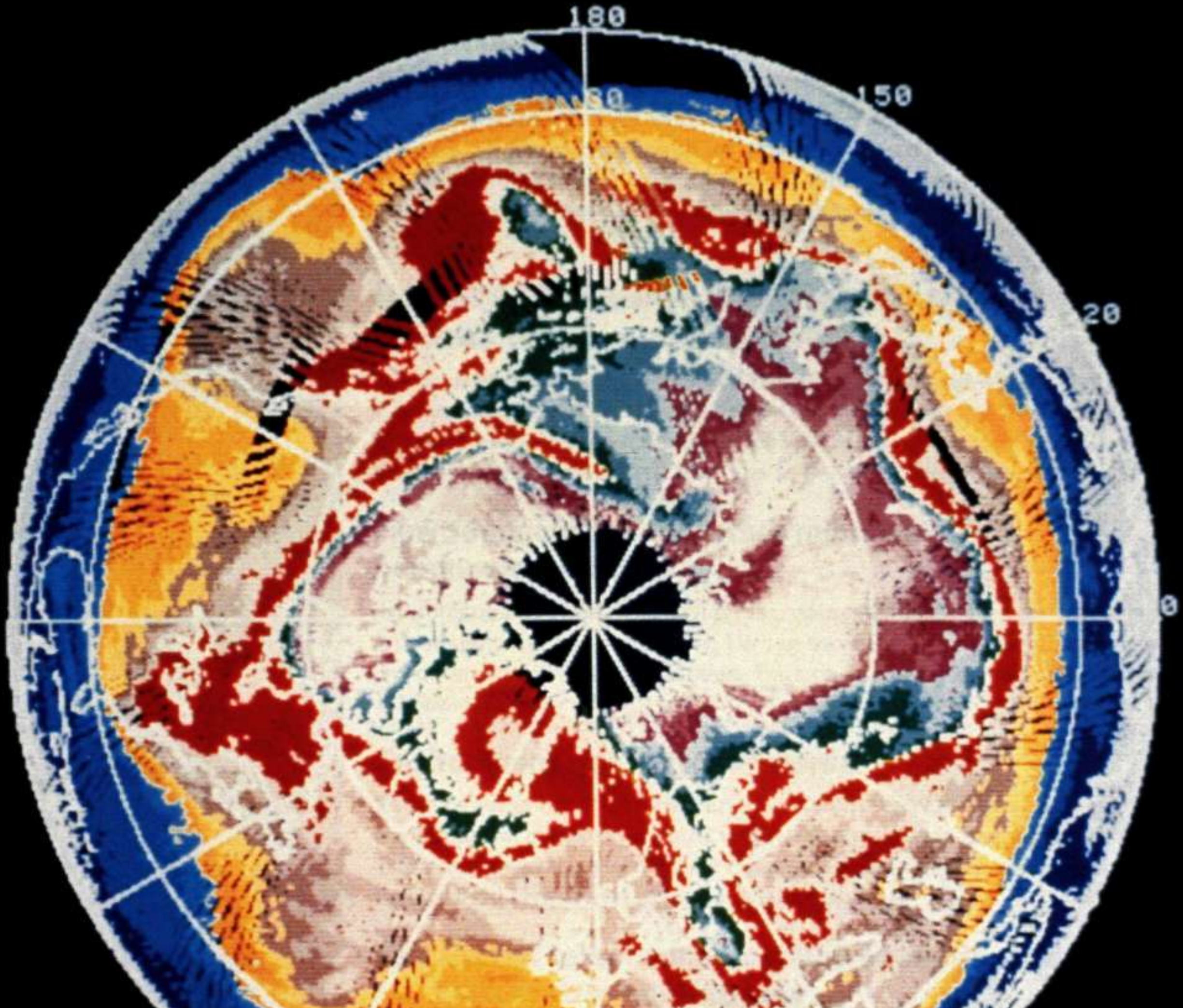
Telefonaktiebolaget L M Ericsson
SE-126 25 Stockholm, Sweden
Tele: +46 8 7190000
Fax: +46 8 6812710

ISSN 0014-0171
Ljungföretagen, Örebro 1997

Ericsson REVIEW

THE TELECOMMUNICATIONS TECHNOLOGY JOURNAL

NO 3, 1997



ERICSSON

Ericsson REVIEW

The telecommunications technology journal

The purpose of Ericsson Review is to report on the research, development and production achievements made in telecommunications technology at Ericsson. The journal is published in English and Spanish and distributed quarterly to readers in more than 130 countries.

Address:
Telefonaktiebolaget LM Ericsson
S-126 25 Stockholm, Sweden

Tel: +46 8 719 00 00

Fax: +46 8 681 27 10

Internet address:
http://www.ericsson.com

Publisher: Håkan Jansson

Editor: Eric Peterson

Production manager: Eva Karlstein

Layout: Paues Media, Stockholm

Printer: Ljungföretagen, Örebro

ISSN: 0014-0171

Volume: 74, 1997

Cover: North Pole total ozone map. Being an environmentally progressive company, Ericsson is the first telecommunications corporation in the world to test life-cycle stressor-effect assessment methods.

Contents

Contributors

In this issue

Olle Källström

Magnus Nielsen

Tord Stureborg

Mats Ek

Martin Löfgren

Herbert H.G. Zirath

Mats-Olov Hedblom

Magnus Ekhed

Olle Källström is employed as Strategic Product Manager of GSM data communication at Ericsson Radio Systems AB. He joined Ericsson in 1995, but has worked in the cellular industry as a telecommunications consultant since 1985, gaining far-reaching experience of operational engineering, product management, and marketing.

Magnus Nielsen is Manager of New Technologies, Network Engineering for AMPS/D-AMPS at Ericsson Radio Systems AB. His group is responsible for switch planning and for developing methods, tools, and services for radio frequency engineering and transmission engineering. He holds an MSc in electrical engineering from the Royal Institute of Technology, Stockholm.

Tord Stureborg currently works as Program Manager in the Network Performance Programs department at Cellular Systems, American Standards, Ericsson Radio Systems AB. He began working with cellular systems in 1986, as manager of implementation projects and marketing, Oceania. Since 1992 he has worked with the introduction of, and performance issues relating to, D-AMPS. He holds a BSc in electronic engineering.

Mats Ek is System Design Manager of the mobile base station subsystem in CMS 88 and CMS 30. Since joining Ericsson in 1983 he has worked in Sweden and in the US as a systems design engineer. He has also served as a manager of digital hardware design. He received an MSc in electronic engineering from the Lund Institute of Technology.

Gunnar Genell, who joined Ericsson in 1994, is Product Manager of radio network products at Ericsson Radio Systems AB. Before joining Ericsson he worked with software development, project management and international marketing. He holds an MSc in electrical engineering from Chalmers University of Technology and an MBA from the School of Economics and Commercial Law in Göteborg.

Dag Jungenfelt is responsible for the Microwave Access range of products in the Microwave Communications Division of Ericsson Microwave Systems AB. He has worked in systems design for the MINI-LINK series of radio relay products since joining Ericsson in 1991. He holds an MSc in electrical engineering from Chalmers University of Technology in Göteborg.

Martin Löfgren joined Saab Ericsson Space AB in 1993, where he worked with microwave circuit design. In 1996 he moved to his current position at the High-speed Electronics Research Center at Ericsson Microwave Systems AB, where he works with IC design and ATM WLAN system-realization issues. He holds an MSc in engineering physics and a PhD in microwave electronics from Chalmers University of Technology, Göteborg.

Herbert H.G. Zirath, who has served as a part-time technical advisor to Ericsson Microwave Systems AB since 1995, holds MSc and PhD degrees in electrical engineering from Chalmers University of Technology, Göteborg. In 1980, he began working at Chalmers as a researcher of cooled millimeter-wave Schottky diode mixers and the properties of millimeter-wave Schottky barrier diodes. In 1986, he became responsible for the development of active millimeter-wave devices and MMICs, such as InP- and GaAs-based MESFETs and HEMTs, including their modeling and related circuits – active and resistive HEMT mixers, low-noise amplifiers, and harmonic generators for frequencies up to 120 GHz. In 1996, he was named Professor at the Department of Microwave Technology at Chalmers.

Mats-Olov Hedblom has held the position of Environmental Manager at Telefonaktiebolaget LM Ericsson since 1995. Prior to joining Ericsson, he amassed extensive experience of environmental management, business management, research, and engineering while working in the fields of environmental analysis, environmental and chemical technique, bleaching, chemical performance, and product development. He holds a PhD in organic chemistry and biochemistry awarded by Uppsala University.

Magnus Ekhed is Product Manager of the TMOS IntraWeb Gateway in the Implementation Services organization at Ericsson Hewlett-Packard Telecommunications AB. His primary responsibility, since joining

From the editor

Eric Peterson

Gunnar Genell

Dag Jungenfelt

Peter Gundersen

Olav Queseth

EHPT in 1994, has been to manage customer projects. He holds an MSc and an MBA from Chalmers University of Technology, Göteborg.

Peter Gundersen is responsible for customizing and implementing operations support systems at Ericsson Hewlett-Packard Telecommunications AB. He has worked in system design and project management since joining Ericsson in 1986. He has also participated in creating the development environment for embedded avionics applications. He holds an MSc in computer science and engineering from Chalmers University of Technology, Göteborg.

Olav Queseth is responsible for systems on the TMOS IntraWeb Gateway development team at Ericsson Hewlett-Packard Telecommunications AB. Before joining the team in 1994, he worked with the customization, integration, and installation of TMOS. He holds an MSc in computer science and engineering from Chalmers University of Technology in Göteborg.

Each quarter, we have the privilege of reporting on new products and developments in research at Ericsson. However, only a tiny fraction of many possible articles makes its way into the pages of this journal.We feel a keen responsibility to select and publish articles that describe the most important technological happenings at Ericsson. Over time, the editorial board and staff will be considering various ways of including more articles in Ericsson Review – in this issue, for example, we are pleased to include six articles, compared with the usual five.

One of the most satisfying aspects of my work is that nearly every new development we report on contributes towards making the world a more open, democratic society. In particular, mobile telephony and data communication help set people free, enabling them to stay in touch with family, friends and work while on the move. Ericsson's new GSM Internet/intranet direct-access solution, for example, introduces an access server at the GSM exchange site, creating a direct end-to-end digital connection that bypasses the PSTN. This solution gives mobile phone users universal access to the Internet and to corporate intranets. Needless to say, it may also be applied in other kinds of digital wireless network, such as a D-AMPS network – all the more reason for operators of wireless AMPS networks to migrate to digital wireless services. In this issue you can read, step-by-step, how digital migration is achieved.

Even with the increased capacity offered by digital networks, high concentrations of people using their mobile phones to carry out a broad range of tasks create congestion and increase demand for bandwidth. Solutions to these problems are at hand, however. For instance, by introducing a hierarchical cell structure into their networks, operators can provide support for more subscribers and greater traffic volume without the need of additional radio spectrum. Hierarchical cell structures also enable operators to extend the reach of their networks into locations where it has previously been difficult to guarantee satisfactory wireless access. Ericsson's newly expanded family of RBS 884 radio base stations covers every requirement of a hierarchical cell structure in wireless D-AMPS networks.

The implementation of new network configurations, such as a hierarchical cell structure, shortens the distance between adjacent base stations. This puts new requirements on the transmission network needed to backhaul base stations to switching networks. Fortunately, Ericsson's microwave division has developed monolithic microwave integrated circuits (MMIC), which they will use to reduce the size and improve the performance, reliability and producibility of their MINI-LINK products. In addition, they are applying the MMIC technology to develop fixed broadband radio access and ATM-based wireless LAN systems, which will satisfy demands for bandwidth in wireless broadband systems.

The technology we report on also enables people to stay put. For instance, one of my "office colleagues" resides in Virginia, while the rest of us live and work in Stockholm. The solutions described in this and coming issues of Ericsson Review will enable a great many people to live where they want, regardless of where their company offices are located. Companies, on the other hand, need not maintain as many offices if fewer people commute to work each day. Moreover, with fewer people commuting, the roads will be less crowded, which means that discharges of carbon dioxide (CO_2) and nitrous oxides (NO_x) will decrease. With scientists telling us that the developed economies of the West must reduce their share of environmental load by 10 to 20 times present levels within the next 50 years, we begin to understand that these solutions are not only convenient, but also gravely necessary. Happily, we learn that by applying information technology we can lower the total environmental load. Of course, a company's environmental responsibility includes more than just permitting its employees to telecommute. Manufacturers, for example, must begin designing for the environment, and reclaiming products at the end of their life. In a special three-part article beginning in this issue, we are pleased to show that Ericsson is doing just that.

To round off the issue, we are excited to introduce the TMOS IntraWeb Gateway, which exploits Web technology to enable persons outside an operation and maintenance center – for example, customer-care and marketing departments – to access important network-related information in near real-time.

Eric Peterson
Editor

Internet and intranet connections over GSM

Olle Källström

Despite early predictions of an explosive growth in the use of mobile data communication, actual results have been quite moderate. Today, however, driven by the business sector's increasing need of access to the Internet and corporate intranets, the demand for mobile data services is growing. With its strengths of a global footprint, international roaming capabilities, reliable connections, secrecy and encryption, GSM is an ideal bearer of mobile data communication.

The author describes how Ericsson's GSM Internet/intranet direct-access solution – whose implementation benefits end-users, companies, and network operators alike – overcomes the fundamental problems of connectivity between GSM and the Internet.

The data communication segment of the wireless GSM market is poised to compete for mobile data customers. Many countries already have more than one dedicated wireless data operator, and as the GSM data communication roll-out moves ahead, competition will intensify. By the end of the decade, a significant portion of the mobile subscribers around the world is expected to work with Internet-based wireless data communication applications. In some cases, these applications will be integrated with ordinary voice services; in others, they will consist of dedicated data solutions.

Ericsson is exploiting the potential of wireless Internet access markets for GSM-based cellular networks. Their expertise in digital wireless technology will enable operators to develop wireless versions of electronic mail (e-mail), Internet and intranet access, information services, and many other functions. With the introduction of the GSM Internet/intranet direct-access solutions discussed here, wireless data access becomes the perfect complement of many applications. Some applications will be hosted by GSM operators, whereas others will simply use the wireless transport mechanism. In either case, GSM operators will gain increased revenues from the services they sell as well as from increases in the use of air time.

Once the fundamental problems of connectivity between GSM and the Internet are solved, it is anticipated that a creative surge within the Internet community will generate a wide array of GSM and data services and applications. Internet access will become a basic feature of GSM systems. Operators will then turn their focus away from plain access services to the potential of value-added services.

Box A
Abbreviations

GPRS	General packet radio services
GSM	Global system for mobile communication
HSCSD	High-speed circuit-switched data
ISP	Internet service provider
IT	Information technology
LAN	Local area network
PCMCIA	Personal computer memory card international association
PDA	Personal digital assistant
PPP	Point-to-point protocol
PSTN	Public switched telephone network
TCP/IP	Transmission control protocol/Internet protocol
VPDN	Virtual private dial-up network

The Internet

Today most people have heard of the Internet – the global communication network that allows people to communicate, cheaply and nearly instantaneously, with individuals and information resources all around the world. Like the telephone system, which is essentially a collection of telephones connected by a worldwide communication network, the Internet is a collection of computers that are also linked together by a global network. The computers communicate with one another by means of accepted standard protocols.

Once users have connected to the Internet, they may access many useful services; for example, they may

- send and receive e-mail;
- transfer files;
- browse and retrieve information;
- communicate socially;
- gather news.

Individuals and organizations also conduct business over the Internet: schools, government agencies, industrial and commercial organizations, and political parties all connect to and communicate over the Internet on a daily basis.

No one has a monopoly on access to, or on the use of, the Internet. This is because the Internet has no central computer. Instead, the many networks and computers that make up the Internet act as peers in communicating and exchanging information.

One reason why the Internet is so successful is that its developers have committed themselves to producing open standards. Also, the specifications, or rules, that computers need in order to communicate amongst themselves are free and publicly available to anyone who wants them. The standards used on the Internet are known as the TCP/IP protocol suite. TCP/IP, which stands for transmission control protocol/Internet protocol, is the "language" of the Internet.

Access to the Internet is usually provided by an Internet service provider (ISP), which may be a telecommunications operator, such as Telia, BT, AT&T, or an independent organization, such as CompuServe. As the name suggests, an Internet service provider supplies – for a fee – such Internet services as e-mail, file transfer, access to the World Wide Web, and so on.

Intranets

The technology that was used to create the Internet has also been applied within public and private organizations; for example, to disseminate corporate information quickly and inexpensively to a widely spread workforce. An intranet is the equivalent of a miniature Internet within the boundaries of an organization's information technology (IT) environment. It may be accessed by anyone within the organization whose computer is connected to the corporate data network. The most common intranet applications provide access to

- business intelligence;
- catalogs and price lists;
- corporate e-mail systems;
- corporate phone directories;
- corporate policies and news;
- document databases;
- employee bulletins;
- job postings;
- job procedure manuals;
- order status and tracking;
- product bulletins.

Access to a corporate intranet is generally restricted to company employees. Very often, a connection to the Internet is provided from an intranet. These connections are usually made through a security barrier (such as a proxy server or a firewall) that prevents persons outside the company from obtaining access to confidential corporate information. Nonetheless, some employees are authorized to access their company network remotely; for example, from a laptop computer over fixed or mobile telephone networks. In this way, while they are away from the office, employees may continue to access up-to-date corporate information, and exchange e-mail, documents and faxes with customers and colleagues. Remote access to a company intranet boosts customer service, improves productivity and infuses the workforce with new elements of flexibility.

Delayed reception of mobile data communication

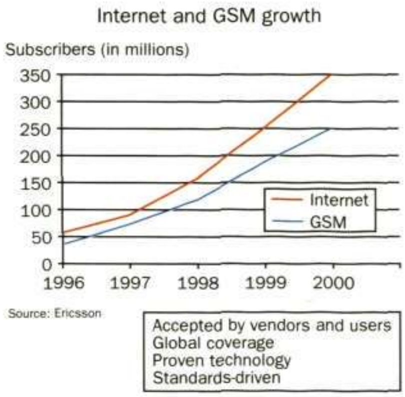
For some time, analysts have predicted an explosive use of mobile data communication. Actual growth in this area has been disappointingly slow, however. Current figures (1996) show that the number of data subscribers constitutes less than one percent of the total number of GSM subscribers. An industry analyst suggests that despite a high adoption rate for portable computers, many companies are failing to tap into the true potential of their equipment. In Europe, for example, less than half the companies using portable computers support remote communication – including e-mail applications. In the US, the number is lower still, at less than 20%. The main objections to GSM data communication are low data transfer rates, cost, and a shortage of applications.

Low data transfer rates

The maximum data transfer rate of 9.6 kbit/s over GSM compares unfavorably with typically 33.6 kbit/s for fixed modem access and even higher transfer rates over corporate networks. To counter this situation, software vendors are now adapting their applications for wireless transmission, optimizing the transfer protocols for accessing Web pages, and removing graphical information from Web pages in order to reduce the amount of data that must be transferred. However, until new higher-speed wireless data services are offered, bandwidth will remain a limiting factor.

Cost

The entry cost of mobile data communication is high: user equipment includes a portable computer or a personal digital assistant (PDA), a PCMCIA card, and a mobile phone that is capable of data communication. Although the smart-phone concept – a combined mobile phone and PDA – simplifies mobile data communication, the cost of this type of product will proba-



Source: Ericsson

Figure 1
Internet and GSM technologies–which are proven, standards-driven technologies–are accepted by vendors and enjoy global coverage. The market forecast shows that the number of Internet and GSM subscribers will continue to grow very rapidly.

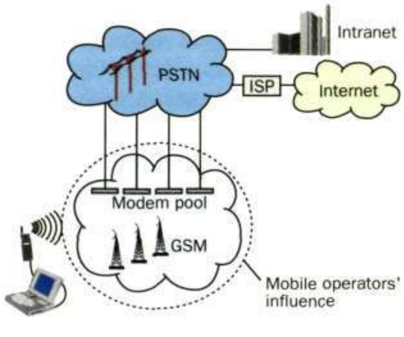


Figure 2
Internet/intranet access from GSM through the PSTN. Currently, it takes from 20 to 40 seconds to establish a modem connection through the PSTN. Also, operators–who are charged for the PSTN interconnect–have little influence in creating value-added services.

bly remain beyond the reach of the average consumer for some time. Also, compared with data calls over the public switched telephone network (PSTN), which are commonly charged at local call rates, the cost of GSM data calls is still high.

Shortage of applications

With any new technology, it takes time before large numbers of people begin using it. In the case of mobile data communication, this is partly due to

- a lack of knowledge of the technology and related applications;
- cost;
- a shortage of suitable applications specifically aimed at the mobile user.

Today e-mail makes up the largest segment of GSM data. With the growing number of intranets, however, more and more people will be accessing the information on their corporate networks. Nonetheless, a "killer application" – that is, a single service or application that suddenly convinces masses of people to use, in this case, their mobile phones for data communication – has yet to surface. In all likelihood, usage will increase gradually as services are improved and as more people become aware of them.

The turning point – Internet and intranet connections over GSM

The demand for local desktop and remote access to intranet services is growing. Over time, as more employees carry laptop computers and GSM mobile phones, and as GSM-compatible PDAs and hand-held computers become commonplace, the GSM network will become an everyday medium for linking remote users to the Internet and company intranets.

Until recently, however, the only way remote users with a laptop computer-mobile phone combination could access the Internet or a corporate network was through the PSTN, using modem pools in the GSM network and at the corporate site. Accordingly, data signals had to be converted from digital to analog (to pass through the PSTN) and then back to digital. The hardware required to make these conversions is expensive, and call set-up is slow: establishing a modem connection to the PSTN can take from 20 to 40 seconds. Moreover, GSM operators are charged for their use of the PSTN (Figure 2).

The Ericsson Internet/intranet direct-access solution introduces an access server point at the GSM exchange site. The server terminates GSM data calls directly, eliminating the need for a modem pool (Figure 3). Thus mobile users bypass the PSTN, obtaining direct end-to-end digital connection. In addition, the call set-up time is reduced to under ten seconds. And as an added benefit, GSM operators avoid having to pay for PSTN connections.

Today's GSM circuit-switched data services support asynchronous and synchronous 9.6 kbit/s analog modems (V.32) with or without V.42/V.42bis data correction and compression and unrestricted digital information (V.110).

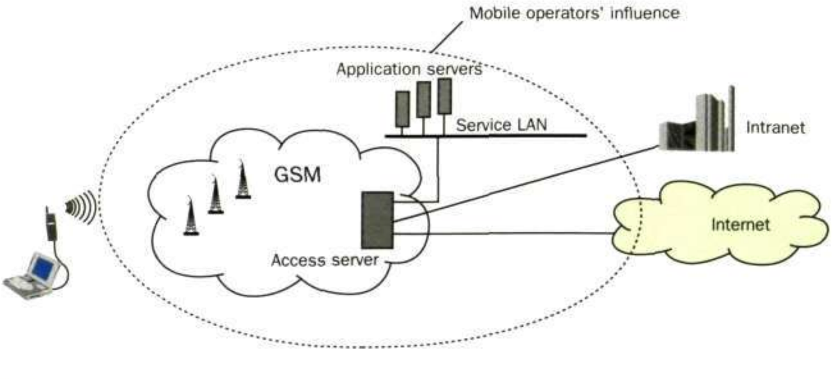
The operator as the ISP

Because the GSM Internet/intranet direct-access solution terminates data calls within the GSM network, operators gain new opportunities to offer data applications and services targeted at specific market segments (Figure 4). For instance, they might build up a service LAN to which they can connect an array of servers (supplying e-mail and access to the World Wide Web). Some operators may even become ISPs in their own right, generating revenue from the interconnection and from the services they provide.

Universal access

Since the access server terminates PSTN and GSM data calls in the same chassis, users retain fixed access to the services provided by the mobile operator. To access these services, users simply dial in through the PSTN (Figure 5). Subscribers who roam into another

Figure 3
Direct access to the Internet/intranet. A centralized access point terminates mobile users' GSM data calls, bypassing the PSTN. This yields a direct digital end-to-end connection with shorter call set-up time.



country use the local GSM network. However, in areas where a direct link to the home GSM network is not available, calls are routed through the PSTN and terminated by modems within the access server.

As mentioned above, until now access to corporate intranets from a GSM network has been achieved by connecting a laptop computer to a GSM phone and dialling into the company modem pool. The direct-access server, however, provides a better, less-expensive solution that benefits end-users, corporate customers and mobile operators alike. Drawing upon solutions from external partners, the GSM Internet/intranet direct-access solution has been developed as a supplement to Ericsson's mobile business product portfolio. Box B describes the different methods of connecting from a GSM network to a corporate intranet.

Virtual private dial-up networking and tunneling

The direct-access solution uses virtual private dial-up networking (VPDN) techniques. The goal of a VPDN service is to permit many separate, autonomous protocol domains to share a common access infrastructure. A key feature of VPDN services is tunneling (Figure 6). Tunnels act as vehicles for encapsulating packets inside a protocol that is understood at the entry and exit points of the network. Tunneling permits multiple protocols to be transported. It also permits unregistered addresses to be used over shared or public networks, thereby allowing users to gain secure access to their corporate gateway. The following example describes what happens when a remote user initiates access using a VPDN service:

1. The remote user dials the access point through the GSM network.
2. The access point accepts the connection and establishes a point-to-point protocol (PPP) connection.
3. The direct-access server uses a local security server to identify the remote user – the user name indicates whether or not a virtual dial-up service is required. User names must be structured, specifying a corporate gateway endpoint (for example: lars.larsson@ericsson.se).
4. If a tunnel connection to the desired corporate gateway does not currently exist, then the access server initiates one.
5. The direct-access server allocates an identification number to the session (several dial-up sessions may coexist in a single tunnel).
6. The corporate gateway authenticates the remote user and accepts or rejects the tunnel.
   Note: Corporations impose their own security and policy on the remote users who access their networks. Thus an organization need not rely on the authentication that was performed at the direct-access server.
7. The corporate gateway exchanges PPP negotiations with the remote user.
8. At this point, the corporate gateway may assign the remote user an IP address, which eliminates the need for service-provider addressing.
9. Once a virtual link has been established, end-to-end data is tunnelled in both directions between the remote user and the corporate network.

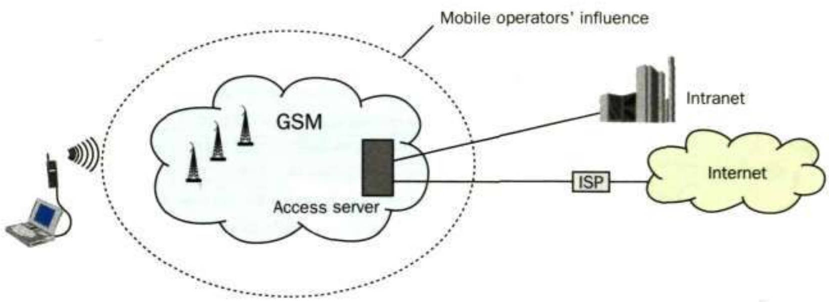
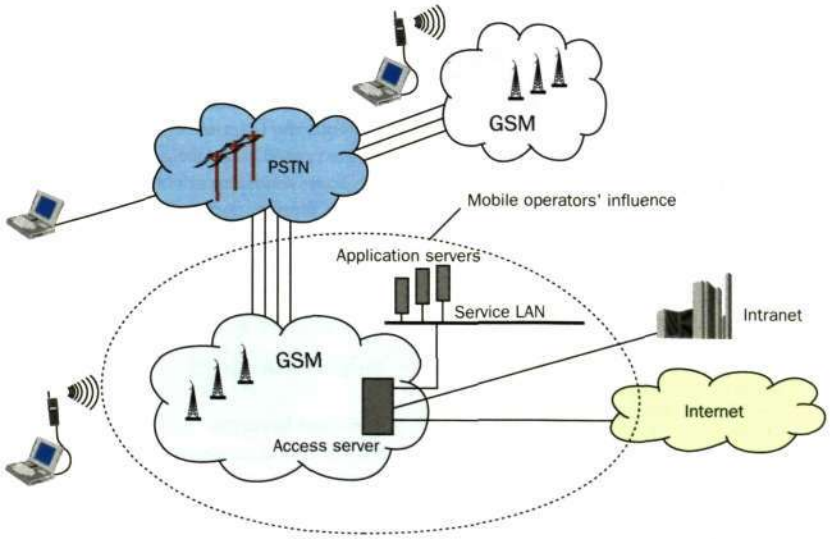


Figure 4
The mobile operator as the Internet service provider (ISP). Data calls are terminated within the GSM network, which enables operators to become ISPs in their own right, targeting data applications and value-added services to specific market segments.

Figure 5
Fixed, mobile and roaming access. Subscribers who roam into another country use the local GSM network. In areas where a direct link to the home GSM network is not available, calls are routed through the PSTN and terminated in the access server.



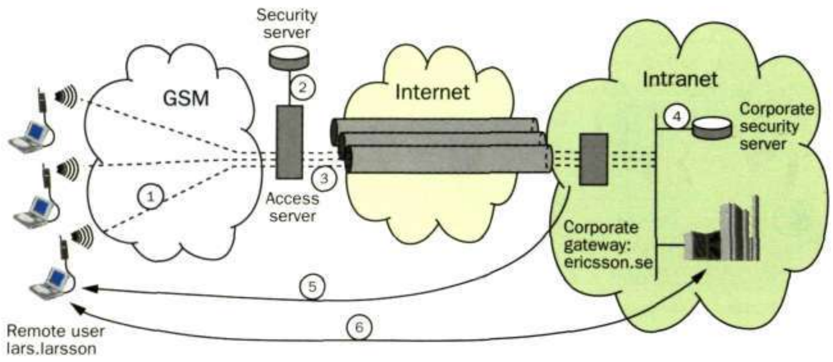


Figure 6
The direct-access solution uses virtual private dial-up networking techniques. This service, which provides secure intranet access, permits many separate, autonomous protocol domains to share a common access infrastructure. A key feature is tunneling.

Security

Obviously, organizations must carefully consider the issue of security when they permit remote users to access their corporate networks. The promotion of the Internet as a global, unregulated public network contributes to the perception that it is an insecure and inappropriate medium for sending sensitive company data. So do highly publicized reports of "hackers" who gain unauthorized access to the databases of supposedly well-protected organizations. Nonetheless, by applying suitable security measures, GSM access over Internet is more secure than many other more common methods (which use standard telephone lines to send data, often unencrypted) over the PSTN. The GSM network, for example, has many built-in security measures, including

- equipment and subscriber authentication and confidentiality;
- data encryption over the air.

Furthermore, the tunneling techniques described above demonstrate that secure virtual private networks may be created through the Internet. "Tunneled data" may be encrypted to prevent unauthorized parties from snooping at it from within the Internet. Finally, corporate organizations may implement additional security mechanisms of their own to authenticate remote users.

Benefits and market opportunities

End-user benefits

To the end-user, the most obvious benefits of the GSM Internet/intranet direct-access solution are as follows:

- Reduced call set-up time when connecting to a data service – typically under ten seconds compared with up to 40 seconds when the call is switched through a modem to the PSTN.
- Remote (off-site) employees gain secure access to company intranets.
- New services – in the long term, mobile operators will provide subscribers with a wide array of specially targeted services.
- The direct-access solution is easy to use – users, who are unaware of how their calls are routed, access the corporate network as if they had dialled directly.
- The direct-access solution supports standard client software, such as Windows 95 dial-up networking.

Company benefits

Besides allowing their employees greater mobility, companies who implement the GSM Internet/intranet direct-access solution reap additional benefits as well:

- Modem pools are replaced with a single router.
- Internet tunneling eliminates expensive leased lines.
- Multiple protocols may be transported over a single tunnel.
- IP addresses may be allocated from within the company.
- Security may be managed from within the company.

Network operator benefits

To the mobile network operator, the primary benefits of the GSM Internet/intranet direct-access solution are:

- Targeted services – mobile operators currently have no control over who uses their network for data communication. Data calls merely use the GSM network as a transit route to the PSTN and an Internet service provider. By contrast, the direct-access solution enables mobile network operators to introduce attractive new subscription and tariff plans that are targeted at specific market segments.
- Differentiation – by providing a unique offering of customer services, network operators can differentiate themselves from their competitors. As the number of operators increases, being able to stand out will matter more and more.
- Market image – for a limited time, mobile operators can enhance their market image by cashing in on the hype currently surrounding the Internet, just as many fixed operators are doing. In the long term, however, connecting to the Inter-

net will be a regular part of every operator offering. The direct-access solution gives mobile operators the means of providing this service.

- Reduced cost – because the direct-access solution terminates data calls within the GSM network (bypassing the fixed PSTN network), operators can save an estimated 5% to 10% on the volume of interconnect fees, depending on the circumstances of the local operator and the interconnect agreement.
- Migration path – the direct-access server accommodates analog and digital calls, allowing operators to offer parallel services and to migrate their dial-up data services from analog to digital.
- Immediate availability – the direct-access solution is not dependent on future GSM network hardware or software; instead, it is available for immediate implementation.

Conclusion

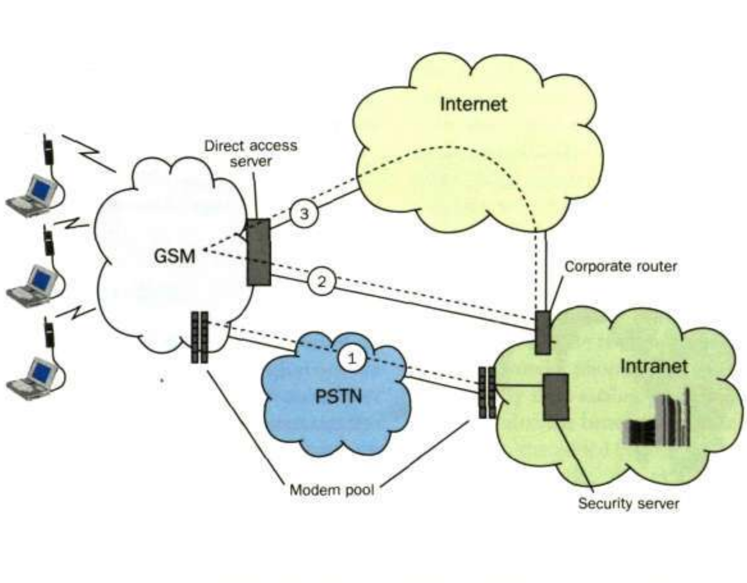
The market for mobile data communication is growing. Compared with early estimates, however, growth in this area has been disappointingly slow. Nonetheless, the success of GSM will doubtless have a significant, positive impact on mobile data communication. The commercial introduction of new GSM bearer services, such as high-speed circuit-switched data (HSCSD) services and general packet radio services (GPRS), will further boost the mobile data market.

Ericsson's GSM Internet/intranet direct-access solution introduces an access server at the GSM exchange site, creating a direct end-to-end digital connection that bypasses the PSTN. This solution gives mobile users universal access to the Internet and to their corporate intranets. As an added benefit, the call set-up time is reduced from between 20 and 40 seconds to under 10 seconds.

The direct-access solution eliminates the need for modem pools, saving corporations money. Moreover, it puts the allocation of IP addresses and the administration of security firmly in the hands of designated company personnel.

The GSM Internet/intranet direct-access solution enables network operators to target specific customer groups, allowing them to differentiate their services and market image from those of their competitors; it enables them to cut operating costs; and it gives them an analog-to-digital migration path.

Finally, the GSM Internet/intranet direct-access solution is not dependent on future GSM network hardware or software, but is available for immediate implementation.



Box B
Intranet connectivity methods

Route 1
Previously, the only way remote users could access a corporate network was through the .PSTN, using modem pools in the GSM network and at the corporate site.

Route 2
The direct-access server is placed in the GSM network, where it terminates data calls digitally, connecting them directly to a router in the corporate network – perhaps using a leased line.

Route 3
The Internet or other public data networks may be used as an inexpensive transport medium through which a secure "tunnel" may be established to carry data protocols. As a result, a virtual dial-up connection is created that mobile users can use to access a corporate network just as if they had dialed directly. As an added benefit, companies no longer need to construct a private access network, which can be prohibitively expensive.

Migrating to digital wireless services in an analog AMPS network

Magnus Nielsen and Tord Stureborg

The AMPS and D-AMPS wireless standards were originally developed in North America. Today they have been deployed in 97 countries. The AMPS standard allows operators to provide subscribers with analog and digital services using the same network infrastructure. Thus when operators of AMPS networks want to migrate to digital services, they may do so in a cost-effective, step-by-step fashion that protects their previous investments in infrastructure.

The authors describe what operators must do to successfully plan and introduce D-AMPS digital services into analog AMPS networks.

Box A
Abbreviations

ACA	Adaptive channel allocation
AMPS	Advanced mobile phone service
BER	Bit-error ratio
CLI	Calling line identification
D-AMPS	Digital AMPS
HSC	Hierarchical cell structure
IFAP	Improved fringe-area performance
IRC	Interference rejection combining
MAHO	Mobile-assisted handoff
MSC	Mobile switching center
SMS	Short-message services
TDMA	Time-division multiple access
VPN	Virtual private network
WIN	Wireless intelligent network

New technology and commercial considerations have heightened operator interest in migrating from analog wireless services to digital wireless services. The main commercial attraction of digital air interfaces is that they make more efficient use of radio spectrum; that is, digital networks accommodate more subscribers than do analog networks. Digital air interfaces also permit operators to offer a greater range of advanced revenue-generating services.

Although many people believe the advanced mobile phone service (AMPS) standard is strictly a US standard for mobile telephony, networks based on AMPS and digital AMPS (D-AMPS) technology support over 75 million subscribers. As the appeal of digital wireless services grows, the benefits of the AMPS standard become more and more apparent to operators all over the world. In particular, the AMPS standard allows analog and digital channels to co-exist within the same frequency spectrum in the same network. Thus operators may introduce digital wireless access into an existing analog AMPS network infrastructure in a graceful, cost-effective, step-by-step fashion, thereby protecting their investments in infrastructure and reaping the benefits of additional network services.

The functions and features of the latest IS-136 AMPS specification, finalized in 1995, enable operators to provide a wealth of new services, including wireless offices and virtual private wireless networks. The digital control channel in the IS-136 AMPS standard allows "sleep-mode" signalling between base stations and handsets, significantly reducing battery drain in the handsets. Operators in different countries have begun introducing global roaming between AMPS and D-AMPS networks. Together, these factors are encouraging a growing number of AMPS operators to introduce digital services into their networks. By the end of 1997, nearly 100 million subscribers worldwide are expected to benefit from digital wireless services (all standards). Some services will have been introduced into existing AMPS networks, while others will have been made available through new digital networks.

In order to free up capacity in the analog network, some operators are offering high-end users premium digital services. Dual-mode portable phones, which work on analog and digital wireless channels, enable network operators to mix analog and digital services in various parts of the network while providing subscribers with seamless access wherever they go.

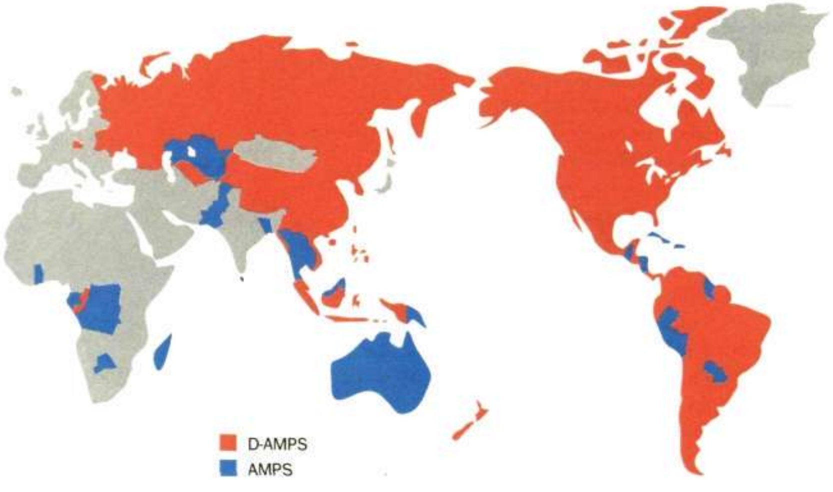


Figure 1
The AMPS and D-AMPS wireless standards are currently deployed in 97 countries worldwide, supporting over 75 million subscribers. The map shows countries where AMPS or D-AMPS systems are in commercial service.

Two underlying principles

Analog-to-digital migration is very straightforward in the AMPS network infrastructure – indeed, the IS-136 digital wireless standard was designed specifically for this purpose. Nonetheless, experience has shown that several practical issues influence the success of migration. Before we describe some of these in greater detail, we want to point out two very important considerations on which operators should base the entire migration:

- Operators will increase the success of their digital migration projects by synchronizing their technical and marketing activities during the planning and implementation phases. The commercial success of a digital migration project entails much more than simply getting the technical aspects right. Equally important are the marketing aspects. Timing, for example, is especially critical in meeting customer expectations of coverage, capacity and voice quality. Before moving ahead with implementation plans, operators may need to prepare the market, so that user demand picks up quickly once the digital services become available. If the marketing plans call for high-end users to migrate from analog to digital services, then the sooner they can be encouraged to do so the sooner operators can ease traffic loads on the analog network.
- Operators should begin digital migration by ensuring that their analog networks function correctly. If a network is not maintained to keep pace with increases in traffic, then the signal quality in some areas of the network may drop below optimum design levels. If this is the case, operators will encounter problems when they try to introduce digital channels into the network: voice quality in digital networks is critically dependent on signal quality.

Digital drivers

The chief rationale for migrating from an analog wireless network to a digital one are increased network capacity, new subscriber services, and new network functions and features. Nonetheless, analysts expect operators to continue investing in the AMPS network infrastructure, which indicates that these networks will remain in operation for some time. As one speaker at a conference staged by the UWC consortium put it: "Every dollar invested in my AMPS network is a dollar invested in my D-AMPS network." Even so, most development of new wireless network functionality is driven by requirements for D-AMPS networks. Wireless intelligent network (WIN) solutions, for example, will have an increasingly important role in the development and deployment of new services.

Another incentive for introducing digital services into AMPS networks is added revenue from digital subscribers who roam into the network from other networks.

Business and marketing plan

Each digital migration project should begin with a business and marketing plan for developing wireless services. Every development in the network should support this plan. The work begins by defining the scope of digital migration, and by describing the usage profiles of network subscribers: Which subscribers are to be targeted? How often do subscribers call? What time of day? What is the average duration of their calls? To make the most of network resources, operators must base their business cases on the answers to these and similar questions.

One migration strategy that has yielded positive financial results involves targeting high-end business users. By virtue of the call traffic they generate, these users – who generally welcome the increased range of services that digital access offers – are especially valuable to the network operator. Not only do they use new digital services, but by migrating from analog to digital access they free up capacity in the analog network for other subscribers.

In the initial planning phase, operators must also consider which services they will offer. Such services as voice-mail, short-message services (SMS), calling line identification (CLI) and virtual private networks (VPN) will probably require modifications to the network structure. Operators must also offer incentives to encourage subscribers to migrate to digital services. Digital dual-mode phones, for example, are more costly than analog phones. Will subscribers value the benefits of digital services more than the added cost?

Assuming the digital services are launched successfully, how will network traffic be affected? For digital transmission, operators usually appropriate channels from the analog radio environment. But if the

Figure 2
With this dual-mode, dual-band phone, subscribers can access analog and digital AMPS channels in the ordinary 800 MHz cellular band as well as in the 1900 MHz PCS band. This facilitates roaming in any AMPS/ D-AMPS network worldwide.

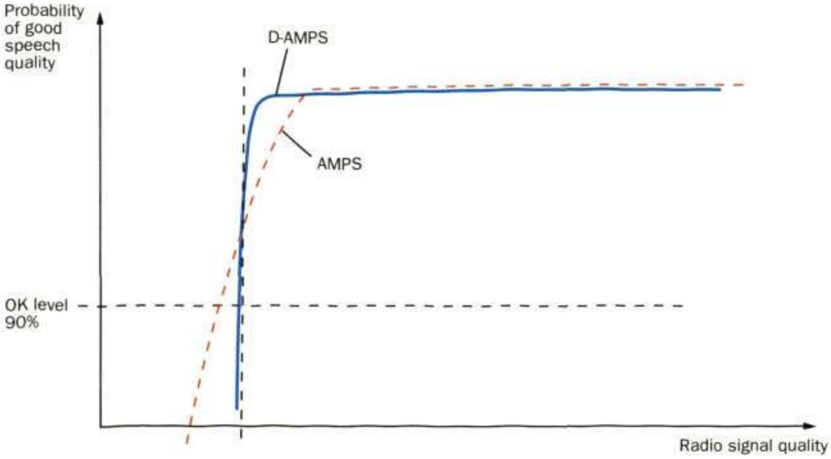
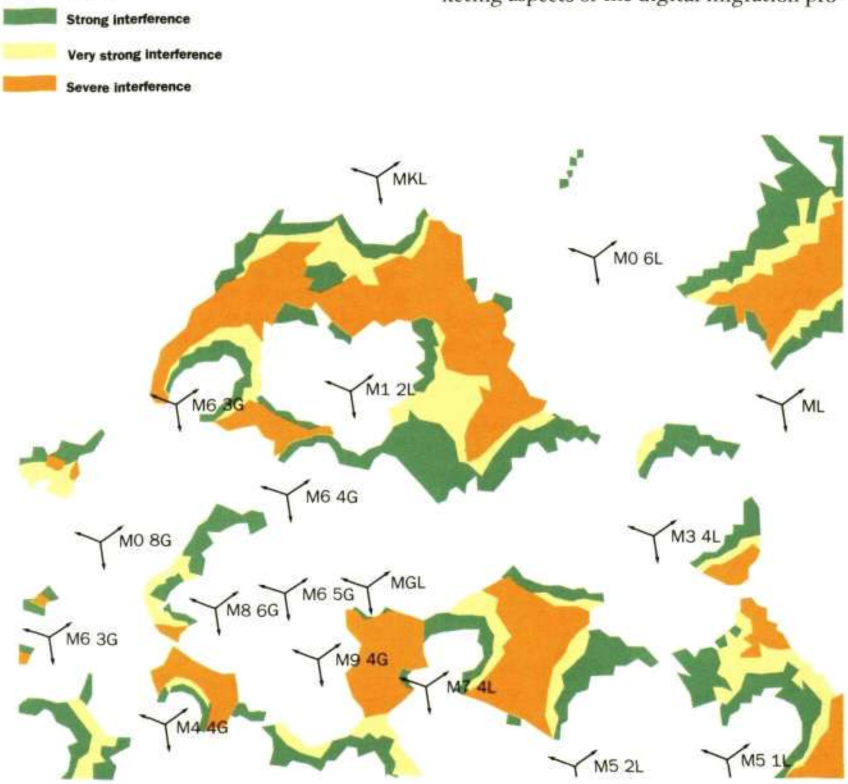


Figure 3
Under normal conditions, analog and digital channels provide good speech quality. As the signal quality decreases, the speech quality of analog channels decreases gradually, whereas for digital channel it drops off more sharply. Subscribers often accept deteriorating speech quality due to reduced radio quality on analog channels. On digital channels, however, reduced radio quality results in intermittently muted speech, which is difficult to understand.

Figure 4
Existing tools should be used to predict and prevent potential problem areas. The plot shows the probability of interference within the coverage area.



analog network is already heavily occupied, then taking channels from it will strain network capacity and reduce trunking efficiency. Therefore, if the engineering and marketing aspects of the digital migration project are not properly synchronized, then subscribers of the new digital services as well as subscribers of the analog services are sure to be dissatisfied.

Another important consideration at the outset is the geographical area in which the digital services are to be introduced. Large digital areas are generally preferable to small ones, since they cut down on the number of handoffs between analog and digital services as subscribers cross back and forth between the analog and digital areas of the network. Some network capacity is usually lost in the border zone between the digital part of the network (where both digital and analog access are available) and the rest of the network (where only analog access is available). Digital services are therefore best deployed in one cohesive area, instead of in small pockets. In this way, services that are unique to the digital part of the network remain accessible to subscribers. When operators plan their networks, they need to ensure that the border zone lies outside the high-traffic area. Many operators choose to provide coverage for an entire city, or to start in that part of the city where the majority of traffic is located.

Digital implementation plan

After having defined the business and marketing plan, operators should assess the quality performance of the analog networks. Most information for this assessment is available in the form of statistics from the mobile switching center (MSC). Data on signal strength and interference may be obtained on a per-cell or on a per-channel basis.

The criteria that subscribers use to judge the quality of service are accessibility, retainability and voice quality. Other criteria are cost and customer services, such as support and billing statements. Experience shows that analog AMPS networks furnish a very robust radio environment, even outside design criteria. Indeed, when the signal quality of the base station drops below recommended levels, the networks usually continue to provide useable services. That is, subscribers are often willing to tolerate increased levels of static and background noise provided they can still make and receive calls.

Digital channels with poor signal quality behave completely different: intermittent muting of the speech creates an unpleasant effect that is difficult and tiring to listen to. Figure 3 compares tolerance in analog and digital radio channels. The curve for analog radio channels shows a gradual drop in quality. The corresponding curve for digital channels, however, is characterized by a sharp drop as the signal quality falls below the minimum threshold.

Operators need to plan for good radio coverage before they begin to implement a digital migration project. If necessary – that is, if there is any doubt regarding network performance in certain geographical areas – then they should make field measurements of signal quality using mobile field-test terminals.

Operators also need to evaluate traffic distribution in the current analog network, identifying the high- and low-volume areas. This information is necessary for planning the location of digital services. Traffic-related information is generally available from the mobile switching center.

Finally, operators must consider what impact digital migration will have on the existing analog service. The implementation of digital services is likely to stretch analog network resources thin – since some analog channels will have been allocated to digital services. To manage the transition efficiently, engineering and marketing specialists need to jointly prioritize this phase of the project.

Radio environment plan

Two key factors in determining voice quality are signal strength and radio interference. In contrast to the performance of analog systems, which degrade gracefully, the degradation of voice quality in digital systems is very steep when parameters for signal strength and radio interference drop below the minimum design criteria. Experience shows that analog systems are often being stretched past the limit of good performance. A good rule to follow when migrating to digital services is to design the system for good digital service performance. In doing so, analog service performance will also be good.

The same frequencies should not be used for analog and digital channels. Instead, by splitting the frequency spectrum into designated analog and digital frequencies, operators can avoid the degradation of voice quality that arises from interference between the two types of signal. Splitting the frequency spectrum, however, requires operators to review and improve the frequency reuse pattern. If the same pattern is applied to analog and digital services, then any interference-related problems that arise can be resolved in a common manner for analog and digital channels. Redefining the frequencies used by individual cell sites often involves retuning all or part of the network.

Before operators can successfully introduce digital services, many find that they must first improve network performance by adding more base stations to the network. If done properly, this increases capacity in congested areas and brings service to areas where coverage had previously been poor. Although some analog network capacity may be lost after the upgrade – since analog channels are converted to digital channels – analog AMPS networks are generally robust enough to cope with such losses during the transition.

Operators must also plan how they are to increase network capacity. One technique, called cell-splitting, may be applied as needed, step-by-step. The first stage of cell-splitting may be achieved without installing additional cell sites. Existing omnidirectional sites are converted into sectorized sites. Further cell-splitting requires additional sites to be installed between existing sites, according to a planned geographical pattern. From the very beginning, network engineers must very accurately comply with the grid when choosing site locations. Once cells have been split, any offset from the correct location – which might initially have been acceptable – will be magnified, growing to an unacceptable relative value. That is, the distance between two sites is halved with each successive splitting of cells.

If more capacity is required for a smaller area close to the site location, then the overlay/underlay function may suffice. This function enables operators to borrow one or more radio channels from other cells. To avoid interference, the cell radius of the extra channels must be smaller than for ordinary channels. The overlay/underlay function hands off calls to the extra channels as soon as a mobile phone moves into their coverage area.

Another technique is to introduce small cells (microcells and picocells) into a large cell (macrocell) where extra capacity is needed[1]. The hierarchical cell structure (HCS) concept, which boosts coverage considerably, is especially useful in bringing services to indoor areas.

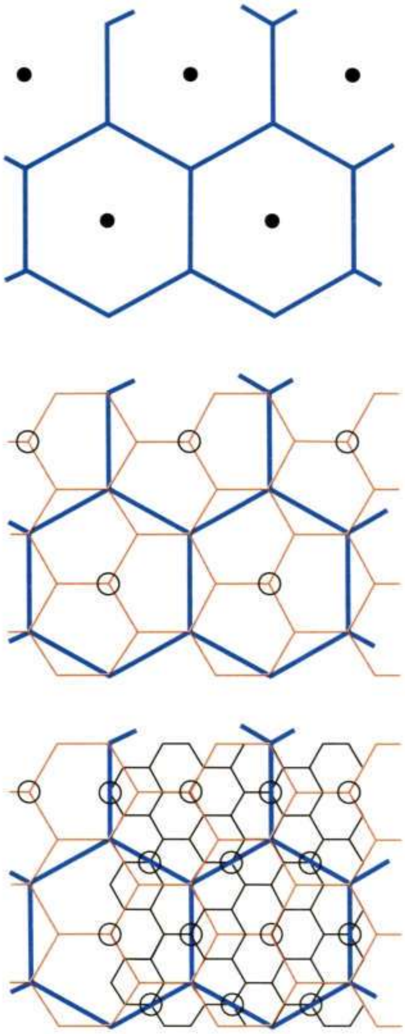


Figure 5
The introduction of new sites to increase capacity within the same coverage areas is called cell splitting. The tolerance for poorly positioned sites declines each time a cell is split. That is, the relative offset from the cell's ideal position is magnified as the cell size decreases, resulting in stronger interference.

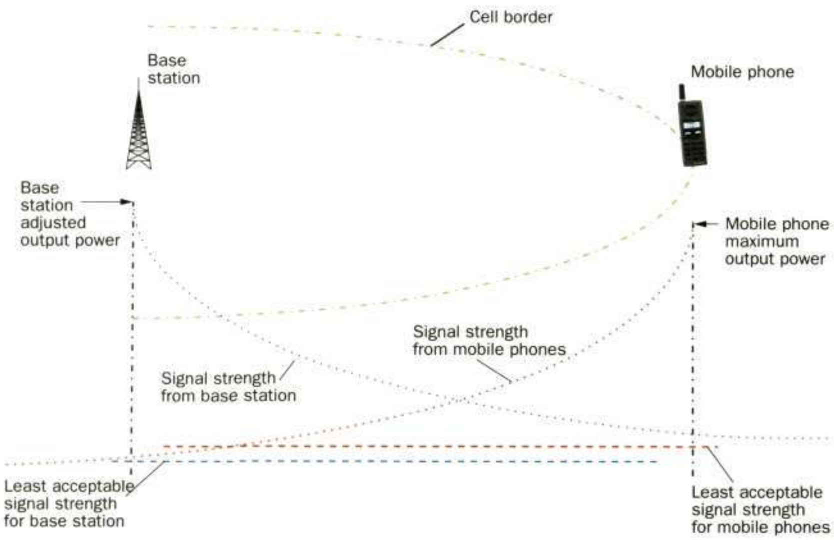

Cell border
Base station
Mobile phone
Base station adjusted output power
Mobile phone maximum output power
Signal strength from mobile phones
Signal strength from base station
Least acceptable signal strength for base station
Least acceptable signal strength for mobile phones

Figure 6
The output power of the base station should be adjusted to provide acceptable signal strength for mobile phones at the cell boarder. Excessive output power in large cells provides coverage in only one direction, since the maximum output power of the mobile terminal is too weak to reach the base station.

Path balance is another important issue. In a balanced path, the signal strength from the base station to a mobile phone at the border of the cell is just sufficient to yield good speech quality. The received signal strength at the base station is adjusted up to the maximum transmit power of the mobile phone. In this way, equally good speech quality is provided at the base station. Obviously, the cell may not be so large as to prevent the radio signal from the mobile phone from reaching the base station.

Voice quality and service availability depend on two-way communication over the wireless link. Excessive output power from the base station, or from mobile terminals, creates unnecessary interference in other cells that use the same or adjacent channels. In large cells, excessive output power might fool mobile phones into believing they are within the coverage area of the base station, although the maximum output power of the mobile terminal may not be strong enough to reach the base station.

Handoff between base stations must be transparent. When noticeable, different voice characteristics, muted speech, different audio levels, and echo are factors that negatively influence user perception of digital service. The implementation and configuration of the network must enable users with dual-mode analog-digital phones to travel to and from analog and digital access without noticing significant changes in the voice characteristics. Operators must control the audio levels in different parts of the network, keeping them constant and at a correct level, regardless of which trunk line external calls are being routed over. This is particularly important for avoiding echo in digital calls: for the echo cancellers in the switching center to operate as intended, the audio level must be set correctly in both directions.

Although a great deal of care goes into optimizing the radio environment for digital transmission, operators should always evaluate network quality. Findings from the evaluation are fed back into the network design and optimization process. If discrepancies are found, the algorithms and parameters in use must be modified to provide the proper results. In order to maximize system quality, the network frequencies must occasionally be retuned.

Supporting tools and functions

In spite of thorough planning, operators cannot always completely avoid interference. In order to minimize the effect of interference, one of the system functions measures the radio channel before call set-up. If an interfering signal is present, another channel is selected. If every digital channel is occupied, then an analog channel is chosen. Regardless, the subscriber perceives good speech quality.

Digital service introduces many new parameters that must be defined and maintained. In the mobile-assisted handoff (MAHO) function, for example, the mobile terminal measures signal levels from defined neighboring cells. At each new location the mobile phone must be updated with a list of the neighboring cells it is expected to measure. Incorrect information could result in a handoff to the wrong cell – resulting in poor signal quality and deteriorated speech quality; or worse, if handed off improperly, the call might even be dropped.

The measurement channel number automation function automatically defines a correct list of channels for each cell in the network. If necessary, operators can override this function and manually define the channels.

Ericsson will soon introduce a solution (called improved fringe-area performance, IFAP) that makes use of robust analog service to boost performance in areas where digital radio coverage is poor. This service is intended for use mainly in rural areas. When the signal strength drops below a given threshold, the call is automatically handed over to an analog channel; if signal strength increases, then the call is handed back to a digital channel.

To help operators plan the radio network of their digital systems, Ericsson has published a comprehensive document with recommended procedures, detailed information on design parameters, and associated functions. The document, entitled RF-Guidelines, explains the fundamentals to beginners, recommending the initial settings of critical parameters, and describing the effects of varying these parameters.

RF-Guidelines is updated for each new system release. The updates include new parameters and necessary changes to existing settings.

Valuable lessons have been learned from migrating to digital services. These lessons have been compiled into a collection of hints and recommendations. Digital Best Practices Recommendations, as the compilation is popularly called, has been published on the Ericsson Web, and is available to every operator who has signed a service agreement with Ericsson. The Digital Best Practices Recommendations will be updated with important new observations and lessons gained from the field.

Planning for growth

In coping with growing subscriber demand, operators may choose from a variety of tools that help them to increase capacity and simplify frequency planning as they add new radio base stations to existing wireless IS-136 networks. Two tools in particular are hierarchical cell structures and adaptive channel allocation (ACA).

Hierarchical cell structures

Using hierarchical cell structures, operators may plan the radio environment around three types of cell: macrocells, microcells, and picocells. As their names imply, these cell types vary in size. Macrocells cover large geographical areas. A single macrocell may cover a rural area, a small town, or a district of a large city. Depending on the design criteria, the size of a macrocell may range from less than one kilometer to as much as 30 kilometers in radius. Microcells and picocells are used to provide coverage to locations where subscriber density is high, including indoor areas. A microcell, for example, might provide coverage to a single street, whereas a picocell might cover a single building, such as a shopping mall, or the floor of a skyscraper. If operators require additional capacity – for instance, to provide coverage and capacity to a sports arena or to an airport terminal – then with a hierarchical cell structure they may add a microbase or a picobase station specifically for that location.

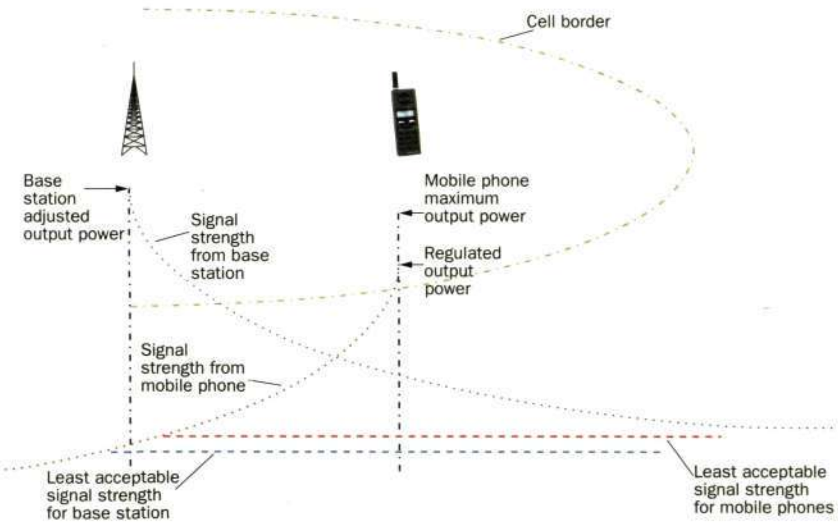

Cell border
Base station adjusted output power
Signal strength from base station
Mobile phone maximum output power
Regulated output power
Signal strength from mobile phone
Least acceptable signal strength for base station
Least acceptable signal strength for mobile phones

Figure 7
The output power of the mobile phone is automatically regulated, up to its maximum power, to provide sufficient signal strength at the base station regardless of the distance to the base station. If the cell size is reduced, the output power of the base station should be adjusted accordingly to reflect the shorter distance to the cell boarder.

The three-level cell structure enables operators to offer subscribers different services and charging tariffs according to their physical location. This differentiation of services is based on identifying certain "preferred cells" to which subscriber phones always try to connect, if available. All necessary communication between the phone and the wireless network is handled by the digital control channel.

Adaptive channel allocation

Adaptive channel allocation affects overall network capacity and quality, and makes it easier to expand the capacity of existing base stations and to add extra base stations. With adaptive channel allocation in service, the best available radio channel in the entire network is automatically allocated to each new call. Channel selection is based on the measurement of bit error ratio (BER) and channel interference. By fully implementing adaptive channel allocation, operators can increase the capacity of their wireless networks by as much as 100%.

In the past, when operators added cells to increase network capacity, they generally had to plan to retune the network frequencies as well. Adaptive channel allocation eliminates the need for manual retuning. In-

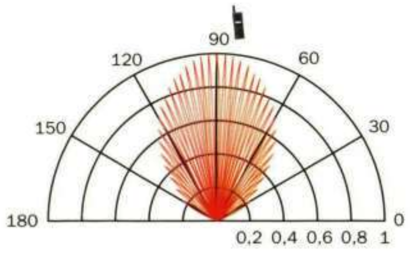


Figure 8
The directivity pattern forms an array of beams using a diversity combination of two or more receiver antennas at the base station. In normal diversity, the equalizer in the receiver adjusts the beams to point with one finger towards the mobile phone, thereby maximizing the signal strength. With interference rejection combining (IRC), the equalizer adjusts the beam pattern, focusing the direction to the strongest interference in a minimum between two fingers. This maximizes the relationship between the correct signal and interfering signals.

stead, operators simply install a new transceiver and switch it on. By applying this technique dynamically to the day-to-day operation of their wireless networks, operators may allocate radio frequency capacity where it is needed most, in response to changes in traffic patterns.

Launching the network and services

Ericsson recommends that operators first launch wireless digital service on a test basis. For instance, operators might provide a few hundred digital phones to subscribers who agree to report their impressions and experiences of the service quality, especially in terms of coverage, capacity and speech quality. Subscribers who are accustomed to using their analog phones in fringe areas – for instance, indoors – may be disappointed with digital service: instead of increased static and background noise, the digital audio signal might intermittently be muted (cut out), making conversation difficult. As subscribers grow accustomed to using their mobile phones in office buildings, elevators, and underground car parks, their expectations rise. The hierarchical cell structure is a good way of improving indoor wireless access to difficult radio environments.

Timing and speed are key factors when launching new commercial digital services. Some network operators subsidize the cost of mobile phones to stimulate a rapid market take-up of their services.

Subscriber expectations are another important factor. The voice characteristics of digital phones differ from those of analog phones. If the network has been optimized properly, however, then the speech quality obtained from a digital phone will not include any static or background noise. Operators should be forewarned that many consumers associate the term "digital" with "high-fidelity." While the sound quality of digital phones is good, it is definitely not hi-fi class. Marketing messages need to take this into account.

New developments

Today Ericsson is working on several solutions that will give wireless network operators more flexible and effective ways of achieving better coverage, capacity and voice quality. For example, an enhanced full-rate vocoder, which will become standard in new digital phones this year, significantly improves voice quality. Similarly, Ericsson will soon introduce two new techniques for reducing interference: interference rejection combining (IRC) and downlink power control.

Interference rejection combining, which can be added to existing base stations, introduces a new way of handling signals. In the past, two incoming signals were combined to produce the maximum signal level. However, by means of IRC, the signals' phase and amplitude are adjusted to reject the greatest interference. This results in a slightly weaker signal level, but because the interfering signal has been reduced more than the correct signal, the net result is improved signal quality.

Downlink power control is a path-balancing technique that automatically adjusts the transmit power level of the radio base station relative to a subscriber's physical location. This technique decreases the risk of network interference by reducing the transmit power to terminals that are near the center of the cell.

Conclusion

Because analog and digital AMPS services may be combined in the same network infrastructure, operators have complete flexibility in deciding when and where to start the digital migration process.

Current analog AMPS networks constitute a strong platform for introducing wireless digital services.

Operators' main reasons for migrating to a digital wireless network are increased network capacity, new subscriber services, and new network functions. Successful migration, however, involves more than simply implementing new technology. Operators must also consider marketing aspects very carefully.

Each digital migration project should begin with a business and marketing plan. Next, operators need to assess the performance of their analog networks. Assessments of this kind help operators to draw up an appropriate radio environment plan. As part of this plan, operators may elect to introduce hierarchical cell structures and adaptive channel allocation – two effective means of ensuring that network capacity and quality can and will keep up with subscriber demand.

References

1 Genell G. and Ek M.: New RBS 884 base stations take IS-136 D-AMPS wireless services into new areas. Ericsson Review 74 (1997: 3), pp. 111-115.

New RBS 884 base stations take IS-136 D-AMPS wireless services into new areas

Mats Ek and Gunnar Genell

By introducing a hierarchical cell structure into D-AMPS IS-136 wireless networks, operators may significantly increase network capacity and extend wireless access into difficult locations.

Ericsson's newly expanded family of RBS 884 radio base stations covers every requirement of a hierarchical cell structure, giving operators the flexibility to tailor-fit network capacity indoors and out.

The authors describe the RBS 884 radio base stations, which transform the concept of hierarchical cell structures into an elegant networking solution.

In the initial phases of deploying a mobile phone network, wireless network operators are primarily concerned with achieving good geographical coverage of the population area. Once the network infrastructure is in place, however, operators broaden the focus of their attention to take in other aspects. Network capacity soon becomes every bit as important as coverage, since operators are under pressure to expand their networks in order to cope with growing numbers of subscribers and increasing traffic volume. At the same time, the issue of coverage takes on a new dimension as operators seek to extend the reach of their networks into completely new locations, especially indoors.

Several techniques may be used to increase network capacity. One technique, called cell-splitting, involves splitting large cells into smaller ones. Cell-splitting creates a network infrastructure that is characterized by cells of different sizes. The smallest cells are generally deployed in city centers, where subscriber density and traffic volume are greatest. The main drawbacks to cell-splitting are inefficient use of radio spectrum, and requirements for re-engineering the cell plan each time a cell is split.

A more elegant solution leaves the existing network of radio base stations intact by introducing into it a hierarchical cell structure (HCS). The result is a multilayered radio environment with base stations of varying output power in differently sized cells.

Capacity and coverage

Hierarchical cell structures offer three significant benefits.

- They increase network capacity and provide support for more subscribers and greater traffic volume without requiring additional radio spectrum. Where demand is especially high, operators may add extra capacity to specific geographical areas of the network.

Figure 1
The hierarchical cell structure concept and the family of RBS 884 products bring wireless communication into new indoor areas, enabling D-AMPS users to make or receive calls in virtually any location.

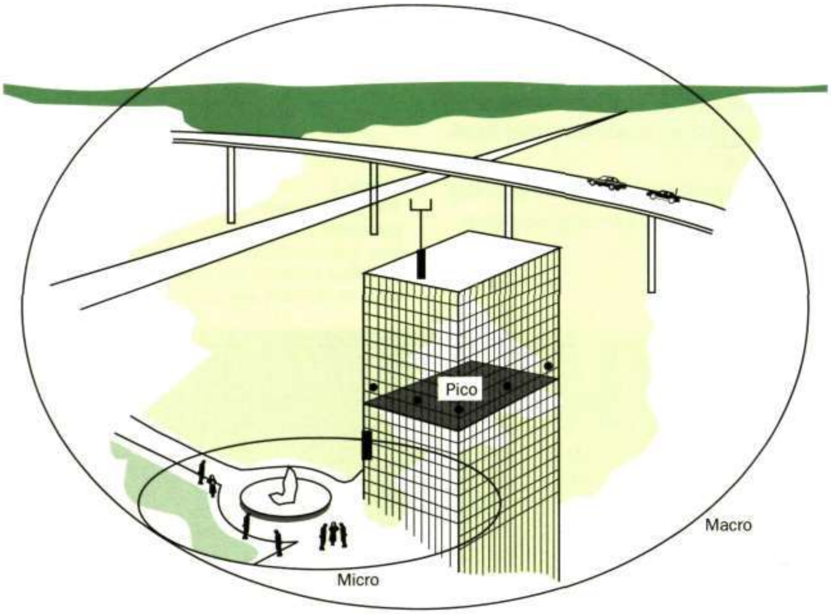


Figure 2
With a hierarchical cell structure, an operator can tailor-fit capacity to match traffic demands, and use the available frequency spectrum more efficiently. Mobile terminals moving at high speed are directed to macrocells. Hot spots with mobile terminals moving at low speed are covered by small cells (microcells and picocells).

- They allow operators to extend, in a cost-effective way, the reach of their networks into locations where it has previously been difficult – if not impossible – to guarantee satisfactory wireless access.
- They enable wireless network operators to introduce a new generation of location-based services.

Box A
Abbreviations

ACA	Adaptive channel allocation
ACC	Analog control channel
AMPS	Advanced mobile phone service
AVC	Analog voice channel
CDPD	Cellular digital packet data
D-AMPS	Digital AMPS
DBC	Downbanded cellular
DCCH	Digital control channel
DTC	Digital traffic channel
HCS	Hierarchical cell structure
IRC	Interference rejection combining
IS-136	Mobile telephony standard on which AMPS and D-AMPS are based
MCPA	Multicarrier power amplifier
PCS	Personal communications services
SR	Signal strength receiver
TDMA	Time-division multiple access
VER	Verification unit

Hierarchical cell structures enable operators to define different cell types and different relationships between cells. Network designers may allocate capacity at a detailed level, creating customized solutions for specific traffic demands. For example, an operator might choose to insert a new base station to alleviate a high volume of wireless traffic at a congested highway intersection. In a hierarchical cell structure, this measure would require only minor changes to the frequency plan.

The coverage capabilities afforded by a hierarchical cell structure are particularly interesting. The upper levels of high-rise buildings, for instance, have traditionally been a problem area for wireless access, since the conventional wireless network infrastructure was designed for subscribers at or near ground level. With a hierarchical cell structure, operators may deploy wireless access on a floor-by-floor basis.

Besides high-rise buildings, other problem areas for wireless access include tunnels, underground car parks, underground railway stations, and areas deep inside large buildings. Depending on their shape and construction, single indoor areas may also constitute problem locations for wireless access. Hierarchical cell structures provide a solution to each of these situations.

Another important feature of a hierarchical cell structure is that its network architecture does not have to be deployed from scratch. Instead, this technique may be introduced into existing digital advanced mobile phone service (D-AMPS) networks in a step-by-step fashion, increasing capacity and coverage as needed.

The hierarchical cell structure represents one of many techniques used in tuning the radio environment to provide greater capacity, reduce radio interference, and improve service quality. Other techniques that may be applied in conjunction with a hierarchical cell structure include adaptive channel allocation (ACA), interference rejection combining (IRC), and downlink power control[1].

Three cell levels

Being the subscriber point of access, base stations are key links in a cellular wireless communication network. Up to three layers of cells are envisaged for hierarchical networks based on the D-AMPS (IS-136 TDMA) wireless standard.

Macrocells

Ordinarily, the first step in rolling out a network involves deploying base stations with high output power and receiving capability. The wide-area cells, or macrocells, that these kinds of base station serve, might cover an area of between 20 and 40 kilometers. Macrocells are the largest cells in the wireless network.

Macrocells are placed next to one another to form a contiguous area. As mobile subscribers move through the area, leaving one cell and entering another, their calls are handed off. Handoff is accomplished by comparing signal strength to determine which radio base station provides the strongest signal.

Microcells

Microcells are used to provide additional capacity for small areas within macrocells, generally where traffic density is greatest. Providing service to an area of approximately one kilometer, microcells may be deployed to alleviate heavy traffic on a single street, at a busy intersection, or at other sites with similar requirements for increased capacity.

Picocells

Picocells, which are inserted within microcells or macrocells, are generally intended for use in indoor locations such as offices, hotel lobbies, and underground railway stations. Operators may also deploy picocells on each floor of a high-rise building to provide seamless wireless access to the entire environment.

Cell selection and new subscriber services

The D-AMPS IS-136 wireless standard accommodates several procedures for selecting cells; that is, cell selection is not based solely on signal strength. When mobile phones are in use, they constantly collect measurements from surrounding cells – even when handoff is not imminent. Neighboring cells are defined as preferred, non-preferred, or standard. Mobile phones always seek preferred cells first, staying with them as long as the signal strength and other factors exceed the levels defined by the wireless operator.

Small cells that have been designated as the preferred cells within a larger cell (microcells and picocells) attract traffic, offloading traffic from the larger cells. This holds true even when the smaller cells are located near the center of a large cell. As a mobile phone leaves one of the smaller cells, the signal strength approaches the bottom threshold and the phone automatically searches for other preferred cells in the vicinity.

Overall network capacity may be increased significantly by designating every picocell and microcell as a preferred cell. That way, most subscriber traffic could be handled at the lowest levels of the wireless network hierarchy. Special algorithms in the radio network determine whether or not a mobile terminal is travelling very rapidly – for example, by car – in which case radio access is maintained at the macrocell level.

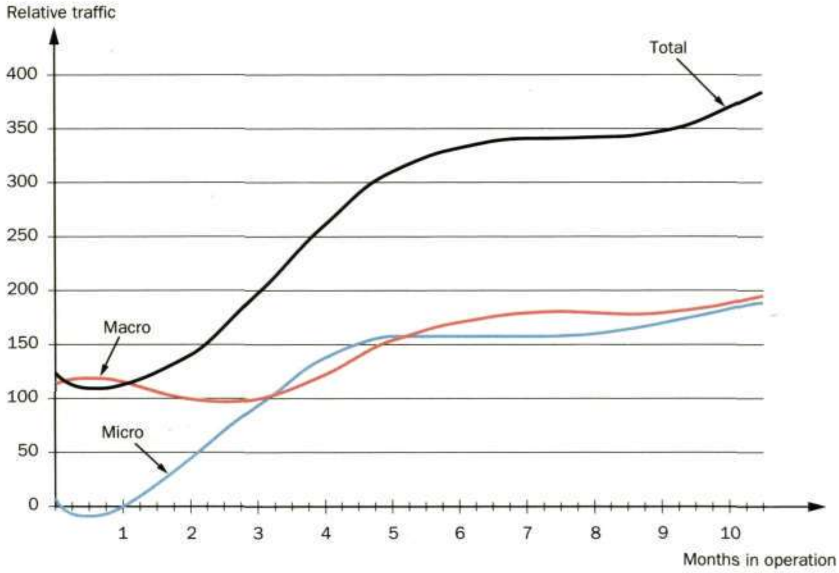


Figure 3
The diagram illustrates the average traffic levels in a cellular network where capacity has been increased by introducing microcells into the most traffic-demanding areas. The three lines represent traffic in macrocells and microcells as well as total traffic.

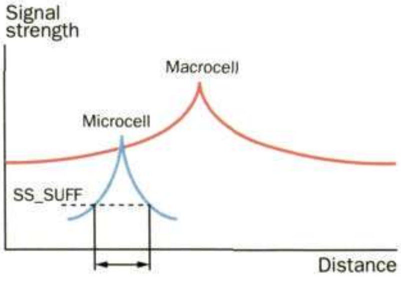


Figure 4
Compared with traditional single-layer cell structures, hierarchical cell structures enable operators to introduce new criteria for handoff. For example, a microcell may be defined as the preferred neighbor to a macrocell – in which case the mobile terminal will hand off to the microcell as soon as the signal strength exceeds the programmable limit SS_SUFF.

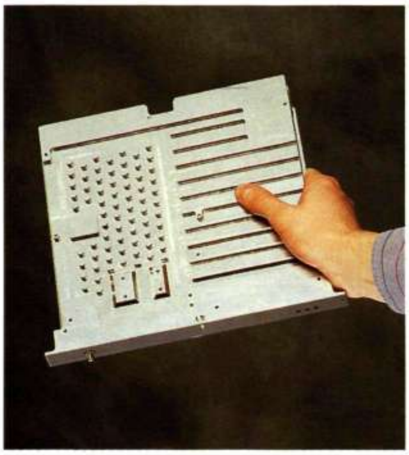


Figure 5
The main building block in each RBS 884 base station is the multimode transceiver (TRX), which is capable of serving every channel function: AVC, ACC, DTC, DCCH, VER, SR and CDPD. The TRX does not require any on-site calibration since all parameters are controlled by software.

The cell-selection procedures also provide a mechanism that enables wireless network operators to target specific segments of the user market, by offering different communication services on the basis of their location. By deploying picocells to provide coverage for a cluster of offices in a large complex, for example, an operator could provide the offices with a unique array of business communication services. The "wireless office" is expected to play an important role in helping businesses to integrate their fixed and mobile communication resources into a seamless communication solution.

The combination of hierarchical cell structures and special cell-selection procedures allows for seamless wireless networks to be created indoors and outdoors, which is the objective of personal communications services (PCS).

Ericsson RBS 884 portfolio

Ericsson is the largest supplier of wireless networks based on the IS-136 TDMA standard. Ericsson's CMS 8800 cellular mobile system includes the mobile switching centers and radio base stations needed for networks operating in the 800 and 1900 MHz frequency bands.

In 1996, Ericsson expanded its line of CMS 8800 radio base stations to include service of outdoor and indoor cells for hierarchical cell structures. The emphasis of this expansion was on delivering an array of base stations that provide the most cost-effective coverage of any outdoor or indoor location. The smallest units in the new family of RBS 884 base stations are very easy to install.

One of the most important developments in the new family of base stations is the introduction of a versatile multimode, multiapplication transceiver. For example, whereas the previous generation of base stations (RBS 882) contained separate modules to handle analog voice channels, digital voice channels, location functions, and verification functions, the new RBS 884 radio base stations handle these functions using a single module. In analog mode, the voice modulates an FM carrier using 30 kHz radio channel bandwidth. In digital mode, speech is coded and a digital bit stream modulates the carrier using phase-shift keying. Three voice paths are transmitted in the 30 kHz bandwidth.

The RBS 884 transceiver operates in different modes. For analog access, it can be used as a control channel, an analog voice channel, or as a signal strength receiver. For digital access, it can be used as a digital control channel, a digital voice channel, or as a verification unit. The transceiver also supports cellular digital packet data (CDPD) services.

The RBS 884 base stations are available for IS-136 wireless networks operating in all frequency bands:

- D-AMPS/AMPS cellular – in the 850 MHz frequency band;
- D-AMPS downbanded cellular (DBC) – in the 800 MHz frequency band;
- D-AMPS 1900 – in the 1900 MHz frequency band.

Outdoor applications

The RBS 884 Macro is designed to support every level of network requirement, from very great to very small capacity, including different output power classes for various coverage demands. This base station facilitates easy expansion to accommodate rising traffic demands. In standard configurations, one base station handles up to 32 carriers per cell.

The balance between coverage and capacity in rural locations is different from that of urban locations. The RBS 884 Macro provides optimum coverage of rural areas without compromising transmit and receive performance. Its advanced antenna system increases signal sensitivity, which guarantees high connection reliability, improves voice quality, and reduces terminal battery usage – since less output power is required of the mobile terminals.

The configuration of the base station varies according to the frequency band in use. For example, in the 1900 MHz frequency band, an arrangement of four-branch antennas ensures maximum gain of received signals.

The RBS 884 Micro is a microbase station designed to provide extra capacity where wireless coverage already exists. It is a complete stand-alone unit for outdoor installation. A multicarrier power amplifier (MCPA), which increases the strength of several transceiver signals simultaneously, helps account for the small size of the base station.

Indoor application

The RBS 884 Micro without MCPA is a microbase station designed to provide extra capacity indoors – for example, in large shopping malls, airports, and convention centers.

Ericsson has solutions for creating picocells to extend wireless network coverage to indoor radio environments – such as hotel lobbies, underground railway stations, and offices – that are otherwise difficult or impossible to reach.

Conclusion

Hierarchical cell structures enable operators to extend wireless coverage into locations where it was not previously possible or cost-effective to do so. They also enable operators to increase network capacity to match subscriber growth and local traffic requirements, providing coverage and capacity where it is needed most.

The family of Ericsson RBS 884 base stations gives operators of a wireless D-AMPS network a complete unified array of units for every network requirement in a hierarchical cell structure. They give operators a dynamic and flexible way of increasing wireless network coverage and capacity.

When deployed in a hierarchical cell structure, the new base stations bring wireless mobile communication one step closer to the level of true personal communications services.

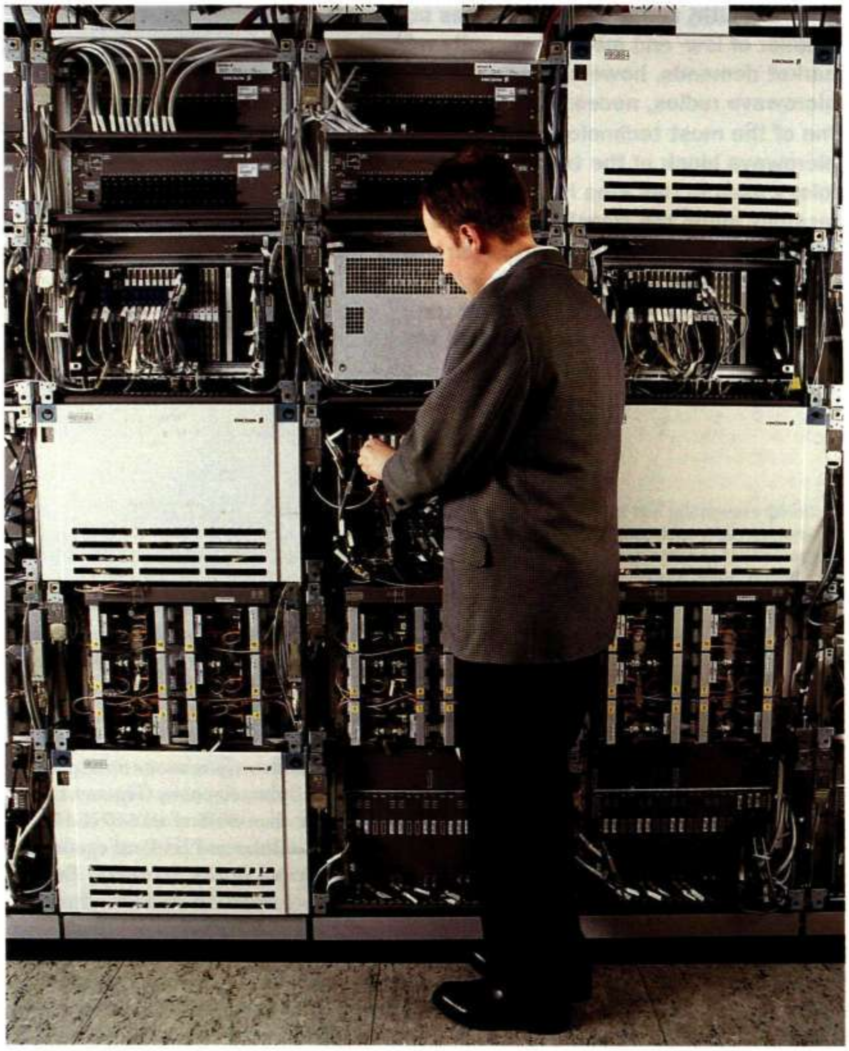


Figure 6
RBS 884 Macro is a modular product concept that allows the same building blocks to be used in many different standard configurations. The base station is available in different power classes and frequency bands, which means it may be installed in any customer application.

References

1 Nielsen, M. and Stureborg, T.: Migrating to digital wireless services in an analog AMPS network. Ericsson Review 74(1997:3), pp. 104-110.

New technologies for future microwave communication

Dag Jungenfelt, Martin Löfgren and Herbert H. G. Zirath

The MINI-LINK series of products has made Ericsson the world's leading supplier of low- and medium-capacity microwave radios. Rapidly changing market demands, however, together with fierce competition in the field of microwave radios, necessitate that these products continue to evolve.

One of the most technology-intensive portions of microwave radios is the microwave block of the transmitter-receivers. Recent advances in the technology used in this area have led to the development of MMICs. Having successfully made the transition from the research lab to mass production, Ericsson is now poised to bring their own MMICs to market.

The authors describe how the integration of MMIC technology is certain to influence the provision of radios for transport and access networks. They then describe the technology behind MMICs, covering various design and production aspects.

MINI-LINK products

Thanks to the MINI-LINK series of products, which first appeared on the market in 1980, Ericsson has become the world's leading supplier of low- and medium-capacity microwave radios. Today MINI-LINK products account for 25% of all such microwave radios sold.

The rising number of MINI-LINK installations is directly related to the growth of mobile cellular telephony (Figure 1). Currently, more than 80% of all MINI-LINKs are used in cellular and personal communications services (PCS) networks. Besides tremendous growth in mobile telephony, other factors that contribute to the market success of the MINI-LINK are:

- its modular concept – being modular, the MINI-LINKs may easily be adapted to individual customer needs using a limited number of unit types;
- its compact size – fairly extensive MINI-LINK nodes may be integrated into ever-shrinking base station cabinets;
- its high degree of reliability – engineers at Ericsson draw on field experience to create robust designs.

By cooperating very closely with major customers, Ericsson have steadily continued to develop the MINI-LINK, improving such characteristics as mean time between failures (MTBF), transmission capacity, availability in different radio frequency bands, functionality, transmission performance, and mechanical design.

Traditionally, the microwave block of the transmitter-receiver – which consists of up- and down-converters, local oscillators, and power amplifiers – has been the most technology- and labor-intensive part of microwave radios. Fortunately, the rapid evolution of digital and analog low-frequency application-specific integrated circuits (ASIC) has now carried over into microwave technology (Figure 2), giving rise to the development of microwave monolithic integrated circuits (MMIC).

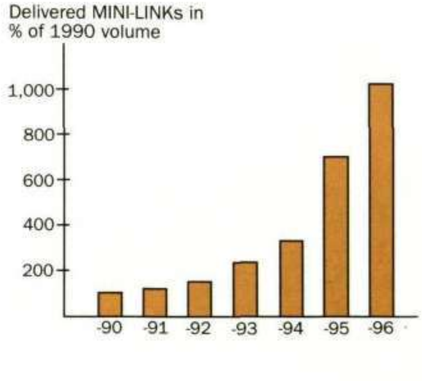


Figure 1
Delivered MINI-LINKs per year since 1990.

The implications of MMICs

In terms of design, MMIC technology greatly affects the physical size, power consumption, and production methods associated with microwave products.

First, by introducing MMICs into their designs, engineers at Ericsson Microwave Systems have succeeded in developing smaller radios with outdoor parts whose volume is about half that of the previous generation. In spite of the size reduction, the radios may contain much greater functionality than before. Moreover, being unobtrusive, the smaller-sized products are more readily accepted by the general public, enabling operators to expand their access networks into residential areas.

Second, the introduction of integrated circuits helps reduce power consumption, since the performance of MMICs is optimized for specific applications. The reduced size leads to lower transmission-line loss, which further reduces power consumption.

Third, the new MMIC-based microwave technology facilitates production processes. Present-day manufacturing techniques, for example, require skilled technicians to tune microwave products manually. By contrast, the techniques used for manufacturing MMIC-based products can be automated almost entirely. Consequently, unit production costs are not limited by technology, as they are today, but by volume.

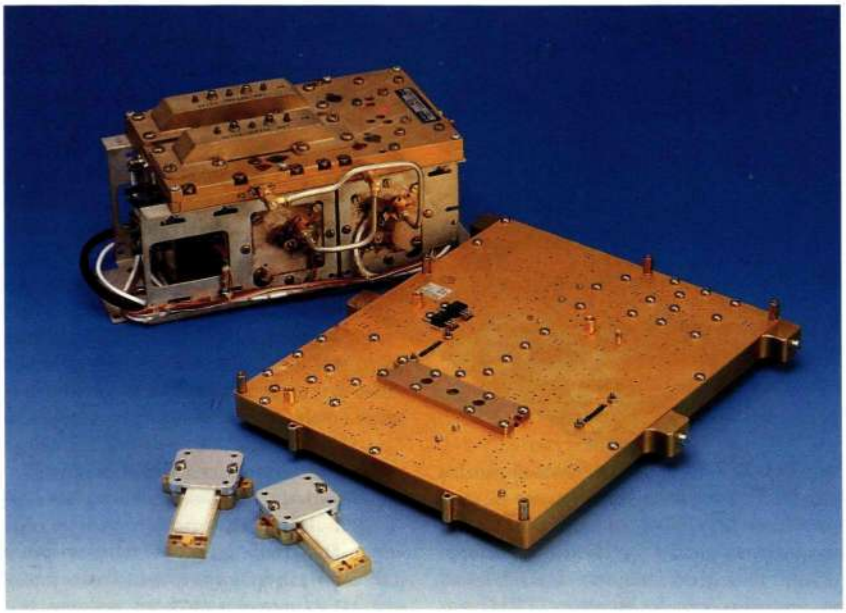


Figure 2
Three generations of the microwave block: back, waveguide technology (first generation, 1985); middle, microstrip technology (current generation, 1992); front, MMIC technology (next generation).

New MINI-LINK applications

Because the majority of MINI-LINK radios find their way into cellular and PCS networks, each new generation of MINI-LINK products must keep pace with the evolution of these networks. For instance, many cellular and PCS operators are currently introducing smaller cells (microcells and picocells[1]) into their networks in order to increase traffic-handling capacity. As a result, the distance between adjacent base stations is shrinking. Operators are also installing smaller, more highly integrated base stations in order to minimize their costs per site.

Considering the transmission network that is needed to backhaul base stations to switching centers, trends of this kind impose new requirements on MINI-LINK products. Denser installations of smaller base stations dramatically increase the relative cost of the transmission network. To Ericsson and other system vendors, this implies that providing cost-effective transmission solutions is a key issue. In response to the new market demands, Ericsson Microwave Systems is working

- to introduce MMICs and other new technologies into MINI-LINK products – which helps reduce size and improves performance, reliability, and producibility;
- to introduce new system concepts, such as dual-purpose, all-outdoor radios which, being software-configurable, operate as conventional point-to-point radios or as terminals (outstations) in a point-to-multipoint (PMP) system.

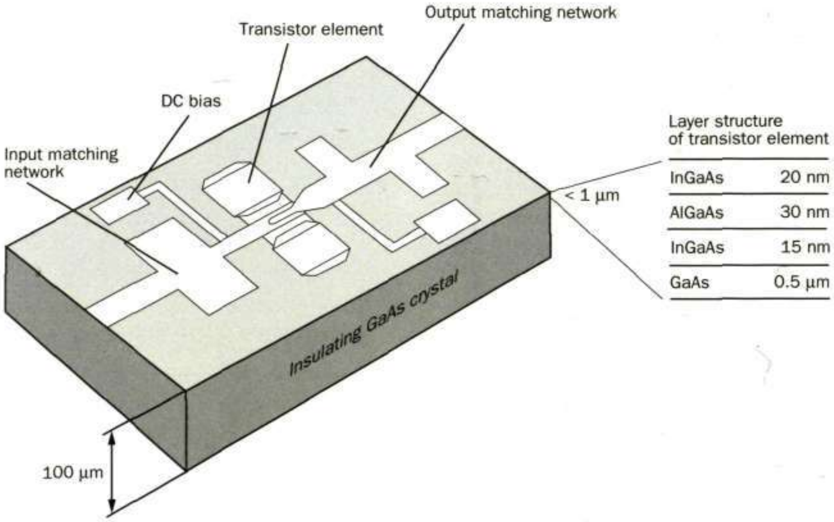
As base stations diminish in size, designers find it increasingly difficult to make room for transmission equipment in the cabinets. Present-day transmission solutions for cellular and PCS backhaul include a relatively high degree of physical integration, but have limited functional integration. However, the next generation of MINI-LINK radios will have more features in common with base stations: the systems, which remain physically separate, will have greater functional integration, especially in terms of network management. With increasing emphasis being put on the transmission network, functional integration is certain to add value not only to the MINI-LINK, but to Ericsson cellular and PCS systems as well.

Radio access applications

The introduction of multimedia services and the liberalization of the telecommunications market have created demands for broadband wireless systems. Public broadband services over radio, however, require bandwidth that is available only at relatively high microwave frequencies. Thus Ericsson Microwave Systems is actively de-

Box A
Abbreviations

ASIC	Application-specific integrated circuit
ATM	Asynchronous transfer mode
CAD	Computer-aided design
CTE	Coefficient of thermal expansion
DFM	Design for manufacture
DFT	Design for testability
FET	Field-effect transistor
GaAs	Gallium-arsenide
HBT	Heterojunction bipolar transistor
HEMT	High-electron-mobility transistor
IC	Integrated circuit
InP	Indium-phosphide
InP-HEMT	Indium-phosphide high-electron-mobility transistor
LNC	Low-noise converter
MBE	Molecular beam epitaxy
MCM	Multichip module
MESFET	Metal semiconductor field-effect transistor
MMIC	Microwave monolithic integrated circuit
MTBF	Mean time between failures
PCM	Process control monitor
PCS	Personal communications services
PHEMT	Pseudomorphic high-electron-mobility transistor
PLL	Phase-locked loop
PMP	Point-to-multipoint
SiGe	Silicon-germanium
VCO	Voltage-controlled oscillator
WLAN	Wireless local area network



Layer structure of transistor element

InGaAs	20 nm
AlGaAs	30 nm
InGaAs	15 nm
GaAs	0.5 µm

Figure 3
Schematic view of an MMIC. The semiconductor layer structure for high-frequency transistors are combined with "semi-insulating" GaAs which acts as a good dielectric. All active and passive components are built on the top of the GaAs crystal. The ground plane is at the bottom of the wafer.

veloping two access solutions for these frequencies: fixed broadband radio access systems and asynchronous transfer mode-based (ATM) wireless local area network (WLAN) systems.

The fixed broadband radio access systems will primarily address the professional business community. Network operators in this category are mainly expected to be new operators without an existing access infrastructure.

The ATM-based WLAN systems will provide high-speed multimedia services to individual users. They will also provide some support for mobility. Public and private network operators are envisaged.

Each of these areas of application represents a potentially huge market, where Ericsson Microwave Systems intends to play a major role. Given their base of the MINI-LINK, new MMIC technology, and Ericsson's market presence, their prospects are bright.

Development of microwave devices and MMICs

Microwave monolithic integrated circuits combine passive and active microwave components on a single chip. The principle behind MMICs is to combine "active" semiconductor materials (for transistors and diodes) with "passive" insulating materials (for transmission lines and passive components) in the same material structure. The entire structure is a single crystal semiconductor. Transistor layers are placed on top of the insulating layer. Where they are not needed, the transistor layers are etched away – in order to minimize parasitic capacitance.

Transistors are made by growing very thin (2 to 300 nm) semiconductor layers with different bandgaps (heterojunctions) on top of insulating gallium arsenide (GaAs). This technique, called bandgap engineering, revolutionized modern semiconductor physics by making it possible to grow such structures with molecular beam epitaxy (MBE). Typically, the overall thickness of the transistor layers is less than one percent of the total thickness of the chip, which is usually 50 to 100 µm. The active layers are magnified in Figure 3, which depicts the structure of a high-electron-mobility transistor (HEMT). For active field-effect transistor (FET) or HEMT-type components, the properties of the transistors depend mainly on the semiconductor material structure in combination with the geometrical dimension of the gate electrode.

Passive circuits may be a combination of lumped passive components, such as overlay capacitors, resistors and spiral inductors. At higher frequencies, components of a distributed transmission line are often used; for instance, open and shorted transmission line stubs.

The top of the insulating layer is used as the component plane. The bottom of the layer is used as the ground plane for microstrip transmission lines. Semi-insulating GaAs or indium phosphide (InP) is used as a dielectric for the transmission line.

Early attempts at making MMICs on silicon failed because of silicon's poor resistivity. The real breakthrough came when GaAs was used.

General MMIC development

Recent demonstrations have shown that MMICs based on GaAs and InP semi-insulating substrates may be used for frequencies up to and above 100 GHz. Current indium-phosphide high-electron-mobility transistor (InP-HEMT) technology represents state-of-the-art gain and low-noise performance: maximum frequency of oscillation is as high as 600 GHz, with a noise figure of 1.2 dB at 100 GHz.

Thanks to MMICs, engineers can now design and produce affordable microwave systems for the consumer market, including automotive radar (77 GHz), short-range communication links for automobiles (60 GHz), and wireless local area networks (at frequencies up to 60 GHz).

Various circuit functions must be implemented in microwave systems, including oscillators, low-noise amplifiers, mixers, power amplifiers, frequency multipliers, frequency dividers, and power detectors. The choice of semiconductor technology determines how well circuit functions can be reliably implemented at a reasonable cost.

Today's most common, mature device technologies for active microwave components are the metal semiconductor field-effect transistor (MESFET) and the pseudomorphic high-electron-mobility transistor (PHEMT) made on GaAs substrate. These devices make excellent low-noise amplifiers, power amplifiers, mixers, and multipliers up to millimeter-wave frequencies. Moreover, several foundries are able to produce them. Foundries are semiconductor device manufacturers who offer their processes to external customers, usually system houses.

MMIC VCOs

Because of simultaneous demands for voltage-frequency linearity, operating frequency bandwidth, and low phase noise, one of the most critical circuits in an MMIC is the "on-chip" varactor voltage-controlled oscillator (VCO). Ordinarily, an MMIC MESFET oscillator cannot achieve less phase noise than -70 dBc at 100 kHz from a 40 GHz carrier.

Although the performance of most reported PHEMT-based VCOs is similar to, or below, MESFET-based VCOs, recent results from an MMIC PHEMT VCO indicate very promising phase-noise results: -89 dBc at 100 kHz from a 25 GHz carrier. Other approaches involve silicon-germanium-based (SiGe) heterojunction bipolar transistors (HBT) made on silicone substrate, or HBTs based on GaAs substrates.

Since the HBT technology is relatively new, very few MMIC foundries currently offer it. SiGe HBT technology is even more exotic. Ericsson Microwave Systems is participating in a European project that aims to explore the potential of SiGe HBT technology for microwave radio circuits.

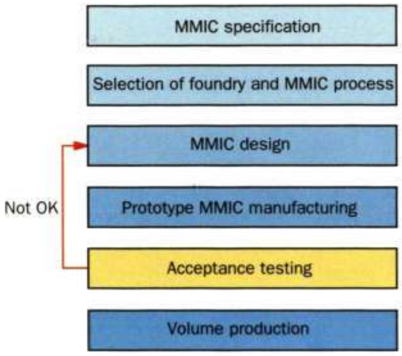


Figure 4
Steps in the MMIC design process.

Designing and manufacturing MMICs

A few years ago, the GaAs-based electronic market was considered a niche market restricted to special applications in the lower gigahertz range. Today, however, despite stiff competition from improved silicon-based devices, GaAs-based technology is well-established, having gained acceptance in several commercial markets, including the market for wireless communication. In response to growing demands for GaAs-based MMICs, a relatively large number of GaAs device manufacturers with worldwide operation currently offer foundry services.

Being a key technology, it is both important and cost-effective to have in-house design competence. Consequently, Ericsson Mircowave Systems design their own MMICs for the MINI-LINK program.

The MMIC design process requires a great deal of interaction between Ericsson Microwave Systems and the GaAs foundries (Figure 4). Ordinarily, the first steps of the design process include defining circuit specifications and deciding on the convenience

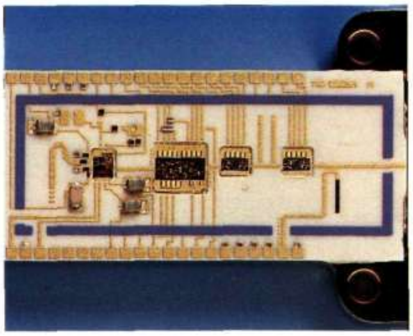


Figure 5
Photographed layout of the transmitter module. The frequency generator, power amplifier and multiplier MMICs are located in this module.

and suitability of an MMIC implementation. Input for these decisions includes the known capabilities of different MMIC technologies and their associated advantages/disadvantages.

An important issue to consider when selecting a foundry and MMIC process is the electrical performance of the process. Processes that yield devices with good electrical performance often cost more than others. However, they generally lead to simpler designs, which require fewer amplifier stages and consume less power. Thus, such processes result in devices that are more cost-effective.

Good electrical models – in the computer-aided design (CAD) tools – are also very important, since design iteration is expensive and time-consuming. In order to adequately account for the intrinsic parasitics of circuit elements, the elements need to be modeled by fairly complex equivalent circuits, which are frequently obtained through semi-empirical fittings of direct current (DC) and microwave measurements.

After the design is complete, a prototype mask set is manufactured and wafers are processed. Next, electrical performance and tolerances are checked by electrical measurements on various circuit elements of the wafer process control monitor (PCM). Approved wafers are then diced and delivered.

At Ericsson Microwave Systems, MMIC performance is measured and evaluated at the circuit and system levels. If the MMIC does not meet specified criteria, it may become necessary to alter the design and repeat the process. Accounting for results is not easy, since faults are often caused by several combined factors. Preferably, fault detection uses a combination of "cut-out" measurements and simulation tools (on single chips, cut-outs are key elements of the MMIC).

Once the MMIC prototype has been verified, preparations begin for volume production. A production mask set is manufactured, and the set of electrical verification tests for qualifying the processed wafer is revised. Ideally, only a few tests are needed to reflect the value of several parameters.

Actual configuration

Relevant frequency bands from 13 to 40 GHz are covered by means of several different MMICs in the transmitter and receiver modules:

- Frequency generator – one frequency generator, which is common to all sub-band modules, covers the frequency range 5 to 10 GHz. It includes a divide-by-eight pre-scaler that directly interfaces to the phase-locked-loop (PLL) divider.
- Multipliers – three multipliers are used to cover different frequencies.
- Power amplifiers – three power amplifiers are used to cover different frequencies. Secondary functions of the power-amplifier MMICs are to detect and control output power.
- Low-noise converters – three low-noise converters (LNC) are used to cover different frequencies. They include a resistive image-rejection mixer, which eliminates the need of a selection filter. The noise figures range from 4 to 7 dB, depending on the frequency band.

Engineers first began designing the MMIC chip set for the MINI-LINK transmitter and receiver modules in the autumn of 1996. Two different MMIC technologies were chosen. The multipliers, power amplifiers, and low-noise converters were based on the GaAs PHEMT technology (which gives low noise, good power, and high gain throughout the entire frequency band), whereas the

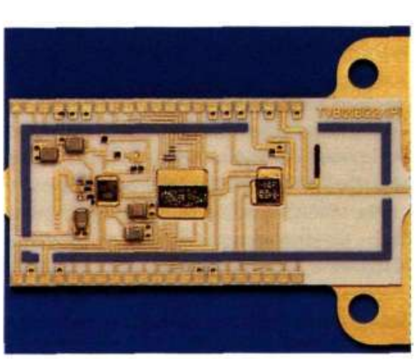


Figure 6
Photographed layout of the receiver module. The frequency generator, low-noise converter and multiplier MMICs are located in this module.

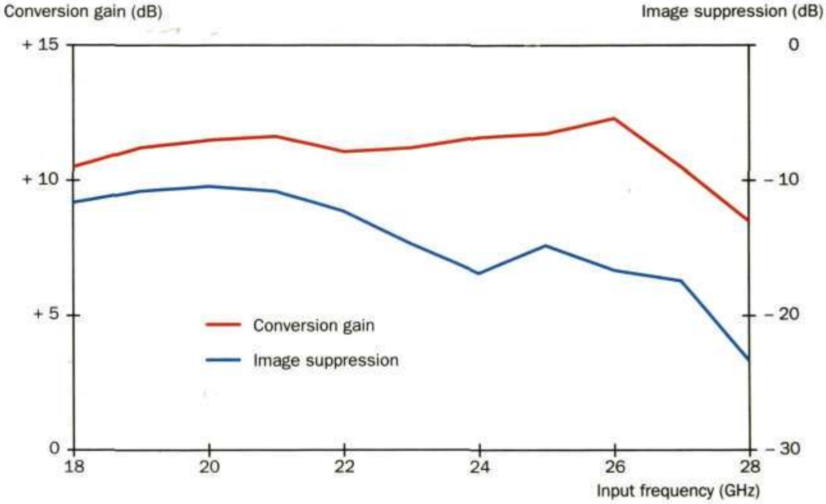


Figure 7
Measured conversion gain and image suppression for the low-noise converter. Upper and lower sideband measurements.

frequency generator was based on the GaAs HBT technology (because of its excellent low phase-noise characteristics).

The configuration layout of the transmitter and receiver modules are given in Figures 5 and 6, respectively. These show two of the MMICs, which are part of the modules. The low-noise converter amplifies and down-converts incoming signals to an intermediate frequency of 1 GHz. Measured conversion gain and image suppression (from the first version of the low-noise converter) are shown in Figure 7. Image suppression is better than 10 dB throughout the frequency range. Conversion gain is constant at about 11 dB up to 27 GHz, where it starts to decrease due to the limited bandwidth of the low-noise amplifier.

The first version of the multiplier, which was designed for an input frequency range of 5 to 10 GHz, has two output frequency ranges: 10 to 20 GHz and 20 to 40 GHz. Figure 8 shows the measured conversion gain of the x4 path.

Further integration

Compared with products based on present-day technology, the highly integrated MINI-LINK microwave modules are smaller and more reliable. Further integration

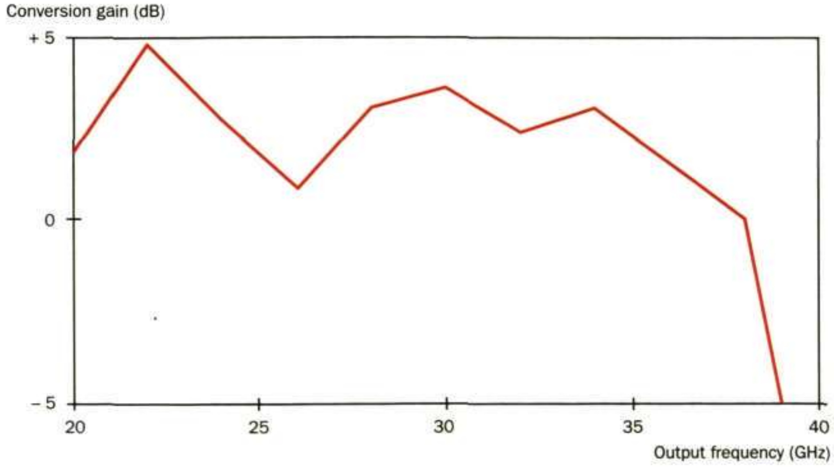


Figure 8
Measured conversion gain of the multiplier x4 path.

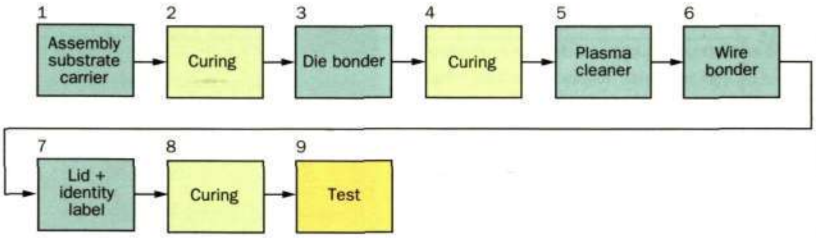


Figure B-1

Box B
Fully automated multichip module (MCM) production line

Process steps

1. Substrate adhesive is printed (screen printing) on metal carrier; mounting of substrate on carrier.
2. The adhesive cures.
3. Chip-mounting adhesive is dispensed; the chips are positioned (mounted).
4. Oven curing of adhesive.
5. The substrate and chips are cleaned with plasma.
6. Wedge bonding of chips to substrate.
7. Lid adhesive is dispensed; lids are placed over the substrate; the modules (MCM) are labeled.
8. The adhesive cures.
9. The modules (MCM) are tested (electrical testing) and packaged.

would result in larger chip areas and more complex circuitry, thereby decreasing the processing yield. An alternative approach, which would reduce the size significantly, is to integrate bias networks, phase-locked-loop circuits, and other low-frequency analog functions associated with the microwave modules.

Typical GaAs-based MMICs usually give the appearance of being very simple circuits. This is because they contain relatively few active devices: most of the chip area is occupied by passive circuitry. However, new developments in circuit topology and MMIC technology may allow an increase in functionality per area, which determines the overall cost of MMICs.

One way of increasing the package density of functionality involves three-dimensional MMIC structures, in which the active devices, resistors, and capacitors are fabricated on the GaAs substrate. Thin dielectric films and conductor layers, which form transmission lines and distributed circuit elements, are stacked on the substrate. Published experimental results indicate that the chip area of three-dimensional MMICs may be reduced to one-third the size of corresponding two-dimensional MMICs.

The TX-RX module

Packaging and interconnection technology

At present, the MINI-LINK program is one of very few advanced microwave product programs that uses mass-production principles. The introduction of MMIC-based microwave modules should not impair production, but should allow even more efficient processes. This puts stringent requirements on the interconnection and packaging system of the modules. For example:

- the component carriers must be designed to yield an infrastructure for GaAs-based MMIC chips whose
  a) microwave performance (low loss, high mechanical tolerance) is good;
  b) coefficient of thermal expansion (CTE) may be controlled;
  c) heat conductivity is good;
- the solder joints and the surface treatment of conductors and solder pads must withstand a reflow soldering process.

In order to meet these demands, Ericsson Microwave Systems is currently developing a new production system to manufacture multichip modules (MCM).

Production

To assemble the MMIC chips into complete MCMs, Ericsson Microwave Systems is installing a fully automatic production line (Box B) at their MINI-LINK production facility in Borås, Sweden. The focus of the new production line is on lead time and quality – two important parameters for improving production facilities.

The MCMs will be assembled in small batches. The assembly line can change MCM types between each batch with no set-up time.

Short lead time throughout the production flow frees up capital, shortens time to delivery, and facilitates more rapid feedback on the production status of previous stages or processes. This feedback may result in corrective actions, or be incorporated into a "best practice" for ongoing improvement.

Good quality and low production costs are mainly determined in the design phase. Therefore production engineers participate in designing the MCMs from the very start, ensuring that it is possible to test (design for testability, DFT) and manufacture (design for manufacture, DFM) the MCMs in an efficient way.

The quality of incoming materials – especially of the MMICs – is the most critical factor affecting the yield of MCMs. Good quality is achieved by establishing a good relationship with the foundries who supply MMICs, and by continually following up all incoming components.

Equipment suppliers for the assembly line were chosen on the basis of their technical experience and on their ability to work in a partnership.

The key parameters for the assembly line are reliability and proper function. Similarly, traceability and process control are the primary factors for achieving good mean time between failures. Thanks to the new production system, it will be possible to trace each MMIC to the supplier foundry from which it originated – even after a complete MINI-LINK product has been delivered and installed in the field.

Conclusion

The rapid evolution of digital and analog low-frequency ASICs has carried over into microwave technology giving rise to MMICs, which combine passive and active microwave components on a single chip. The MMIC technology enables designers to build greater functionality into much smaller, more power-efficient and reliable microwave products.

The benefits of this technology are many. Greater functionality and lower power consumption mean that operators can minimize their costs per site by installing smaller, highly integrated microwave radios. Further, being unobtrusive, the smaller-sized products are more readily accepted by the general public, giving operators the opportunity to expand their access networks into residential areas. Moreover, Ericsson Microwave Systems is applying MMIC technology to create fixed broadband radio access systems and ATM-based wireless LAN systems for public and private broadband services over radio.

References

1 Ek, M. and Genell, G., New RBS 884 base stations take IS-136 D-AMPS wireless services into new areas. Ericsson Review 74(1997:3), pp. 111-115.

Environment, for better or worse

Mats-Olov Hedblom

People and industries the world over are acknowledging that we can no longer take the environment for granted. The resolutions that were passed at the Rio Conference in 1992 charge every world society with the task of attaining a proper balance between their social, economic and environmental responsibilities. Partly in response to that conference, world industries have agreed to develop a suite of tools for environmental management – the ISO 14000 series of standards.

Corporations who have identified the environmental dimension as the most important aspect of good global citizenship are currently seeking tools that will help them to condense this dimension into strategic scenarios of the future. One such tool is the life-cycle assessment.

The author describes how Ericsson, as a member of the IT industry, is in the process of establishing a sound environmental platform on which to base their operations, now and in the future.

Part 1 of this three-part series of articles deals with various international perspectives of the environmental issue. Part 2 describes the life-cycle assessment in depth. Part 3 demonstrates how Ericsson can apply findings from the life-cycle assessment, designing for the environment and labeling products according to the emerging ISO 14025.

According to the Oxford English Dictionary, the word environment means: (a) "that which environs; the objects or the region surrounding anything"; (b) "the conditions under which any person or thing lives or is developed; the sum-total of influences which modify and determine the development of life or character."

According to the International Standardization Organization (ISO), environment means: "surroundings in which an organization operates, including air, water, land, natural resources, flora, fauna, humans, and their interaction."

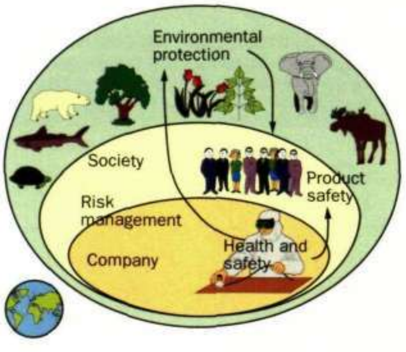


Figure 1
The external environment deals with all measurable changes that result from human activities.

From the biological perspective, the environment is "the total stock of genetic material produced from ecological evolution throughout the past billion years." Mankind has only recently begun to understand that we have a collective, controlling influence on the fate of this genetic capital – which also happens to be the basis of well-being in future generations.

Great confusion

The environment represents one of the greatest challenges with which society will have to deal. We are as yet unable to describe in any detail the enormous complexity and interdependence of the organisms of the world's many ecosystems. And as if that were not enough, each new day brings word of new environmental problems or incidents that can be attributed to human activities either directly or indirectly. These problems/incidents usually fall into one of three categories, of which the first comprehends environmental problems whose character has a global impact. Examples include:

- the diminishing stratospheric ozone layer – due to the use of chlorinated or brominated low-molecular-weight hydrocarbons;
- global warming – due to the extensive use of fossil fuels.

The second category involves problems that have a regional impact on the environment. For example:

- acid rain in the southern parts of Scandinavia and in the north-eastern parts of the US;
- a high concentration of ground-level ozone-forming compounds that diminish crop yields – due to discharges of nitrogen oxides and hydrocarbons (mainly from the transportation industry).

The third category includes problems that have a local impact on the environment. Very often these relate to the unwanted presence of "dirty" industries or landfill areas which, being perceived as polluters of ambient air, water or land, have a negative influence on the esthetic value of neighboring areas.

The environmental debate is heavily laden with major problems. On the one hand, we know too little about too many issues to be able to piece together a holistic view of the environment. For example, we can only guess at the outcome of global warming. After ten years of extensive research, which involved thousands of researchers, the International Panel on Climate Change (IPCC) concluded that we are already living in an era of global warming. However, the findings of another forum of researchers, the Global Climate Coalition (GCC), flatly repudiate this conclusion. Who is right? What are we to believe? Clearly, everyone agrees that we cannot afford to adopt a wait-and-see approach until after disaster strikes, but we are currently without a mechanism for dealing with the issues.

On the other hand, because of the complexity of major environmental issues, and because of the uncertainty that surrounds them, the issues are frequently debated separately instead of as parts of a greater interdependent whole. Moreover, environmental issues are sometimes deliberately lifted out of context – in order to sway political sympathies. For instance, when listening to a debate on the consequences of the peaceful use of nuclear power – which does not contribute to the global warming effect but could potentially contaminate land with relatively permanent radioactive waste – we are seldom told that a sustained increase in carbon dioxide concentrations in the air over the next 50 years will assuredly result in the flooding and subsequent loss of lowlands. This loss – which will be brought on by a general change in global weather patterns – will be many times greater than any potential outcome of the peaceful use of nuclear power.

Today in Sweden, many politicians advocate that a large percentage of nuclear and fossil fuels may be supplanted by biofuel derived from forest products. At the same time, however, in order to preserve the biological diversity of the forests, Scandinavian forest industries are negotiating a unique eco-labeling scheme for their products, which is based on sustainable forestry. If they succeed, then smaller quantities of biomass will be allowed out of the forests.

There is also a historical aspect to environmental issues. We regularly see this in north-south dialogues, where the industrialized nations of the northern hemisphere condemn the developing nations of the southern hemisphere for cutting down their rain forests, calling such action bad environmental behavior: However, the developing nations in the south legitimately wonder: "How is it that after having converted your forests into wealth and an improved standard of living in the north, you deny us the same opportunity?"

Box A
Abbreviations

EMAS	Eco-management and audit scheme (European environmental management system)
ENGO	Environmental non-governmental organization
EOL	End-of-life
EU	European Union
GCC	Global Climate Coalition
GNP	Gross national product
IPCC	International Panel on Climate Change
ISO	International Standardization Organization
IT	Information technology
LCA	Life-cycle assessment
NGO	Non-govermental organization
TCDD	Tetrachlorodibenzodioxin
WBCSD	World Business Council for Sustainable Development

The corporate perspective

From a corporate point of view, environmental issues fall roughly into four categories (Figure 1):

- health and safety in the workplace;
- product safety;
- environmental risk management;
- the external environment.

In this context, health and safety matters generally relate to production sites; pertinent guidelines and regulations mainly deal with protecting workers from direct or indirect harm. Environmental health- and safety-related issues are subject to national legislation. Recently, member countries who participate in the ISO rejected a proposal to draft an international health and safety standard.

The aim of product safety is to ensure that products pose neither direct nor indirect threats to the physical well-being of the persons who use them. The area of product safety is mostly governed by general laws, and in many countries these laws are tested in the courts. In the US, for example, several hundred people have filed a lawsuit against IBM for having manufactured keyboards whose use has caused irreversible damage to plaintiffs' arms, hands and fingers. Product safety-related issues that affect the information technology (IT) industry include the electromagnetic radiation from equipment that is shown to cause measurable changes in living cells.

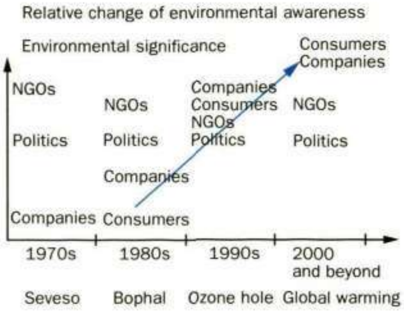
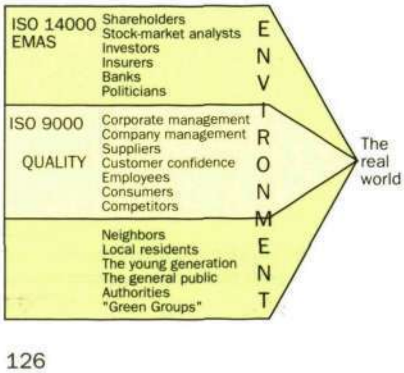


Figure 2
Well-educated "world citizens" from developed countries will force industry to declare their environment profile.

Figure 3
Environmental solutions must satisfy many important stakeholders.



Environmental risk management, which is part of a corporation's overall risk-management program, attempts to foresee and prepare against the negative environmental effects that a potential accident or disaster might cause. For instance, if dangerous chemicals are handled in any phase of a manufacturing process, then exceptional care must be exercised to ensure that the chemicals are not released into the ambient environment. The well-known Seveso accident – in which an aerosol mixture of sodium trichlorphenate, caustic soda, solvent, and other toxic compounds (dioxines, TCDD) escaped, severely contaminating neighboring regions – is one of numerous examples that serve to remind us that disasters do sometimes occur.

This article deals with what is specifically referred to as the external environment. From the Ericsson corporate perspective, the external environment must be considered for every interaction by the company, positive or negative, with the local, regional or global environment. Accordingly, as a member of the IT industry, Ericsson is in the process of establishing a sound environmental platform on which to base their operations, now and in the future.

Political agendas and the environment

People all over the world are acknowledging that we can no longer take the environment for granted. As a consequence, environmental issues are being added to all sorts of agendas. Not long ago, world politicians finally agreed to ban the production and use of freon, which destroys the stratospheric ozone layer. It is hoped that the politicians of the world will be equally effective in dealing with the even greater challenges of the future.

The Rio Conference (1992) is generally considered a shifting point in the worldwide political perspective of the environment. Its message – that nothing short of a complete transformation of attitudes and behavior will bring about necessary change – reflects the complexity of the problems facing us. Governments have recognized that poverty as well as excessive consumption by affluent populations create damaging environmental stress. We must redirect national and international plans and policies, to ensure that every economic decision takes environmental impact into full account. In light of this knowledge, governments adopted three major resolutions, whose aim is to change the traditional approach to development:

- Agenda 21 – a comprehensive program of global action for preparing the world for the next century.
- The Rio Declaration on Environment and Development – a series of principles that define the rights and responsibilities of individual states.
- The Statement of Forest Principles – a set of principles for the sustainable management of forests worldwide.

In addition, two conventions were opened for signature: the united framework convention on climate change, and the convention on biological diversity.

Because of its implications on the burning of fossil fuels, the convention on climate change remains an issue of great controversy, encountering strong opposition from all major oil-producing nations, including the US. The outcome of a summit meeting in Kyoto, Japan (December 1997) will be crucial to the issue of global warming.

The Agenda 21 resolution is expected to have the greatest influence on IT industries.

Local programs for implementing Agenda 21 have already started to pervade procurement requirements from communities and other political institutions in certain parts of the world, particularly in northern Europe.

Although a great many stakeholders frequently refer to this vision, no one knows for certain what is or is not sustainable: the environment is simply too complex to be treated in depth by a single stakeholder. Therefore, although they may address environmental issues, stakeholders tend to discuss them one at a time, frequently using the environmental debate to disguise more primitive interests.

Market uncertainty

How market forces will respond to environmental issues is still an open question. Relatively indifferent in the 1970s, consumers began to show increasing awareness of environmental issues during the 1980s (Figure 2). Today environmental awareness is quite high among consumers, especially in industrialized nations, which account for most of the world's gross national product (GNP).

Environmental non-governmental organizations (ENGO), which maintain a neutral and very credible position among numerous stakeholders, constitute a major force for environmental awareness and greatly influence many consumers.

Through their purchasing behavior, well-educated consumers in developed nations are expected to pay much more attention to environmental, and possibly ethical, labeling.

The greening of industry

Partly in response to the political overtones of the Rio Conference, world industries have agreed to develop a suite of tools for environmental management – the ISO 14000 series of standards. Although only environmental management and auditing standards have been produced to date, a growing number of large corporations, including Ericsson, have elected to introduce ISO 14001 (the environmental management standard) before the end of this millennium. This collective action is certain to have a great positive influence on the environment.

One part of ISO 14001 states that customers are required to verify the environmental profile of their suppliers. In this sense, ISO 14001 is very similar to ISO 9001 (the ISO standard for quality management), which requires customers to verify that their suppliers fulfill certain basic requirements for quality. Shortly after its introduction, many of the companies who adopted ISO 9001 began requiring their suppliers to implement ISO 9000 systems. The general response to ISO 14001 is expected to evolve in much the same way. By comparison, the environmental standard will be much more difficult to implement than ISO 9001, since environmental issues are more complex than quality-related issues.

Stakeholders interested in environmental issues include neighbors; local, regional and state political bodies; environmental non-governmental organizations; and representatives from the stock market, financial institutions, and insurance companies (Figure 3). In particular, indicators suggest that financial stakeholders have a growing interest in the environment:

- Major banks throughout the world are hiring their own in-house environmental experts.
- Numerous banks are issuing environmental questionnaires to their business customers.
- Financial analysts are requesting environmental information in order to complete their evaluations.
- Insurance companies are teaming up to draw attention to the conceivable relationship between disasters and global warming.

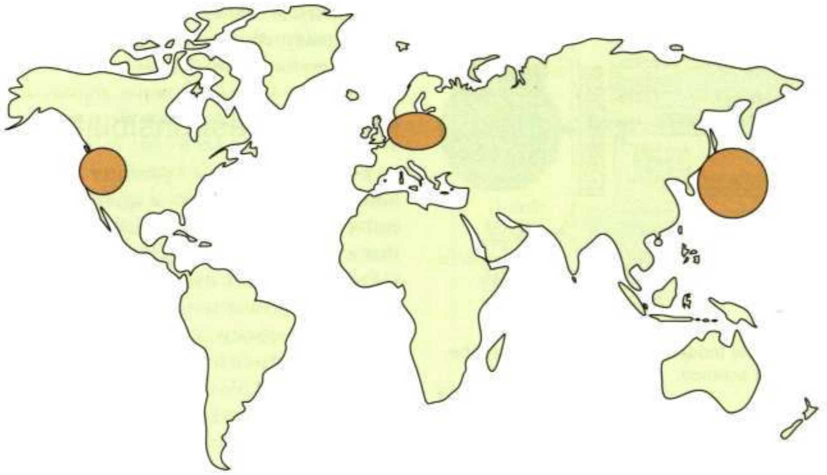


Figure 4
In terms of IT, the most developed environmental areas also happen to be the most significant ones.

Environmental map

Interest in environmental issues varies greatly in different societies.

Today Germany is the "greenest" of industrial societies. In the beginning of 1997, more than 600 German organizations had registered for the Eco-management and Audit Scheme (EMAS), which is the European environmental management system.

Running a close second, Japan expects as many as 700 companies to be certified according to ISO 14001 by the end of 1997. Initially the electronics industry will account for more than two-thirds of all certification. However, every major Japanese company intends to have their foreign and domestic facilities certified by the end of 1998[1]. The general observation is that Japanese industries are preparing themselves for any future environmental contingency.

In North America, the State of California leads the way on environmental policy.

It is interesting to note that the IT industry has congregated in those parts of the world that are most environmentally mature (Figure 4). Even so, some US corporations feel that ISO certification can wait, claiming that the costs of certification outweigh

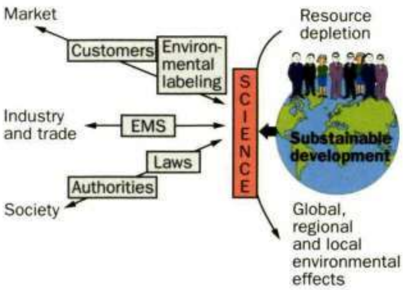
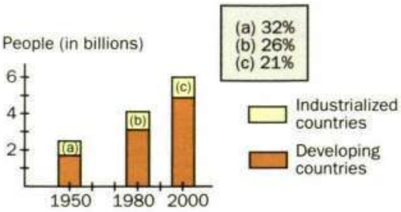


Figure 5
Sustainable industrial development must be based on science.

Figure 6
By the year 2050, ten billion people are expected to populate our world.



its benefits, or that existing legislative pressure already ensures world-class environmental performance.

Producer responsibility

In Europe, the idea of "producer responsibility" has been on political agendas for several years. Producer responsibility means that a process whose chain begins by adding value to a product does not end when the product is sold on the market, but must also include some means of reclaiming the product in an environmentally responsible way at the end of its life (end of life, EOL). This idea was first introduced in countries with limited space for landfills. Today, however, producer responsibility is gaining support elsewhere too, as more and more people balk at the idea of throwing valuable resources into refuse dumps. Moreover, because many landfills contain hazardous or poisonous compounds that are slowly migrating into the environment, the landfills represent significant ecological problems.

Recently, the member countries of the European Union (EU) were unable to agree on a common legislative solution to producer responsibility for electric and electronic products. Instead, each member country in northern Europe, along with Norway and Switzerland, is drafting its own legislation – the unfortunate consequence being that no two laws will be the same.

Regardless of how the laws are worded, producers who design for the environment will benefit, since the emphasis of EOL is on environmentally responsible disintegration. Once producer responsibility laws are passed, deposition in landfills is certain to cost more or be completely forbidden. Consequently, designers of products will be forced to take into account the environmental effects of their products – and the system that produces them – throughout their entire life cycle. Similarly, greater emphasis will be placed on product ownership, since every project for new products will have to assume responsibility for the material content and the cost profile of end-of-life treatment, which may not take place until many years after the product is actually sold.

Science as the environmental platform

Some industries that deal in raw materials, such as the forestry industry in northern Europe and North America, have made great headway in introducing environmental management systems. They are proud of their achievements and want to tell their customers, using various kinds of eco-labeling.

The electronics industry, which is expected to follow suit, will begin marketing its environmental image some time before the year 2000. In Europe, at least one large electronics company (Siemens) has already started to base its marketing strategy on environmental themes. Most organizations, however, are still trying to figure out what environmental performance really means (Figure 5).

While there is no simple answer, the introduction of an environmental management system should be considered for the short term as well as for the long term, making allowances for varying degrees of maturation in different parts of the world. As with any other managerial function, environmental management is valued for its ability to contribute to the bottom line. Today no corporation believes it can sidestep environmental issues without incurring some negative consequences (badwill).

In a newly published report[2], the World Business Council for Sustainable Development (WBCSD) recommends several ways in which the financial community can "improve the quality of its decision-making process by integrating environmental factors into its analysis." The report, which explores how established models of shareholder value can be expanded to include the positive and negative effects of environmental issues, acknowledges that environmental drivers are "one of many dimensions" a company can use to create a competitive edge, and notes further that companies who ignore such drivers "miss an important element of competitive advantage." The report concludes that "the markets will increasingly come to recognize that eco-efficient companies are well-run companies."

In the short term, the first order of business is to introduce environmental management systems that will help corporations to structure and make sense of the complex environmental reality. In the long term, looking beyond the year 2000, many different branches of industry will compete to cater for the functional needs of society at the lowest possible cost to the environment. It is here that the telecommunications and infocommunication industries – indeed the entire IT industry – have their greatest opportunity to expand as a branch, not only because of their technical cost-benefit performance, but also because they offer society a means of approaching the problem of sustaining itself. In qualifying for and remaining in this enviable position, the telecommunications and infocommunication industries must rely heavily on science, ensuring that money and efforts are invested for the purpose of building up a sustainable society. A vital tool for this task is the life-cycle assessment (LCA), in one of its developed forms.

Trying to put the environment into its proper perspective and then agreeing on some common values is no easy task. Fortunately we are shifting from a state in which nature and the environment are either free or where ownership is based on supporting human societies, to a state where energy consumption, emissions, and the extraction of raw materials are judged through newly critical eyes. Obviously there is a strong interdependence between economic output and the normal use, or potential misuse, of the earth's natural resources. Some life-cycle assessments have tried to accommodate this interdependency by introducing "willingness to pay" as the basis on which to judge the results. Needless to say, this approach is risky, since what people are willing to pay to save the environment does not necessarily measure up to what is needed, from a scientific point of view, to produce a sustainable environment.

The environmental challenge

By the year 2000, nearly six billion people are expected to populate our world; by the year 2050, this number is expected to reach ten billion (Figure 6).

What would happen if everyone on the earth today enjoyed the same standard of living as those living in the industrialized parts of the world? Does the crust of the earth contain enough necessary raw materials? Is there enough available land to make this feasible? Researchers of future human societies agree that if ten billion people are to share this planet in the next century with a decent standard of living, then certain fundamental changes must take place. In coming years, various societal activities will be measured in terms of their contribution to the environmental load (Figure 7).

The aggregate measurement of society's activities makes up the total environmental load, which is the sum of all negative influences on the environment. It is also a measure (index) of sustainability.

Although scientific data is incomplete, estimates show that the developed economies of the West must, during the next century, reduce their share of environmental load by as much as 10 to 20 times present levels. No one knows how we are to do this, or what the real consequences will be. However, compared with any other physical solution, researchers agree that the use of information technology for communication may help lower the total environmental load by up to 10,000 times present levels. Thus IT has the potential to advance future societies substantially without requiring them to sacrifice much in the way of living standards.

No one expects that worldwide change will voluntarily take place overnight. Certainly the rate of change will differ in various parts of the world. Regardless of how developments unfold, the short-term priorities of nearly every business will remain a sound economy and customer satisfaction. In the long term, however, businesses – especially large, global corporations – will assign top priority to survivability issues, of which being a good global citizen is one.

Corporations who have identified the environmental dimension as the most important aspect of good global citizenship are currently seeking tools that will help them to condense this dimension into strategic scenarios of the future. The need for such tools is great, as the market wants to know corporate policy, and is demanding to receive actual data on emissions and the consumption of resources. One tool, and the scientific method that best describes the environmental dimension, is the life-cycle assessment.

End of Part 1. Part 2 describes the life-cycle assessment in depth. Part 3 demonstrates how Ericsson can apply findings from the life-cycle assessment, designing for the environment and labeling products under development according to ISO 14025.

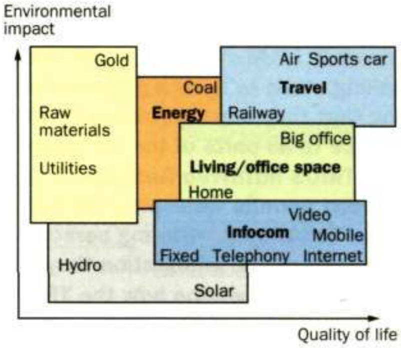


Figure 7
Infocom offers solutions with a low environmental impact that will be further improved through Moore's law.

References

1 Cutter Information Corp.: Business and the environment's ISO 14000 update™, "Japanese Firms Eagerly Adopting ISO 14001 – with Official Encouragement", June 1997)

2 Cutter Information Corp.: "Business and the environment®, "WBCSD Report Finds Strong Links Between Environmental Performance and Shareholder Value"; June 1997)

Distributed telecommunications network access using the TMOS IntraWeb Gateway

Magnus Ekhed, Peter Gundersen and Olav Queseth

In an increasingly competitive marketplace, network operator success depends more and more on network usage. In optimizing usage, operators invest in O&M systems, which control traffic and network quality. Next, in taking steps to uphold good service levels for customers, operators are finding that they must distribute information on network status and performance to all parts of the organization.

The TMOS IntraWeb Gateway from Ericsson Hewlett-Packard Telecommunications permits widespread access to network information through corporate IT networks, allowing persons outside an O&M center to access the network-related information they need.

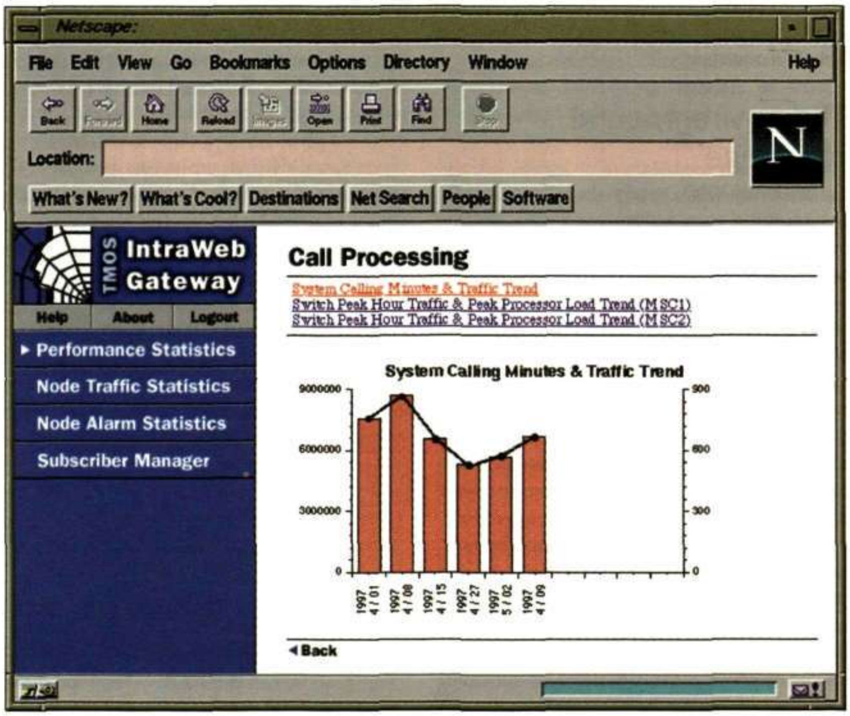
The authors describe how the TMOS IntraWeb Gateway makes new functions and interfaces available anywhere in the corporate network, thereby improving internal work flow and enabling network operators to give their customers better service.

Traditionally, operations support systems (OSS) were created to meet the needs of the operation and maintenance (O&M) organization, whose job is to operate the telecommunications network. Consequently, personnel at O&M centers have a good picture of network status and performance. Direct access to network information has always been their privilege. Other departments have received network status reports distributed on paper, or in some cases, by electronic mail (e-mail).

As the market changes, network operators are refining their business processes to make their operatios more cost-effective, and to provide better service and shorter time to market. In doing so, they want to distribute information on network status and performance to new categories of users within their organizations. A typical example involves distributing reports on traffic performance measurements on a regular basis to different parts of the organization.

TMOS IntraWeb Gateway

The TMOS IntraWeb Gateway (TIG) concept, from Ericsson Hewlett-Packard Telecommunications (EHPT), applies Web technology to give managers, customer-care departments, marketing departments, and network planners direct access to user-specific information and reports from an operations support system. It does so by introducing a gateway – but without creating dependencies – between telecommunications management and operations support (TMOS) and the corporate information technology (IT) network. Thus, by using an ordinary Web browser, company personnel anywhere in the corporate IT network outside the operation and maintenance center (OMC) may access a subset of functions and information in TMOS and execute standard tasks in the telecommunications network.



Call Processing

System Calling Minutes & Traffic Trend
Switch Peak Hour Traffic & Peak Processor Load Trend (M SC1)
Switch Peak Hour Traffic & Peak Processor Load Trend (M SC2)

Figure 1
Information distribution – example of distributed traffic information from TMOS via TMOS IntraWeb Gateway.

The TMOS IntraWeb Gateway helps personnel at the operation and maintenance center to spread network information within the operator organization, and allows authorized persons outside the OMC to perform tasks that help customers. The internal work flow is improved, and customers get better service.

The TMOS IntraWeb Gateway, which is fully scaleable and easily customized, also helps network operators to get the most out of existing IT infrastructure. Web server hardware and customized software, along with tools and system interfaces, make up a controlled, secure, user-friendly link through which staff access TMOS information.

A Web-based solution allows standard Web-client software, such as the Netscape Navigator, to be used for browsing information. Thus information to and from operations support systems (such as the TMOS eXchange Manager), and from business support systems (such as TIMS) may be integrated into and displayed on existing clients. The TMOS IntraWeb Gateway may also be customized to access operations support systems (other than TMOS), seamlessly integrating operations support systems in multi-vendor environments into a single user interface.

Fetching and transferring information

The TMOS IntraWeb Gateway enables employees to fetch information from TMOS as well as to transfer information to network elements that are accessed through TMOS by means of

- scheduled transfers from TMOS to the Web server;
- simple, controlled actions, which are activated from Web pages and sent from the Web server to TMOS.

Information distribution

In the past, when a marketing department needed performance measurement data for planning business activities, they usually had to order that data from the operation and maintenance center – which meant that they often had to wait for the data. Moreover, when the data arrived, the marketing department usually had to reformat it before they could use it.

The TMOS IntraWeb Gateway enables operators to schedule regular data transfers from TMOS to a Web server, which greatly improves quality and speeds up the distribution of information to end-users (Figure 2). Furthermore, by being able to schedule data transfer, operators can better regulate the impact of load on TMOS.

Software in the Web server transforms data into a variety of reports, which are published on the operator's corporate intranet. The format of the reports may vary from simple spreadsheets to automatically updated graphs. Thus, thanks to the TMOS IntraWeb Gateway, when someone at the marketing department wants to view an up-to-date report, he or she can find it on the Web server.

To better support customer processes, managers may also use the TMOS IntraWeb Gateway for creating and distributing customized presentations of TMOS data. Access restrictions guarantee that only authorized persons access the reports.

Furthermore, the TMOS IntraWeb Gateway enables end-users with little or no TMOS experience to obtain information from TMOS over the intranet.

Process automation

Previously, when a subscriber called to report that his or her voice-mail service was blocked, the customer-care department could do little more than submit a work order to the OMC, since their own administrative system did not provide a means of activating the subscription. The TMOS IntraWeb Gateway allows customer-care staff to perform such simple actions. They do so by filling in predefined forms displayed in their browser. A command sequence is then generated and sent through

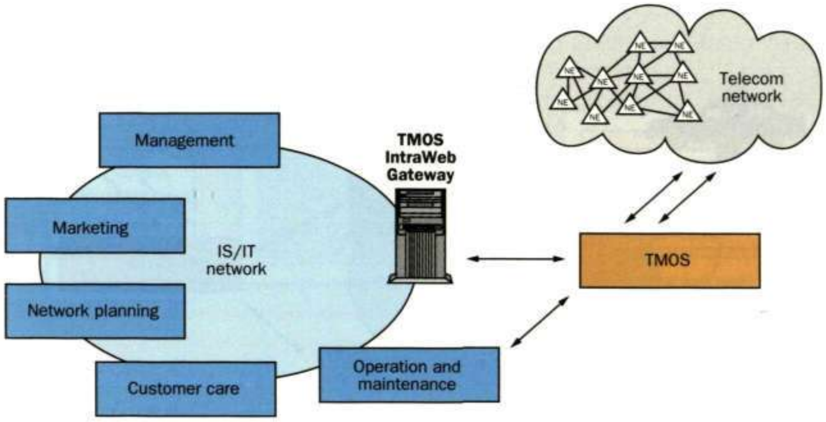


Figure 2
The TMOS IntraWeb Gateway gives managers, customer-care departments, marketing departments, and network planners direct access to user-specific information and reports from an operations support system.

Box A
Abbreviations

CLI	Command-line interface
EHPT	Ericsson Hewlett-Packard Telecommunications
IS/IT	Information systems and information technology
IT	Information technology
MML	Man-machine language
O&M	Operation and maintenance
OMC	Operation and maintenance center
OSS	Operations support system
SSL	Secure sockets layer
TCP/IP	Transmission control protocol/Internet protocol
TIG	TMOS IntraWeb Gateway
TMOS	Telecommunications management and operations support

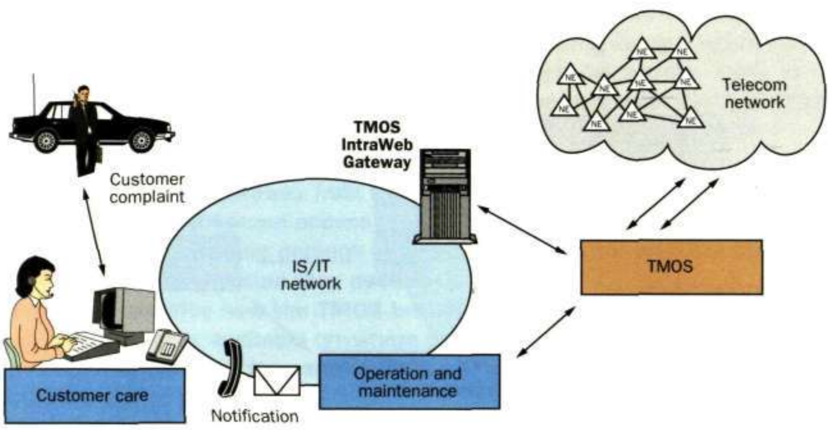


Figure 3
Process automation – the TMOS IntraWeb Gateway enables customer-care staff to perform simple O&M actions, such as restoring a customer's voice-mail service, if blocked.

the command-handling facilities in TMOS. The customer-care staff may observe changes in network status as the commands are executed. Using the TMOS IntraWeb Gateway concept, customer-care staff may actually restore certain kinds of customer service as the problem is being reported (Figure 3).

Product description

The TMOS IntraWeb Gateway is delivered as a basic package (Figure 4) with optional gateway customization consultancy service. The basic package contains software from EHPT that provides high-level interfaces to TMOS functions. The interfaces ensure that customer adaptations will run with any TMOS software.

The basic package also contains standard third-party hardware and software. Currently, these include an operating system (Windows NT), a Web server (Netscape Enterprise server), a Web-authoring tool (Netscape Navigator Gold), a programming language (Java), and optional firewall software (Firewall 1).

TMOS IntraWeb Gateway module

The functionality of the basic package may be divided into three categories:

- TMOS access blocks.
- The TIG toolbox.
- TIG applications.

TMOS access blocks

The TMOS access blocks – which consist of a man-machine language (MML) access function block, a command-line interface (CLI) access function block, and a database access function block – add stable interfaces to TMOS functionality.

The man-machine language access function block enables TIG application programmers to connect to network elements that are linked to the TMOS system. The functionality allows users to send commands, and to choose between immediate or delayed responses.

The command-line interface access function block provides functions for accessing command-line interfaces to TMOS functions. Command-line interfaces are programs that are run from a command line. The program results, which are normally printed on the screen, may be used in the TIG application.

The database access function block gives operators access to databases located in TMOS servers. Data that is retrieved from the databases may either be processed in TIG applications, or displayed directly in the browser.

The TIG toolbox

The TIG toolbox, which is used for creating TIG functionality, contains functions for user and authorization handling, functions for logging and auditing applications, support for generating spreadsheets, and parsing tools.

The user- and authorization-handling function block provides functions for handling user information and some initial login data. The function block mainly handles user authentication by establishing the identity of TIG users. This is done either through secure sockets layer (SSL) certificates or by means of login and password combinations.

When users have been authenticated, the function block generates a list of the applications they are authorized to run. Obviously, the list may differ for different users.

A Web-based interface facilitates the administration of user profiles.

The spreadsheet generation function block helps TIG application programmers to generate output in the form of a spreadsheet directly in the TIG application. It also provides rules and guidelines that help TIG application programmers to program efficiently.

The logging and auditing function block gives TIG application developers a uniform way of logging events that occur within TIG applications. The function block also contains functions that enable users to browse the logs generated by the applications.

Because most TIG applications need to parse input from HTML forms, *the parsing tools function block* contains functions that help TIG application programmers with this task. The function block also screens the input in order to improve security.

TIG applications

The basic package currently includes two applications: one for managing network performance and one for managing faults in the network.

The performance statistics application enables TIG users to view reports on the performance measurement statistics collected by TMOS. Predefined reports are distributed via the Web and displayed in the Web browser.

The alarm statistics application enables TIG users to view TMOS alarm logs. Users may retrieve historical data for a specified time interval by selecting the attributes of the alarms they want to see. The data generated from this application is displayed as a graph in HTML format, or in a spreadsheet (MS Excel).

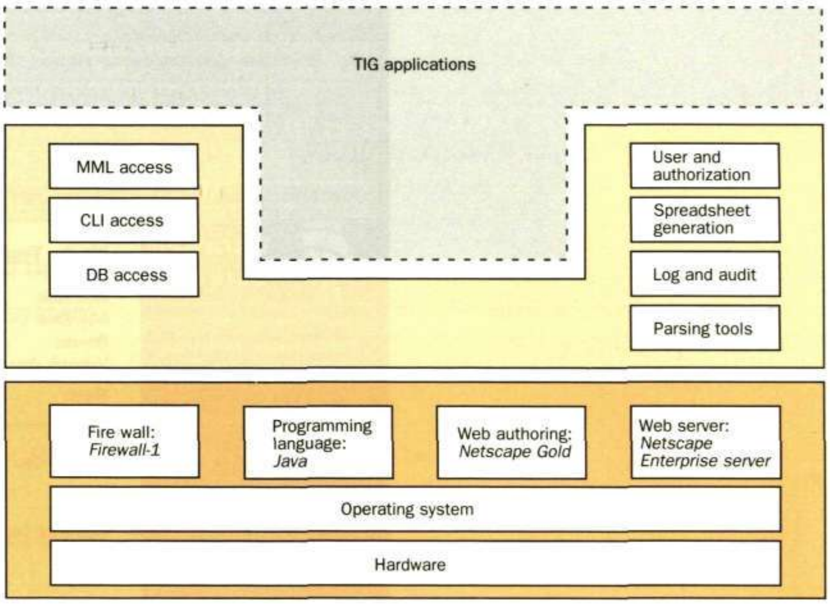


Figure 4
The TMOS IntraWeb Gateway consists of third-party products, access blocks to TMOS, the TIG toolbox, and the TIG applications.

TIG design environment

The design environment forms the basis for a certification program that will allow partners to develop TIG applications.

When used in combination, the TIG toolbox, style guides, templates, and specified design rules speed up the development of TIG applications and customer adaptations, and ensure that each application has the same look and feel. Adherence to the design rules also guarantees the security of TMOS, and keeps maintenance and upgrade costs down to a minimum.

Security

Because the TMOS IntraWeb Gateway allows users to access TMOS from the Web, security aspects must be considered very carefully. There are many ways in which an intruder can harm a computer system. For example, besides damaging hardware through theft or sabotage, intruders may try to attack a server in order to gain access to the super-user account. If they succeed in this, they can modify data, listen to traffic on the local area network, intercept data sent through the network, or even modify the telecommunications network. Thus, every hole through which a would-be intruder might seek to gain access must be closed. To this end, measures (firewall, and screening of user input) have been taken to prevent intrusion, to control access (authentication and authorization procedures), and to shut out eavesdroppers (encryption).

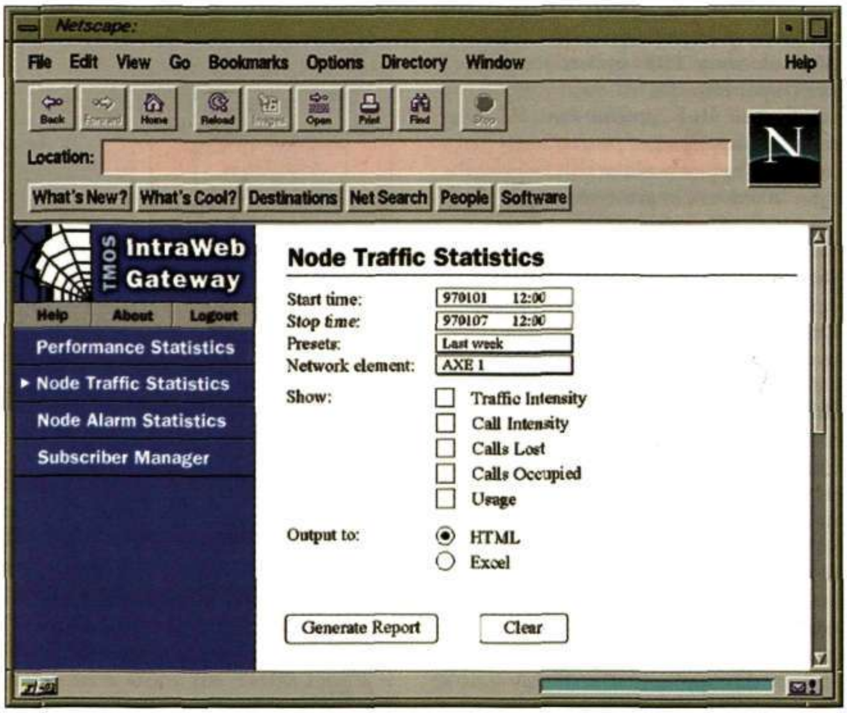
The TMOS IntraWeb Gateway was designed with security in mind. Likewise, the selection of each accompanying third-party product has been based on compliance with the security framework. Firewalls, security levels, and encryption capabilities, which were developed in collaboration with Hewlett-Packard, have been designed to fulfill customer needs. Network and access security are embedded in the TMOS IntraWeb Gateway system.

Authentication

The login procedure, known as the user and authority handling function, screens each person who logs into the system with SSL certificates, or by querying them for a user name and password. This procedure ensures that only authorized personnel view or access data in TMOS.

Authorization

The most valuable part of TMOS is the data it contains – for example, traffic data and customer data are valuable to competitors.


Netscape:
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
Location:
What's New? What's Cool? Destinations Net Search People Software
TMOS IntraWeb Gateway
Help About Logout
Performance Statistics
▸ Node Traffic Statistics
Node Alarm Statistics
Subscriber Manager
Generate Report
Clear

Node Traffic Statistics

Start time:	970101 12:00
Stop time:	970107 12:00
Presets:	Last week
Network element:	AXE 1
Show:	□ Traffic Intensity □ Call Intensity □ Calls Lost □ Calls Occupied □ Usage
Output to:	● HTML ○ Excel

Figure 5
TMOS IntraWeb Gateway gives authorized users easy access to telecommunications network data anywhere in the corporate network.

Therefore, adequate measures must be taken to protect this data from harm. Only authorized persons must be able to access it. And then, only such categories of data as they need to access. The TMOS IntraWeb Gateway uses a built-in security system which ensures that no one but authorized users may view data, change data, or remove it from the system.

Encryption

An encryption technique is used to prevent information leaving the protected TMOS zone from being intercepted or replaced with false information. Encryption is implemented using the secure sockets layer standard and third-party products (Netscape).

Firewall

The TMOS IntraWeb Gateway has two separate network interface boards: one is connected to the corporate IT network, and the other one is connected to the TMOS network. Each unnecessary service is turned off and a Web server is installed on the gateway machine. In addition, firewall software is installed on the gateway machine – the software logs all server activities, generating alarms for unusual events, and filtering out unwanted TCP/IP packets.

Conclusion

The TMOS IntraWeb Gateway concept works with any version of TMOS, giving several user categories outside the operation and maintenance center access to the system. It permits operators to distribute high-quality real-time information to new user categories via a Web browser. It also enables OMC personnel to report on network status, and allows other departments to perform standard network operations. By integrating and automating work procedures, network operators can improve work flow and give customers better service.

Various security measures control information access.

The efficient transfer of information from the telecommunications network to different parts of the operator organization shortens lead time for customer services and makes better use of investments in TMOS. The TMOS IntraWeb Gateway concept also allows operators to extract more from existing IS/IT investments, without creating dependencies between TMOS and corporate IT networks.

Contents

Previous issues

No. 3, 1996

No. 4, 1996

No. 1, 1997

No. 2, 1997

New patents within Ericsson

CLOCK SELECTION

Peter Lundh , Mats Wilhelmsson, Anders Bjenne

Patent number 9503370-0 /P06576

MULTI STANDARD MODEM

Per Stein

Patent number 5628055/P05840

DTX FACCH SYNCHRONIZATION

Francois Sawyer

Patent number 5625872/P06300

TALKING MOBILE PHONE

Björn Ekelund

Patent number 5630205/P06128

MULTI MODE SIGNAL PROCESSING2

Paul Dent

Patent number 5629655/P06252

SELF DIAGNOSTIC OF BUFFERS

Mats Ernkell, Stefan Sahl

Patent number 5633878/P06346

MULTI-CHANNEL AUTOBAUDING CIRCUIT

Simon Mak, William Wong

Patent number 5627858/P05654

LARGED PHASED-ARRAY

Paul Dent

Patent number 5619210/P06064

SUBSCRIBER BEHAVIOR LOGGING

Brian Womble

Patent number 5488648/P05983

MOBILE TERMINAL LOCATOR

Daniel Dufour

Patent number 5613205/P06420

ERROR HANDLING & RECOVERY

Anders Jönsson, Ulf Winberg, Charles Lignell, Chung Lee, Peter Larsen

Patent number 5594861/P06550

PARAGIM OF ACTIVE SUBJECTS2

Håkan Larsson, Kerstin Odling, Åke Rosberg, Håkan Karlsson

Patent number 5572727/P06369

INTELLIGENT NETWORK SELECTION

Patrik Nilsson

Patent number 504866/P06436

DUST DETECTION

Sasan Esmaeili

Patent number 504861/P06436

LOW RESISTANCE OHMIC CONTACT

Bertil Kronlund

Patent number 504916/P06307

UNSYMMETRICAAL DYNAMIC ROUTING

Hans Andersson

Patent Number 504712/P06486

TRACKING RAYS

Jan Färjh

Patent number 504622/P06761

FLOATING LVDS

Mats Hedberg

Patent number 9502715-7/P06495

A CONNECTOR

Uno Henningsson, Michael Lynn

Patent Number 9502896-5/P06930

AN ELECTROOPICAL CIRCUIT

Uno Henningsson, Michael Lynn

Patent number 9502895-7/P06931

FLEXIBLE START SYNCHRONISATION

Anwar Chivi, Richard Adjimah

Patent number 9502216-6/P06473

MAJORITY VOTE FUNCTIONS

Peter Lundh, Anders Bjenne

Patent number 9503421-1/P06571

ATM SWITCH CORE II

Tawfik Lazraq, P-O Bergstedt, Hannu Tenhunen, Mehran Mokhtari

Patent number 9501720-8/P07215

MODULAR COVERAGE EXTENSION

Hans Mähler

Patent number 504992/P06557

ESD/EMC PROTECTION BETWEEN CABINETS

Kent Enström, Hans-Olov Essland, Björn Kassman, Anders Svensson, Erik Torhage

Patent number 505004/P06737

IMPULSE SIGNAL CONVERTER

Svante Axling, Juan Hernandez

Patent number 5638435/P06480

SUPPORTING EXTENDED RANGE

David Smith

Patent number 5642355/P06609

DCCH BACKUP SOLUTION

Richard Brunner, Daniel Dulong

Patent number 5541978/P06165

HIGH FREQUENCY BALUN

Jerzy Dabrowski

Patent number 5644272/P06881

COOLING OF OUTDOOR ELECTRONICS

Hans Mähler

Patent number 504950/P06558

ANTENNA WITH GROUNDPLANE

Christer Andersson

Patent number 505074/P06559

FEQUENCY SELECTIVE RADOME

Stefan Johansson, Tomas Stanek

Patent number 504815/P06508

PHASE LOCKED OSCILLATOR

Björn Lofter, Glenn Sjöberg

Patent number 505090/P06572

NOT NOW MOBILE

Lucia Suarez Haces, Hannu Vainiomäki, Christina Birkhammar, José Pons, Dagmar de Rooy

Patent number 505175/P06265

DYNAMIC INFRASTRUCTURE

Staffan Andersson, Erik Bogren, Torgny Lindberg, Lars Novak

Patent number 9501543-4/P06394

STABILIZED POLYSILLICON RESISTOR

Ulf Smith, Matts Rydberg, Håkan Hansson

Patent number 9503198-5/P06545

INTERFACE MODEL

Ola Smith

Patent number 9402921-2/P06169

Telefonaktiebolaget L M Ericsson
SE-126 25 Stockholm, Sweden
Tele: +46 8 7190000
Fax: +46 8 6812710

ISSN 0014-0171
Ljungföretagen, Örebro 1997

Ericsson REVIEW

THE TELECOMMUNICATIONS TECHNOLOGY JOURNAL

NO 4, 1997

ERICSSON

Ericsson REVIEW

The telecommunications technology journal

The purpose of Ericsson Review is to report on the research, development and production achievements made in telecommunications technology at Ericsson. The journal is published in English and Spanish and distributed quarterly to readers in more than 130 countries.

Address:
Telefonaktiebolaget LM Ericsson
S-126 25 Stockholm, Sweden

Tel: +46 8 719 00 00

Fax: +46 8 681 27 10

Internet address:
http://www.ericsson.com

Publisher: Håkan Jansson

Editor: Eric Peterson

Production manager: Eva Karlstein

Layout: Paues Media, Stockholm

Printer: Ljungföretagen, Örebro

ISSN: 0014-0171

Volume: 74, 1997

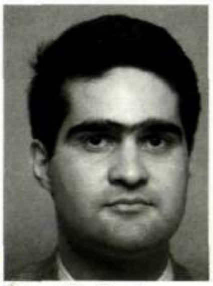
Cover: Geographic information systems are extremely valuable for high-speed boats and ships that operate near the coastline or in an archipelago. They may be combined with a GPS to indicate a vessel's position.

Contents

Contributors

In this issue

Allan Hansson


Robert Nedjeral

Ingmar Tönnby

Jan Bergkvist

Hans Brandtberg

Johan Frössling

Lena Lüning

Mats Åkerlund

Mikael Ronström

Mats-Olov Hedblom

Allan Hansson, has participated in the development of several systems for telephony and data communication, since joining Ericsson in 1974. Today he works as a member of the strategic system planning unit at Ericsson Utvecklings AB. He holds an MSc in electrical engineering from Chalmers University of Technology, Göteborg.

Robert Nedjeral joined the Public Networks Internet Program at Ericsson Telecom AB in July 1997, where he works as the product manager of Phone Doubler at Work. He holds a BBA in computer systems management from Schiller International University.

Ingmar Tönnby has worked with research and development projects since joining Ericsson in 1974. He currently works as a technical expert within the strategic system planning unit at Ericsson Utvecklings AB. He graduated from the Lund Institute of Technology in 1969.

Jan Bergkvist is System Design Manager of Public Intranet, Data Communication Networks and IP Services at Ericsson Infocom Systems. Since joining Ericsson in 1984 he has also worked with system design management of SDH and broadband products. He holds an MSc in applied physics and electrical engineering from Linköping Institute of Technology.

Hans Brandtberg, who joined Ericsson in 1974, is currently a senior advisory specialist in the field of display and reconnaissance systems at the Displays and Reconnaissance Systems Division of Ericsson Saab Avionics AB. Previously, he was manager of the Airborne Display Systems Design Group and of the Systems Development Section. He holds an MSc in electronic engineering.

Johan Frössling has worked as a member of the Map Data Handling Group at Ericsson Radio Systems AB since 1996. In his role as digital map coordinator, he is responsible for procuring and producing digital map data that the organization can use in its cell- and network-planning tools and support systems. He holds an MSc in physical geography from Stockholm University.

Lena Lüning, who joined Ericsson in 1995, currently works with map data handling for mobile network planning at Ericsson Radio Systems AB. She holds an MSc in geology from Stockholm University.

Mats Åkerlund works at Ericsson Telecom AB, where he is responsible for the development and use of geographic information and geographic databases. Since graduating from the Lund Institute of Technology in 1971 he has participated in the development of numerous radar and command and control presentation systems. He also acts as a consultant to the Swedish Department of Defense, and sits on the board of the Center for Geoinformatics at the Royal Institute of Technology in Stockholm.

Mikael Ronström is the senior specialist of distributed databases at Ericsson Utvecklings AB, where he coordinates the technical development of AXE VM and the NDB Cluster. Previously, he has worked with system investigation, research projects, and education and project management. He is currently completing his PhD at the University of Linköping.

Mats-Olov Hedblom has held the position of Environmental Manager at Telefonaktiebolaget LM Ericsson since 1995. Prior to joining Ericsson, he amassed extensive experience of environmental management, business management, research, and engineering while working in the fields of environmental analysis, environmental and chemical technique, bleaching, chemical performance, and product development. He holds a PhD in organic chemistry awarded by Uppsala University.

From the editor

Eric Peterson

Some months back I learned from an acquaintance that a friend of mine was having problems with his computer and hoped I would give him a call. I did, but the line was busy. Twenty minutes later I tried again. The line was still busy. I waited another twenty minutes and tried again. Busy. I persisted. Still busy. Again? Yes, and still busy. "How unusual," I thought. "This fellow isn't really the talkative kind." And then it dawned on me – my friend must have got himself a subscription to the Internet. No doubt he wanted my help configuring his browser and e-mail client. However, judging from the way his phone is constantly tied up these days, it is safe to assume he succeeded on his own or got someone else to help.

This is all well and good – I too spend a fair amount of time surfing in cyberspace – but how would it be if everyone joined in? How would we call or get in touch with one another if our lines were always busy? Through e-mail or Internet-relay chat? By regular post? Fortunately, a group of engineers at Ericsson foresaw this problem and came up with Phone Doubler™.

Phone Doubler is an Internet telephony application whose virtual second line provides temporary simultaneous access to the PSTN and to the Internet. Thus, persons with a telephone can call users on a computer, and users on a computer can call persons with a telephone or other users on a computer. Moreover, by connecting a Phone Doubler server to a PBX and to a local area network that supports remote access, companies can give employees who work from home access to the corporate network environment and to the telephony service of the PBX. Thus, with Phone Doubler, persons who surf the Internet or who connect to the corporate intranet via dial-up access need not worry about missing important calls.

I'm betting a good deal of corporate intranets will be rented from network operators who have implemented Public Intranet. Thanks to its secure, scaleable mechanisms for handling subscription services, charging and quality of service, Public Intranet allows service providers to bring essential business services to the Internet market. And yes, Public Intranet is another Ericsson innovation.

As information and telecommunications network services increase, the role of databases will also take on new proportions. Future databases will offer constant availability, exceptional performance, and real-time response. Once again, Ericsson is at the forefront with a prototype parallel data server for telecommunications applications – the NDB Cluster. This fully scaleable, highly versatile database functions as

- a service-network database – for example, as a home location register, a number-portability database, or a service-control node;
- a management network database – for example, as a charging server;
- an information network database – for example, as a Web server, an e-mail server, or as a data server for a geographic information system.

Geographic information systems are an integral part of the digital-map systems Ericsson designs for advanced vehicles, vessels and aircraft. Ericsson also uses geographic information systems extensively in planning and deploying mobile telephone systems. The application of geographic information systems is probably as varied as the environment. Therefore, it should come as no surprise that they are also used for mapping the environment.

For better or worse, we all have to face up to the environment. We truly do not have a choice. However, Ericsson is doing more than just facing up to the facts – indeed, we are not only a world-leading supplier of equipment for telecommunications systems and related terminals, but we are also the first telecommunications corporation to apply the life-cycle stressor-effects assessment to our design and manufacturing activities.

That is, we are doing our part to reduce our share of environmental load. Moreover, by applying Ericsson products and technical solutions we can further lower the environmental load. For my part, I say we start by implementing Phone Doubler. My friend is home – I can tell because his phone is busy. But since I can't call him, I may have to drive the car over to his place...

Eric Peterson
Editor

Phone Doubler – A step towards integrated Internet and telephone communities

Allan Hansson, Robert Nedjeral and Ingmar Tönnby

Until recently, the Internet and telephony services were thought to be unrelated topics. Today, however, interest in interactive user-to-user applications over the Internet is on the rise, notably for voice applications.

To be really useful for telephony purposes, Internet telephony applications should – without compromising QoS – allow users with computers to call users with telephones, and users with telephones to call users on the computer. Phone Doubler does just this, by creating a virtual second line for the Internet user.

The authors describe the evolution of Internet telephony and Internet techniques for telephony. They then describe four application scenarios. Finally, they describe Phone Doubler, which enables users with dial-up access to the Internet or to a corporate intranet to call and receive calls without having to log off the network. The components that make up Phone Doubler constitute a platform for the support of numerous applications that is a step towards a seamless integration of the Internet and telephony services, including Internet telephony, call centers, and multimedia.

Phone Doubler™

Phone Doubler™ is a trademark of Telefonaktiebolaget LM Ericsson.

The Internet and telephony

Internet telephony and the application of Internet techniques for telephony are hot topics. Interest in these areas is fueled by the dramatic growth of the Internet and the widespread adoption of Internet protocol (IP) techniques for enterprise networks.

The number of Internet users is growing. But so too are available bandwidth and routing capacity. We have seen that IP techniques apply not only to asynchronous data transfer, but also to isochronous services, such as voice and video. Personal computers (PC) and workstations have evolved from off-line-oriented tools for computational tasks and handling documents into full-fledged multimedia devices. They provide also a front end to network services.

Today PC users want nearly continuous access to the network, whereas a few years ago they required only temporary network connections, if at all. From the home, the most common method of accessing the Internet or an enterprise network involves a dial-up service over the public switched telephone network (PSTN). Lengthy sessions (connections to the Internet or to a corporate intranet) over dial-up lines have considerably changed the usage pattern of telephony service providers (TSP). In fact, in terms of duration, Internet sessions more closely resemble television viewing than telephone conversations. Long holding times result in network congestion that could jeopardize other telephone services and reduce call completion rates. And since the telephone line is busy when a dial-up session is active, it has not been possible to use the line for incoming or outgoing calls.

Internet techniques for telephony

Several factors are at work to bring Internet techniques to telephony and to bridge the gap between telephony and IP networks:

- Traffic cost – different tariff principles apply to Internet and telephony traffic; in particular, to long-distance traffic, where telephony traffic is based on distance and usage, whereas Internet traffic is currently based on flat rates or on access time.
- Connectivity – when dial-up access is used often and for lengthy periods, user telephone lines remain busy, which means that – unless the call can be held on the computer – users cannot call or receive calls while they are on-line.
- Integrated applications – multimedia computers are very versatile tools that allow voice applications to be combined with telephony and other services, such as Web surfing, shared applications and multimedia conferencing.
- Integrated networks – in office environments, IP techniques for telephony could greatly simplify the network structure, by transferring all data and telephony traffic

Box A
Abbreviations

API	Application program interface
BRI	Basic rate interface
CDR	Call data record
DECT	Digital enhanced cordless telecommunications
DHCP	Dynamic host configuration protocol
DSP	Digital signal processor
DTMF	Dual tone multifrequency
ETSI	European Telecommunications Standards Institute
FTP	File transfer protocol
GSM	Global system for mobile communication
HTTP	Hypertext transfer protocol
IETF	Internet Engineering Task Force
IN	Intelligent network
IP	Internet protocol
ISDN	Integrated services digital network
ISP	Internet service provider
ITU	International Telecommunication Union
LAN	Local area network
MIB	Management information base
PBX	Private branch exchange
PC	Personal computer
PCM	Pulse code modulation
PoP	Point of presence
POTS	Plain old telephone service
PPP	Point-to-point protocol
PRI	Primary rate interface
PSTN	Public switched telephone network
QoS	Quality of service
RSVP	Resource reservation protocol
SLIP	Serial line interface protocol
SNMP	Simple network management protocol
TCP	Transmission control protocol
TSP	Telephony service provider
UAN	Universal access number
UDP	User datagram protocol
UPT	Universal personal telecommunications
VGA	Video graphics array
VoIP	Voice over IP
WWW	World Wide Web

over a single IP network. Currently, two networks are needed: one for telephony and one for the local-area network (LAN). The application of Internet techniques to telephony involves either supporting telephony on a computer that is connected to an IP network or using an IP network as a carrier of voice information in a telephone call between telephones and computers. Depending on technical and economic considerations, three basic elements are found in various configurations: voice over IP (VoIP), gatekeepers, and voice gateways.

Voice over IP

The Internet protocol is a general protocol for sending packets of digital data between network interfaces that are identified by IP addresses. As such, the protocol is completely transparent to the information transported in the packets. Voice over IP refers to the use of the Internet protocol between applications that handle signals for real-time transmission of voice information over an IP network. At the transmitting end, voice information is encoded into a suitable digital representation, which is divided into packets and sent to the IP address of the recipient. At the receiving end, the information is unpacked and decoded into a voice signal. To reduce the need of bandwidth in the network, compression algorithms are generally used as part of the encoding and decoding.

To ensure quality of service (QoS) for interactive voice traffic, the delay of voice information must be kept low. In interactive sessions, speakers can normally perceive round-trip delays greater than 100 ms, which raises the need for echo cancellation in the network. Delays above 300 ms, which are clearly perceived as a deficiency, require speakers to compensate by waiting for their counterpart to respond.

Some characteristics of IP networks differ from those of isochronous circuit-switched telephone networks. These differences must be considered when IP networks bear interactive voice applications. In isochronous telephone networks, every byte of information is transferred with short and non-varying delay, which means that voice information is transmitted with low latency and low jitter. Transmission errors generally cause only minor degradation of voice quality. Although delay is seldom a problem, it limits the number of satellite hops that can cumulate for a telephone call.

When voice is transmitted over an IP network, every packet contains a time slice (for example, 40 ms) of speech information. To reproduce the voice information, the next sequential packet must be on hand for the receiving application just as soon as it has finished playing back the preceding packet. Otherwise, the voice sounds choppy. Jitter between individual voice samples is not an issue, since the samples in each packet arrive at the same time. Today's IP networks do not guarantee that every packet will arrive, nor do they guarantee that packets will arrive in the order in which they were sent. Some packets may be delayed, depending on router load and on the number of links in the network. Still, in networks where links are few and the load on packet routers is light, delays are short and packet loss is low. For voice over IP, there is little chance that packets will arrive in an order that differs from the order in which they were sent.

Although flawless transmission of IP packets cannot be guaranteed, transport protocols – such as the transmission control protocol (TCP) – guarantee fault-free transmission of data over the network. If necessary, these protocols can request that packets be retransmitted. However for voice over IP, the retransmission of packets would cause disproportionate delay. Therefore, voice over IP makes predominant use of the user datagram protocol (UDP). To give controlled characteristics to voice over IP, the resource reservation protocol (RSVP) is being considered more and more, although it is not yet commonplace.

The delay of voice packets in IP networks is caused by three factors: the length of the time slice, link delays in the network, and router delays. Voice-application designers must consider these factors. To be used broadly, voice applications must be designed to work well over dial-up connections.

Time slice. Before a packet can be sent, voice samples must be collected for the entire time slice. The length of the time slice should balance the need of short delays with the need of limiting bandwidth and router load. The shorter the time slice the greater the bandwidth that is required, since more overhead must be devoted to the packet headers, and routers must handle more packets per time unit. In most cases, the time slice may be 40 ms or less, which is the equivalent of transmitting GSM-encoded speech at a rate of 15 kbit/s. To transmit the same speech information over a 14.4 kbit/s

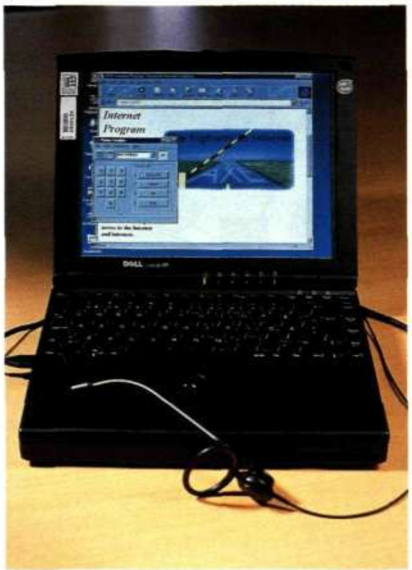


Figure 1
Phone Doubler enables users who surf the Intenet to make or receive telephone calls from their PC.

modem line, the time slice must be extended to 80 ms.

Link delays in the network. Link delays over the network mainly depend on the link with the least bandwidth which, in a dial-up scenario, is usually the dial-up link. The transmission time of voice packets themselves, which are rather short, is not very significant (often less than 10 ms). However, if the voice packets are mixed with traffic that uses long packets, then the voice packets may be delayed significantly by the transmission time of the longer packets. This problem can be alleviated if voice packets are assigned a higher priority over the link. For slow links, segmentation mechanisms might also be necessary, allowing the voice packets to interrupt the transmission of a long packet.

Router delays. Each router that a packet passes while *en route* from sender to recipient delays the packet. How much a packet is delayed depends on the momentary load of the router. Router delay adds up for each router involved. Thus the total delay may vary significantly, from being negligible in fast, low-load networks or networks that employ resource reservation mechanisms, to very significant, as is common today in parts of the Internet. Router delay is the most unpredictable aspect of delay in voice applications.

The gatekeeper

Voice over IP is a technique that enables voice information to be transmitted between interfaces that are identified by IP addresses. By contrast, telephony and other interactive applications deal with communication between users. Basic gatekeeper functionality includes mapping a user identity, such as a user name or telephone number, to the IP address of the interface where the user may be reached, and to the characteristics of the communication with the user at that interface.

The mapping between users and IP addresses is not always static. Dial-up user IP addresses are usually assigned dynamically when access to the Internet service provider (ISP) is established. Similarly, even in address domains with fixed IP-address allocation, users commonly move between computers. Thus, gatekeepers must contain some mechanism for handling dynamic information, by registering current user characteristics.

The voice gateway

The earliest Internet telephony applications solely allowed voice conversation to be held between users on computers connected to the IP network. However, to be really useful for telephony purposes, users with computers should also be able to call users with telephones, and users with telephones should be able to call users on the computer. The role of the voice gateway is to bridge the gap between a telephone network (the PSTN or integrated services digital network, ISDN) and an IP network, by converting address information (assisted by the gatekeeper) and handling call-setup procedures between the networks. When a call is in progress, the voice gateway converts the voice signal between an analog or digital line to the telephone network and voice packets sent over the IP network.

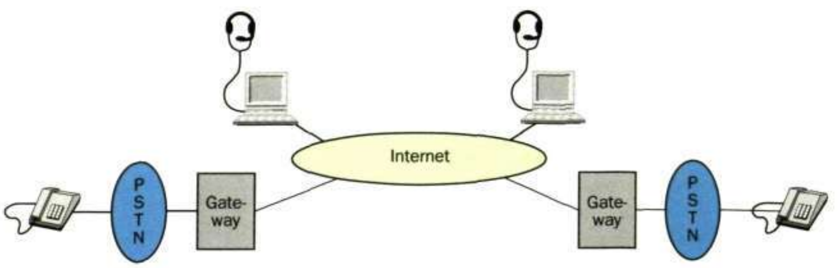
Application scenarios

The following four scenarios emphasize particular driving forces for Internet telephony and the application of Internet techniques for telephony. Real applications and products combine elements from several of these scenarios.

Internet telephony scenario

The most discussed phenomenon regarding telephone and IP networks, notably the Internet, is Internet telephony, where the Internet is used as a long-distance carrier of voice for telephony. Originally introduced between Internet users (many applications exist, some of which have been integrated into Web browsers), Internet telephony has evolved to allow calls from IP users to telephone users, and between telephone users via the Internet. In this scenario, one or two voice gateways are involved in a call. The gateways are located relatively close to the telephone users. Using the IP network for long-distance traffic means that only local calls are made (and billed) over the telephone network (Figure 2).

Figure 2
Internet telephony uses Internet as a long-distance carrier of telephone calls between telephones and computers connected to the Internet.



Interest in Internet telephony is tied to the issue of charging, where different tariff principles apply to IP traffic and long-distance telephony traffic. Although telephony traffic over the Internet requires considerably less bandwidth than a call over the telephone network (actually, this could be debated, since voice compression is often used for telephony traffic on transatlantic lines) there is no strong evidence that the real "production cost" of services for handling traffic over the Internet is less. As tariff principles become distance-independent, this cost relationship will change. In a telephone network where all calls cost the same amount, regardless of distance, there is no cost benefit for using a dial-up Internet access to make long-distance calls over the Internet.

There are two issues to Internet telephony that need to be considered:

- The quality of service associated with an Internet phone call may easily be inferior to that of a call made over the telephone network. In particular, delay can be considerable as well as unpredictable – the result of traffic congestion and of hops between many routers. In well-structured IP networks, the problem of delay can be mastered by controlling load and by limiting the number of router hops.
- For service to be useful for making calls to many destinations from many places, service providers must be compensated in some way for the use of their gateways – which must be placed all over the world.

Intranet with telephony scenario

Companies with a good intranet to which all or a large majority of employee computers are connected might consider a scenario in which the intranet also supports telephony. This would do away with the need of having to manage two networks. Instead, telephony and data services could be provided over a single network. Intranets can be distributed between geographically disparate sites via leased lines or other means that guarantee adequate quality (Figure 3).

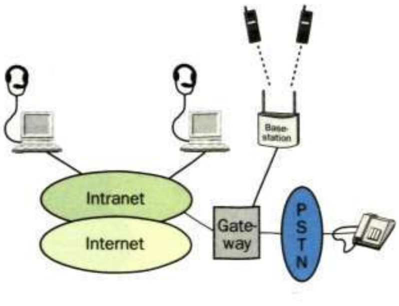


Figure 3
IP-based enterprise network used for data and telephone traffic. A gateway is used for reaching users with hand-held telephones, and for connecting to the public telephone network.

According to this scenario, there are no traditional telephones within the company. The computers on the intranet contain telephony applications and devices that are suitable for providing voice conversation alongside data applications. Stand-alone IP-telephone devices may also be used as well as wireless telephones. Traffic of all kinds is transported over the IP network. Since the network is controlled by one organization, the quality of network services can be predicted and maintained at a level that is sufficient for telephony traffic – even for long-distance calls between different sites. For interactive applications, user names are the obvious choice of address. Directory services keep track of where users are currently logged on, thereby facilitating personal mobility within the network.

Obviously, employees must still maintain telephone contact with users outside the network. Thus users must also have telephone numbers at which they can be reached, and the network must be connected to the telephone network by at least one gateway. The gateway, which may be connected to the telephone network over an ISDN primary-rate interface (PRI) or a private branch exchange (PBX) interface, maintains or calls a directory that matches telephone numbers with users who can be reached on the IP network. The gateway is also used for interconnecting between the IP-based network and a wireless network; for example, a digital enhanced cordless telecommunications (DECT) network. An Ericsson product, LAN-phone, addresses this scenario and supports user mobility between the public network and the LAN by using an intelligent network-based (IN) universal personal telecommunications (UPT) service.

Telephony access over IP

The third scenario represents Internet users' need to stay in touch with telephone users, by being able to call them and to be called. This scenario is particularly interesting to users who use the telephone line for a dial-up connection to an Internet service provider, thereby tying up the telephone line for long periods. The application of Internet techniques can provide a virtual second line that enables users to place and receive telephone calls without having to log off the network (Internet). Phone Doubler allows users to call and to be called from any telephone in the world without introducing a new telephone number or unusual call procedures.

In this scenario, the quality of communication and connectivity are more important than price. Thus the voice gateway is best placed near the PC user where, to the greatest possible extent, voice information travels through the telephone network. The IP connection is used for multiplexing telephony and data traffic.

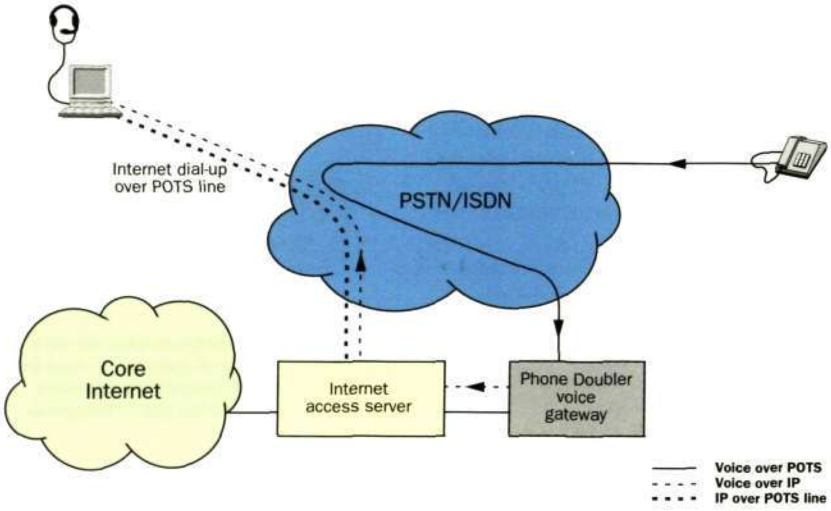


Figure 4
An incoming telephone call to a Phone Doubler user is redirected to the Phone Doubler voice gateway, and forwarded to the user via the dial-up Internet access.

Multimedia telephony scenario

Multimedia telephony depicts the use of voice calls as a complement to interactive multimedia sessions. The scenario encompasses a wide range of applications, including:

- call-center applications, where users navigate through the Web – for example, to the home page of a travel agency. When users want to contact an agent, they simply click on a symbol on the screen to establish a voice call through the Internet or over the telephone network;
- application-sharing, which allows users to interactively share and manipulate documents or images as they carry on a voice conversation;
- multimedia conferencing, where users of computers and telephones interact using the capabilities of their respective terminals.

According to this scenario, the main emphasis is on using IP techniques to bring various media into play on the user's workstation, thus exploiting exciting new ways of communicating. Because quality aspects are very important, telephone networks may be used, if necessary, to ensure the quality of the voice transmission.

In the future, the integration of applications is expected to become the most important driving force for IP techniques.

Interoperability and standardization

A crucial factor for successfully merging the worlds of telephones and computers is the assurance of interoperability between components. Communication must be established seamlessly between users of different devices, on different networks, and using different applications. Much standardization work has already been accomplished in this area. Nevertheless much work remains. As a major player in communications, Ericsson is heavily involved in standardization issues within the frameworks of the International Telecommunication Union (ITU), the European Telecommunications Standards Institute (ETSI), and the Internet Engineering Task Force (IETF).

Phone Doubler

Today the most common way of accessing the Internet is via a modem link over a dial-up PSTN connection. Unfortunately for the user, telephone calls and Internet access via the PSTN access line are exclusive of one another. That is, when someone surfs the Internet, the PSTN access is not available. Similarly, when the PSTN access is in use, the Internet cannot be accessed.

In the past, the solution to this problem was to get a second telephony subscription. Now, however, Phone Doubler offers an alternative: Phone Doubler creates a virtual second line, which allows users to access telephony services of the PSTN while they access the Internet via a modem or ISDN. During an Internet session, users may call any telephone in the world, or they may be called from anyone who dials their regular telephone number. The other party need not know that the user is connected to the Internet.

What is important in this situation is that users are able to stay in contact with the rest of the world without the worry of missing important calls. In this case, the price of telephone calls is not the issue.

How does Phone Doubler work?

Users experience Phone Doubler through the Phone Doubler Client, a software telephone application that runs on personal computers. If a user wants to make a call while connected to the Internet, he or she simply clicks on an icon on the screen, and then enters the telephone number to be called in the dialog box that appears. The client forwards the call request to the Phone Doubler voice gateway, which translates the request into an ordinary telephone call to the PSTN/ISDN (Figure 4). The PSTN/ISDN completes the call to its destination. When the called party answers, conversation is held by means of a microphone and loudspeaker (or a headset), which are connected to the PC.

If a telephone call is made to the user while he or she is surfing the Internet, a window pops up on the screen and an acoustic signal alerts him/her to the incoming call. To establish the call, the user clicks on the answer button.

Gateway near user

The Phone Doubler voice gateway is normally placed adjacent to the access server, in order to provide optimum voice quality; that is, by avoiding delays that are caused by routed networks. The gateway exchanges IP packets with the Phone Doubler client for controlling the calls and for transferring voice over IP. On the network side, the gateway connects to ISDN via a standard primary-rate interface or a basic-rate interface (BRI). The capacity of the gateway is adapted to match the capacity of the access server. Given normal telephony call frequency, a voice gateway supporting 30 simultaneous voice calls is suitable for supporting 200 to 300 simultaneous IP users.

Use of voice over IP

With Phone Doubler, voice traffic is only carried over IP between the user PC and the gateway. Since this takes place over a dial-up connection through the PSTN/ISDN, which is not shared with other IP users, the quality of the voice-over-IP connection is directly dependent on the user's own traffic. Preferably, the access server is configured to give higher priority to voice packets than to other packets. The absence of extensive data traffic during a voice call keeps voice quality and delay to well within acceptable limits. Except for access to the ISP, every part of the call is handled as a normal telephone call. Thus, the quality of the call is not affected by whether it is local or long-distance.

With voice over IP, voice information can be transmitted over the dial-up modem link together with other types of information. However, since the available link is a dial-up modem connection at a rate below 64 kbit/s, speech information must be compressed. Standard GSM compression (13 kbit/s) is sufficient for carrying speech information over a 28.8 kbit/s modem link. Other speech compression algorithms may also be used.

Phone Doubler enables users who access the Internet over ISDN to send and receive IP traffic at the full 128 kbit/s rate while making and receiving telephone calls. Only about 15 kbit/s of bandwidth is used for voice over IP. All remaining capacity is allocated to data traffic. Therefore, data traffic does not have to be downgraded to a 64 kbit/s connection for incoming calls.

Binding and mapping addresses

Telephony and IP networks have completely different address spaces. In order to relate telephony calls and IP packets to each other, the addresses must be mapped and bound to a user.

When the user starts the Phone Doubler client, the current IP address (fixed or dynamic assignment) is passed to the gatekeeper along with the user's telephone number and other necessary authentication data. The gatekeeper authenticates the user and registers the binding between the telephone number and the IP address.

Each time a call passes through the voice gateway, the gatekeeper is consulted about the binding. This information is used for charging outgoing calls. For incoming calls, the information is used to find the IP address at which the user may be reached.

Redirection

To reach a subscriber who uses the telephone line for Internet access, calls need to be redirected to the telephony gateway. Once a call to the user's telephone number is delivered to the gateway, the gateway can forward it to the proper IP address.

Several methods may be employed for redirecting calls. The preferred method is to use the standard PSTN call-forwarding-on-busy service. The PSTN detects that the subscriber line is busy and redirects the call to the voice gateway, which consults the gatekeeper to determine whether or not the subscriber is currently an active Phone Doubler user. If so, the call is routed by the gateway to the Phone Doubler client. Otherwise, the voice gateway signals a busy condition back to the calling party. Another means of redirection is remotely controlled unconditional call forward, requested by the voice gateway.

Applications

Virtual second line

The most straightforward application of Phone Doubler is to provide temporary simultaneous access to the PSTN and to the Internet by creating a virtual second line. The primary benefit of doing so is that users are not excluded from PSTN access when they are connected to the Internet. For example, users who are surfing the Internet may be reached by incoming calls, or they may initiate outgoing calls without needing a second phone line.

By implementing Phone Doubler, Internet service providers can offer their customers better, more competitive service.

Telephone service providers also have much to gain from Phone Doubler: calls to persons surfing the Internet will reach their destination and be answered. Thus, more calls are completed and charged. The alternative to Phone Doubler is often a second PSTN subscription. Phone Doubler helps operators to avoid costly investments in the access network and local exchanges, which means providers can increase revenue without making significant new investments. A telephone service provider who doubles as an ISP may also offer competitive services that combine telephony with the Internet.

Working at home

Companies having a Phone Doubler server connected to their PBX and to LANs that provide remote access service can give their employees much greater flexibility in selecting their workplace. For example, employees working from home gain access, over a dial-up modem, to the ordinary network environment and to the telephony service of the PBX. Incoming and outgoing calls can be routed to the internal number, provided that the user's incoming calls are redirected to the voice gateway. Calls to the user's home telephone may also be redirected to Phone Doubler via the company PBX. Thus, while they use dial-up access, employees can be reached at their office and private (home) telephone numbers.

Web dialing

Objects on a Web page may contain a link to a telephone number. With Phone Doubler, users can call such numbers simply by clicking on the link. A straightforward application of this functionality would allow an Internet user who has run across an interesting commercial offering on a Web site to click on the link on a Web page to call for more information.

The Helsinki Telephony Company is engaged in a project that will create a three-dimensional (3-D) image of Helsinki: the Helsinki Arena 2000 virtual city. When the project is complete, Internet users will be able to move about in a 3-D model of the Finnish capital. A prototype of the model demonstrates how Phone Doubler allows users to click on telephone symbols on the doors and walls of the virtual city, in order to initiate a telephone call. Thus, while strolling along the streets of the virtual city, users who want the services of a certain shop may click on the phone symbol associated with that shop, calling and speaking with someone at the real store. Similarly, if while passing by the home of a friend a user gets the urge to call, he or she may do so.

Call centers

Companies that use telephony extensively as a medium for interacting with their customers often establish call centers, which provide computer support during telephony conversations. Call-center operators get information on individual customers before or at the beginning of the call. Customers who call the center are frequently prompted by an automated attendant to respond, with the push-buttons of their telephone, to a number of conditions or requests. The information is then made available to the operator who takes over the call.

Web techniques and the functionality of Phone Doubler could greatly enhance the value and productivity of call centers. For example, companies that present their products on the World Wide Web (WWW) could ask customers to complete a data form via the Web interface. Users would be asked to give their telephone number and perhaps a reference that is used to look up addressing information and customer preferences. The call center could then dial the customer back as soon as the next agent became available, displaying information on the customer to the agent. If the customer has Phone Doubler, the call can be completed while he or she is connected to the Internet via dial-up access. Moreover, the dialog between the customer and the agent need not be limited to voice, since the connection over the Web may also be used for displaying images or for demonstrating applications. Thus we see that the combination of Phone Doubler and Web technology creates powerful multimedia solutions.

Mobility

Users of a portable computer with the Phone Doubler client and a built-in modem can connect to the Internet over any available telephony line, register with their ISP point of presence (PoP), and then make outgoing calls, which are charged to their account. If remote redirection is supported, incoming calls to their ordinary telephone number can be directed to the telephony gateway, which forwards the calls to the user's computer.

Phone Doubler – the product

There are two versions of Phone Doubler: one for Internet service providers and telephony service providers, and one for enterprise markets. The former was developed for service providers who want to add value to their basic Internet access service. The enterprise version, called Phone Doubler at Work, addresses the needs of corporate environments, where employees use dial-up links for remotely accessing the corporate intranet. Each version offers the same basic functionality for enabling telephone calls and dial-up Internet/intranet access to share a single plain old telephone service (POTS) line. The main difference is in the way the voice gateways process voice. In the high-capacity version, the voice gateway relies on digital signal processors (DSP) for voice encoding/decoding; in Phone Doubler at work, voice encoding/decoding is performed by software.

Voice gateway

The voice gateway, which comprises hardware and software, receives and processes telephone calls and monitors call progress. In essence, it is a transmission device that converts voice to and from G.711 pulse code modulation (PCM) and GSM signals, and terminates ISDN signaling.

The software component, which runs on a Windows NT server,

- receives and processes telephone calls from external parties to Phone Doubler clients;
- receives and processes telephone calls from Phone Doubler clients to external parties;
- monitors call progress and signals connection failures;
- compresses and packs PCM signals from the external network into GSM 6.10 and UDP coding;
- decompresses and unpacks UDP and GSM 6.10 coding into PCM signals for transmission to the external network.

Being scaleable, the voice gateway can easily be expanded if additional voice channels are needed. ISPs and network operators increase the capacity of their systems by installing additional DSP boards. In the enterprise version, capacity is increased by upgrading the voice gateway CPU.

Box B
Product data

Functional
Voice over IP
- GSM 6.10 speech coding (13 kbit/s)
- Advanced echo suppression
- DTMF signaling

Security
- Net mask user validation
- Challenge-response user authentication (RSA MD5-based)
- E.164 address validation
- Gatekeeper design prevents users from downloading other users' passwords from the FTP server
- IP source address restrictions for HTTP and FTP access

Management
- SNMP-based alarm handling and element management, using the MIB-II standard
- HTTP-based subscriber management, using a Web browser
- FTP-based remote software updates and upgrades

Client computer
Operating systems supported
- Microsoft Windows 95
- Microsoft Windows 3.x with TCP/IP stack and SLIP/PPP
- Microsoft Windows NT

Hardware requirements
- RAM: 8/16 MB (Windows 3.11/95, NT)
- CPU: 486DX2-66 or better
- Speakers and microphone
- 2 MB of free hard disk space
- Duplex or simplex audio card (MPC-compliant wave device)
- V.32bis/V.34 modem (14.4 kbit/s or faster)

Gateway hardware
ISP/TSP Phone Doubler
ETSI Gateway
- Industrial-grade, PC-based server

Phone Doubler at work, high-end PC
- Minimum 133 MHz Pentium CPU
- 64 MB RAM
- 1 GB SCSI-II hard disk
- Eicon PRI/BRI interface board (Diva Server BRI, Quadro, S2M)
- 10BaseT, 10Base5 or 10Base2 Ethernet interface
- VGA adapter

Gateway software requirements
Windows NT 4.0 (workstation or server) updated with the latest service pack
Voice gateway features
- E.164/IP address dynamic translation
- Support of dynamic IP address allocation (PPP, DHCP)
- SNMP (MIB-II) management
- Terminal mobility support

Data network
- 10BaseT, 10Base5 or 10Base2 Ethernet interface (access server side)

ISDN network
- ISDN PRI or BRI (network operator side)

Gatekeeper

The gatekeeper handles administrative tasks for several gateways in the Phone Doubler system. Some of its most important functions include

- authenticating users;
- managing services;
- managing subscriber databases;
- locating subscribers;
- collecting charging information;
- managing gateways;
- assisting in call setup;
- providing call data (which is needed for billing).

In its most basic configuration, Phone Doubler needs just one gatekeeper (more may be added for redundancy).

Phone Doubler client

The Phone Doubler client is a software application that runs on the user's computer. It enables users to access the service using a multimedia computer (Box B) and an ordinary 14.4 kbit/s (or faster) modem.

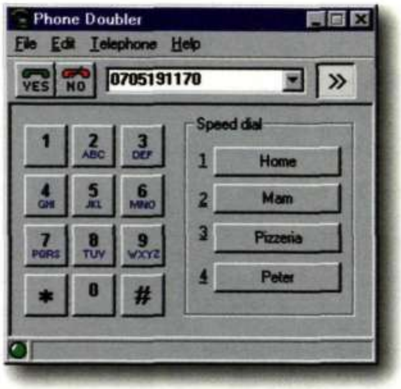


Figure 5
The Phone Doubler client permits users to make, answer and reject calls. The client window pops up on the screen when an incoming call is on the line.

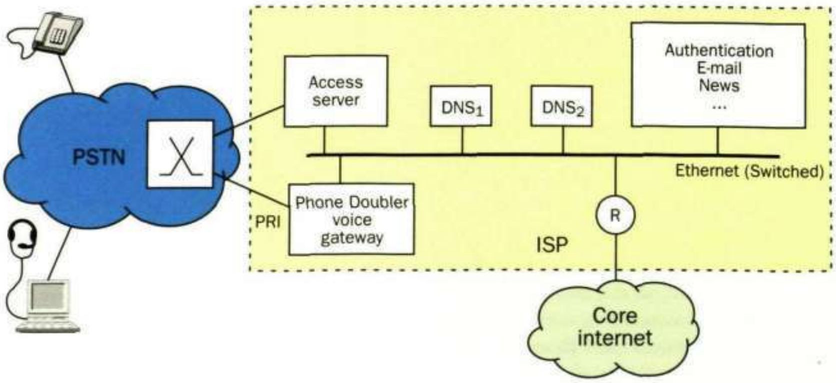


Figure 6
Phone Doubler voice gateway used with a typical Internet PoP configuration.

In terms of appearance, the client application closely mimics the look and feel of a standard telephone set (telephone keypad, programmable speed-dial buttons, and hands-free speakerphone mode), rendering it easy to use for making and receiving telephone calls (Figure 5). The Phone Doubler client

- provides the user interface to the Phone Doubler service;
- handles call-control signaling to the voice gateway;
- transfers the user E.164 and IP addresses to the gatekeeper;
- decompresses and unpacks GSM 6.10 and UDP coding for output to the computer sound card;
- compresses input from the computer sound card to GSM 6.10 and UDP coding.

Application program interface

Phone Doubler telephony functionality may be accessed using the Microsoft telephony application program interface (API), which enables Phone Doubler to be used with Web browsers and other applications. For example, the action of clicking on a link to a telephone number activates Phone Doubler, which initiates the call.

Network scenarios

Network introduction. Although the physical layout of a Phone Doubler system may vary in ISP and network operator configurations, the basic connectivity requirements are the same. Viewed as a whole, the Phone Doubler system needs two external links: one to the ISDN/PSTN telephone network, and one to the data network. The link to the ISDN/PSTN network may be an ISDN PRI or one or more BRIs, depending on requirements for capacity. The link to the data network constitutes a direct network connection to the Internet backbone or to an intranet.

Large-area coverage. If the users of a Phone Doubler system are dispersed over a large geographic area, several voice gateways may be deployed, each one serving a particular area and linked to the local data network access point. Phone Doubler does not route compressed voice data over the Internet, intranets, or over any other IP-based network. Instead, its voice gateways place a Phone Doubler call directly on the ISDN/PSTN telephone network. This limits the voice-over-IP segment of the call to between the user and the local data network access point and maintains sound quality – there are no adverse effects such as delay, jitter, lost packets, and echo, which are present in wide-area, IP-based networks. Every voice gateway is reached by one universal access number (UAN) to which users forward their calls when they use the Phone Doubler system. The routing algorithm for the UAN should be equivalent to the point-of-presence UAN algorithm. No special requirements apply to the location of the gatekeeper, since it is not involved in the transmission of voice information.

In a typical PoP configuration, the voice gateway and the gatekeeper are connected directly to the Internet service provider's local-area network. If there are multiple points of presence, one or more separate voice gateways are installed at each location – to avoid having to route voice data. Each voice gateway in a wide-area system is managed by a centralized gatekeeper (Figure 6).

In AXE systems, an AXE Internet access server is integrated into the switch. The voice gateway has a direct network connection (for example, Ethernet) to the access server; that is, it is connected directly to the AXE cabinet.

Management

All management functions are grouped into a common Phone Doubler site-management application that runs on the gatekeeper. There are two basic function categories: user administration and network management. The management application, which is based on the WWW metaphor, may be accessed using any standard Web browser. This gives system administrators a familiar environment in which navigation is fast and intuitive.

The user administration subsystem generates a list of all Phone Doubler users as well as a list of users who are currently signed onto the system. In addition, it enables administrators to add, delete and modify user data.

The network management subsystem is used for adding, deleting, modifying, and configuring voice gateway modules. It also produces a list of installed voice gateways (Figure 7).

Charging

Charging is handled in a variety of ways, depending on customer requirements. The Phone Doubler system is not limited to, but supports, common billing mechanisms, such as Network-based charging and Radius accounting. In fact, call data generated by the gatekeeper may form the basis of virtually any customized billing scheme.

- In *built-in mode*, the gatekeeper generates call-data records (CDR) of the local file system. Incoming and outgoing calls are recorded. Using the file transfer protocol (FTP), the billing system collects CDRs at appropriate times.
- In *network-based mode*, the local exchange handles charging. For outgoing calls, the gateway sends the user-provided caller ID to the exchange. This ID must be included in the call-data record from the exchange.
- In *Radius accounting mode*, charging data is converted for accounting purposes into Radius messages. Radius is the standard used for charging dial-up sessions.

Software updates

New software releases may be downloaded from a dedicated FTP server. Thus software may be upgraded automatically after the distribution directory on the FTP server has been updated.

Conclusion

Phone Doubler creates a virtual second line to the PSTN for Internet users. Thus, users who surf the Internet or who connect to a corporate intranet via dial-up access need not worry about missing important calls. Instead, with Phone Doubler, they can make and receive calls without logging off the network.

By clicking on an object on a Web page that is linked to a telephone number, users with Phone Doubler can initiate a call to that number. Phone Doubler's strengths are not limited to voice, however. Indeed, the combination of Phone Doubler and Web technology also creates powerful multimedia solutions. For example, as agents converse with their customers via Phone Doubler, they can also display images and demonstrate applications over the Web.

There are two versions of Phone Doubler: one for Internet and telephony service providers, and one for enterprise markets.

Companies that implement Phone Doubler give their employees much greater flexibility in selecting their workplace. Employees who work from home gain access to the corporate network and to the telephony service of the PBX. Thus, they can be reached at their internal company number as well as at home.

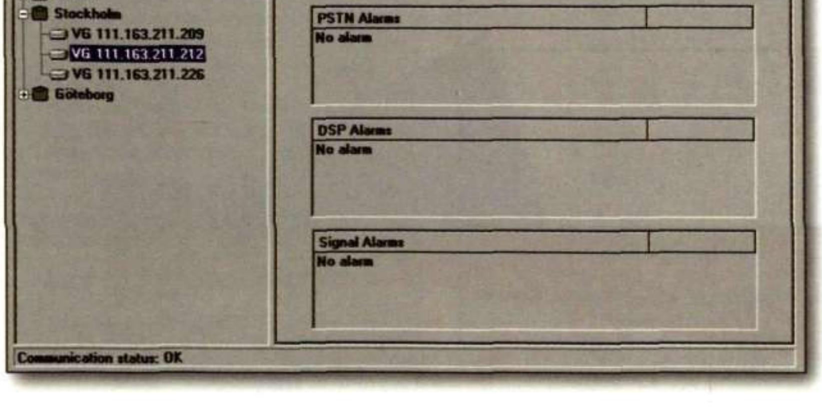
In short, everyone benefits from Phone Doubler:

- Users gain simultaneous access to Internet and PSTN/ISDN services.
- Internet and telephony service providers can offer better service (improved customer satisfaction).
- Telephony service providers can increase revenues (greater call – completion percentage) without making significant new investments in the access.

The present versions of Phone Doubler are only the first of many solutions in Ericsson's growing portfolio of products that facilitate seamless cooperation between telephone and IP-based networks.

The components of Phone Doubler, which support such standards as H.323 and a variety of speech algorithms, constitute a platform that is evolving to support numerous applications, including Internet telephony, call centers, and multimedia.

Figure 7
Phone Doubler voice gateways are managed from a standard Web browser.



VoiceGateway 111.163.211.212

PSTN Alarms
No alarm

DSP Alarms
No alarm

Signal Alarms
No alarm

Communication status: OK

Public Intranet

Jan Bergkvist

Secure, scaleable mechanisms for handling subscription services, charging, and QoS are needed for introducing essential business services into the rapidly evolving Internet market. Public Intranet, which addresses these needs with a component library and an open-standards distributed architecture, gives service providers the tools they need in order to supply top-grade services in large-scale solutions.

The author describes Public Intranet, its system architecture, quality of service and resource management, and some of the services that may be offered through Public Intranet, including a content hotel, a virtual private intranet, Internet access, and multimedia telephony.

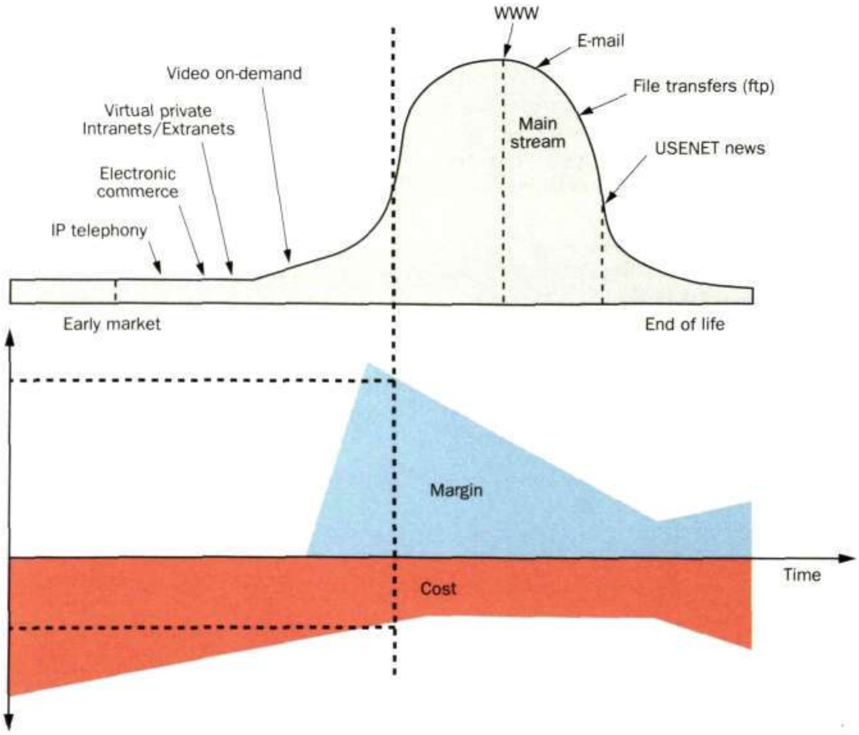
As a pervasive platform for large-scale distributed applications, the World Wide Web (WWW or Web) currently lacks security, subscription services with charging mechanisms, and service-resource management. Nonetheless, as greater functionality and improved performance are being added, Web technology is moving beyond its current role as a document library of static hypertext markup language (HTML) files hosted by Web servers.

In many respects, the Internet may be regarded as just another network that adds new services to the total offering of communication services. But in terms of how technology has been applied to date, substantial improvement is needed for bringing the quality of the Internet to the level of other existing networks. The Public Intranet service network (PISN) architecture introduces characteristics most commonly associated with telecommunications into present-day Web technology.

The Public Intranet defines a business concept that provides a framework for developing and deploying commercial services that use the Internet protocol (IP) as a transport mechanism. The framework includes services for applications (content hotels and multimedia telephony) and for subscription, charging, security, and resource management. An IP-network solution is included in the total offering.

The service network, which may be adapted to meet the needs of different markets, addresses service providers who target business and residential markets. Special versions for the enterprise market may be configured to offer Intranet-in-a-box functionality.

Figure 1
Profit margins for provisioning services diminish once the services become a commodity. Because the cost of provisioning services is usually fairly constant, service providers need to move up the value chain in order to stay profitable.



Service-provider strategies – an evolving market

Today's service providers mainly focus on supplying access to the Internet and associated services, such as e-mail, file transfer, and news. As competition between providers intensifies and profit margins diminish, these services will become commodities. Therefore, for service providers to make a profit, they must constantly add new services, leaving old ones to low-margin, low-cost providers (Figure 1). To survive in this high-end segment, service providers must rationalize their operations and become ever more cost effective. Several established providers – and some new ones – will move higher up the value chain to begin offering more valuable services for which end-users are willing to pay a premium.

Useful business services are those that

- end-users work with most often;
- are critical to end-user business operations.

Today, services of this kind are not provided on IP networks, because the networks are neither secure nor guarantee quality-of-service (QoS) at the transport or application level.

The Public Intranet effectively combines new services with techniques for supplying secure, high-quality service in large net-

Box A
Terminology

ADK
Application development kit

ADSL
Asymmetric digital subscriber line.

API
Application program interface.

ATM
Asynchronous transfer mode.

Broker
The broker is a service network function which presents the service offerings for a given end-user based on factors such as subscriptions, user preferences, service availability, load on servers, and the capability of terminals.

Broker server
A process or a physical machine that executes a broker function in the Public Intranet.

Content
Any information that is useful for an end-user and that can be retrieved to the end-user's terminal equipment, such as AV streams, text and graphics files, catalogs, Web page contents, Web search results, electronic papers and documents, software files, or mobile code.

Content hotel
An end-user service that provides support functions for distributing and retrieving content information. End-users may access the content hotel to search, browse, and retrieve data.

CORBA
Common object request broker architecture.

CPN
Customer premises network.

DSL
Digital subscriber line.

GUI
Graphical user interface.

HTML
Hypertext markup language.

HTTP
Hypertext transfer protocol.

ICMP
Internet control message protocol.

IDL
Interface definition language.

IIOP
Internet interORB protocol.

IP
Internet protocol.

IPsec
IP security, Security IP-tunneling protocol.

ISDN
Integrated services digital network.

ITU-T E.164
ITU-T recommendation on the numbering plan for the ISDN era.

ITU-T H.320
ITU-T recommendation on narrowband visual telephone systems and terminal equipment.

ITU-T H.323
ITU-T recommendation on visual telephone systems and equipment for local area networks that provide a non-guaranteed quality of service.

ITU-T T.120
ITU-T recommendation on data protocols for multimedia conferencing.

ITU-T X.500
ITU-T recommendation on information technology – open systems interconnection – directory: overview of concepts, models, and services.

ITU-T X.509
ITU-T recommendation on information technology – open systems interconnection – directory: authentication framework.

LDAP
Lightweight directory access protocol.

MCU
Multipoint control unit.

MMTS
Multimedia telephony system.

O&M
Operation and maintenance.

OMG
Object management group.

PISN
Public Intranet service network.

PISP
Public Intranet service platform.

PKCS
Public key cryptography standard

PKI
Public key infrastructure.

PSTN
Public switched telephone network.

QoS
Quality of service.

RFC1577
Request for comments no. 1577 – proposed Internet standard on classical IP and ARP over ATM.

Service provider
A service provider uses the mechanisms of the service network platform for deploying, publishing, and marketing a set of applications or services.

SNA
Service network application. Is an application or service deployed in the service network.

SNMP
Simple network management protocol.

SOHO
Small office/home office.

SSL
Secure sockets layer.

Subscriber
An organization or person that has entered into a contractual relationship with the Public Intranet on behalf of a set of end-users. Through this contractual relationship, the subscriber requests access to a set of end-user services, for which the subscriber is willing to pay according to the pricing model defined for these services.

Subscription
A directed contractual agreement with obligations of delivering and paying inherent in the direction of the agreement.

Terminal
Any equipment, such as a personal computer or a workstation, that is connected to the Public Intranet through a line interface.

UPT
Universal personal telecommunications.

VPI
Virtual private intranet.

WWW
World Wide Web, also known as the Web.

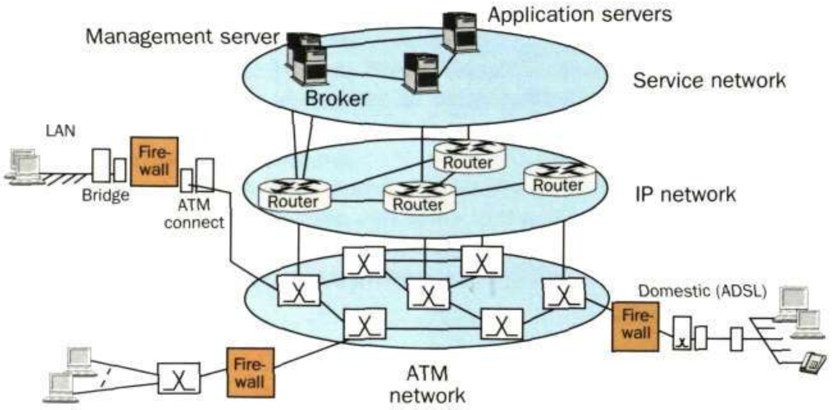


Figure 2
The Public Intranet is a service platform built on top of an IP network that interconnects customer networks (LANs) and various service nodes.
The Public Intranet is independent of the underlying transport network. IP networks that are built on top of ATM can enhance resource handling.

works, thereby giving service providers a foundation on which to develop their businesses. Furthermore, service providers can reduce their operation and maintenance (O&M) costs by integrating the Public Intranet into existing external administrative systems, such as billing and customer account systems.

Service network architecture

The complete Public Intranet includes an IP network and a service network. Each IP network installation is custom-designed, taking into consideration the existing network structure and installed equipment. The Public Intranet offers design guidelines for different solutions. The first installation is based on classical IP over ATM according to RFC1577.

User equipment in customer premises networks is typically connected to IP routers via Ethernet interfaces. User terminals may also be connected to the network through various digital subscriber line (DSL) access networks, which open the way for small office/home office (SOHO) solutions.

The service network (Figure 2) is distributed over several logical and physical nodes, including:

- a management server – which includes a user database and interfaces to external systems;
- broker servers – which authenticate end-users and present the service offering;
- application servers – which run the applications selected by end-users.

The nodes and the customer premises networks (CPN) are interconnected by the IP network.

Before the broker can present services, the service provider must register them with the service network operator. End-users are granted access to services through subscriptions.

End-user terminals connect to the service network via any available broker in the network.

Public Intranet offers an integrated management of network elements; that is, it integrates the presentation of different network element management systems into a single interface. The network-management system also includes a common alarm system that displays the alarms of different network elements.

To make a Public Intranet installation secure, conceivable threats must be modeled and owner and customer needs analyzed. The Public Intranet contains tools for implementing security policies.

Firewalls and packet-filter functions ensure that access to individual machines may be controlled for each protocol; that is, access to network management services, such as telnet and the simple network management protocol (SNMP), may be confined to a subset of management terminals or IP addresses. Packet filters, which are provided by the IP routers, filter traffic that enters and leaves different administrative domains. Service network providers own the packet filters; therefore, they can use the filters to control and protect network traffic within the Public Intranet by screening the Internet control message protocol (ICMP), by screening other protocols, and by validating source addresses.

Firewalls are used when requirements for controlling traffic and logging exceed the capabilities of packet filtering. Requirements of this kind include control functions for Internet gateways and customer premises.

High-capacity Internet gateway and firewall

Internet gateway and firewall functions provide Public Intranet users with high-capacity access to the Internet while protecting the resources of the Public Intranet from unauthorized use or misuse by external users. All external traffic entering the Public Intranet is screened. Depending on customer requirements, outbound traffic may also be screened.

Network planning and dimensioning are important for achieving good quality. Every resource, except the customer premises network may belong to, or be associated with, the Public Intranet service network.

The function for handling resources offers services to the Public Intranet network and its application servers. However, in customer premises networks, resources are managed by the customer. Thus, the end-to-end quality is a combination of these factors. The quality of service offered to a user may be matched with the static properties of that user's customer premises network.

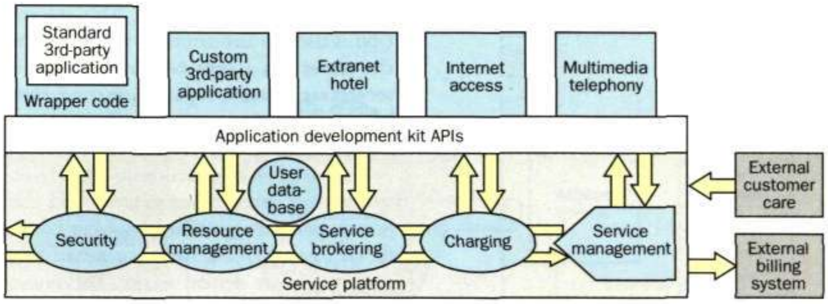


Figure 3
The Public Intranet service network consists of a service platform that provides services for the applications. Interfaces to external systems are also provided through the platform.

System architecture

The Public Intranet service network architecture is divided into two main parts: the Public Intranet service platform (PISP) and service network applications (SNA). The service platform represents a generic set of functions for deploying services, while the service network applications are a set of applications and services that are deployed on this platform to give end-users value-added services (Figure 3).

The service platform, which is the core of the Public Intranet product, provides support for selecting services and controlling their execution. The platform includes functions for supporting subscription, charging, resource handling, security, and management. The functions are made available to services through a set of application program interfaces (API) and tool kits. Applications that are deployed on the service platform use the functions to varying degrees, achieving different levels of integration with the Public Intranet. The platform contains an application development kit (ADK) for creating applications with functions provided by the APIs.

The service platform facilitates the development of distributed applications based on a three-tier model and the CORBA 2.0 standard specified by the Object Management Group (OMG). The common object request broker architecture (CORBA) provides the means for implementing a distributed-object environment and includes the Internet InterORB protocol (IIOP) for object interoperability, and the interface specification language (IDL) for language-neutral interfaces.

The service network, which is also based on distributed-object technology, makes extensive use of Java and Web technology to achieve a platform-independent implementation.

Integrating third-party applications

The APIs in the service platform constitute an open framework for the service environment. Services are integrated into the framework in one of two ways: they are developed directly with the APIs for charging, security, or resource handling or, if services already exist, they are encapsulated (wrapped) into the framework.

The application development kit, which is a set of classes and APIs that offer platform services, enables third-party applications to be deployed in the service network. Third-party applications are integrated into the platform using the API in their code modules.

Configuration management

A service management application enables the service network operator to manage subscribers, content providers, and others. The hierarchical definition of the management application permits distributed management, thereby enabling individual subscribers to manage their own end-users.

Resource management

The resource-handling function of the service network is divided into two parts:

- application server resources;
- network transport resources.

Application-server resources are monitored with respect to load, number of users, and consumption of bandwidth. This information is used in determining whether or not the server can accept new users. Thus, the servers are protected against overload, thereby guaranteeing users a specified quality of service.

The number of concurrent users from a single user group is monitored and compared with the maximum number of users for that group. If the limit is reached, additional users are denied access. This ensures that active users are guaranteed the quality of service that best matches the capabilities of the customer premises network and public network that serves it.

At the network level, resource monitoring and reservation depend on how the IP network and the underlying transport network have been implemented.

Functionality in routers and asynchronous transfer mode (ATM) networks affects the implementation of resource handling.

Security

The Public Intranet security model is based on a public key infrastructure that employs X.509 certificates for authenticating system users and machines. Thus, the authentication procedure does not transmit pass phrases over the network. Customers are given a scaleable system that supports distributed services and that can grow with the needs of an emerging infrastructure for new distributed network services, such as electronic commerce. The Public Intranet provides a multilayer security model that supports a variety of security requirements; that is, it ensures that security can be tailored to fit different markets.

The logon procedure is a three-way authentication routine that provides end-users with a single logon to the service network. Each end-user generates a digitally signed location update and audit-trail message that certifies his or her location (IP address) at any given time.

Using secure sockets layer (SSL) and IPsec, Public Intranet supports strong encryption of data at the network and transport layers. Therefore, encryption functions may be applied to session and stream-based services. The architecture also provides a suite of symmetric encryption protocols, which ensure additional support of different traffic requirements and characteristics.

Data integrity is protected through digital signatures using the same principles as for privacy. IPsec and SSL data-protection services may be added with low-layer public key cryptography standard (PKCS) APIs. Interoperability is ensured by means of standard algorithms for digital signatures.

A public key-certificate management system may be included in the system, which enables service providers to serve as a trusted certificate authority.

Service tickets contain information fields that prohibit replay and man-in-the-middle security attacks. Service tickets are valid for a specific application server, which enables the broker to balance load.

Support for network, transport and application-layer security services allows security to be added transparently to applications and services that were not designed with security in mind. Tool kits and guidelines, which include a set of mechanisms and policies for protecting the customers and owners of the intranet, adequately protect the Public Intranet and its services, resources, and equipment against misuse.

Charging

Public Intranet charging comprises functions for collecting usage data from applications and the service platform. Usage data is stored, formatted and forwarded to an external billing center.

Charging is achieved in a client-server fashion, since the client provides an API for the applications. The application to be charged runs on an application server. A special application generates event-based records using the charging API. Before charging records are sent to the charging server – which runs on a dedicated machine – the charging client encodes them. Charging records may be sent individually, as they are generated, or stored in a file for later delivery. The charging client monitors the connection to the charging server, reestablishing the connection, if lost. During interruptions, charging records are written to a local database. The charging server monitors the connection to the external billing system in the same way.

Records from various charging clients are accumulated in the charging server. These are then sent to an external billing system for post-processing in accordance with the format and protocol requested by the external system.

The charging server and clients use SNMP traps to report special events to service and network management systems. The application uses the service ticket for linking a unique user identity to the use of a service.

Brokers

Brokers have a server and a client side. The client side handles the interface to smart cards, which are used for identifying end-users and for presenting the applications available to them.

The server side handles authentication and controls access to the list of available applications.

End-user profiles, which include the list of application subscriptions, are fetched from the user database.

By using a smart card that includes public key infrastructure (PKI) certificates, end-users can invoke a service network access application that uses a security API for signing messages sent to the broker. Using an X.500 database, the broker verifies incoming logon messages with the security API. The broker then retrieves the user profile from the database. It also updates the user record with data on the user's current location; for example, to serve incoming calls by the multimedia telephony system (MMTS).

Based on user and terminal profiles, the broker performs a filtering function to derive the most suitable service offering. Similarly, based on the profile of the terminal from which a user logs on, the broker may optimally downgrade the available services. For example, if a user has subscribed to a service that requires a video camera, but the current terminal is not adequately equipped, then the broker can disable the service. The service offering is returned to the user's graphical user interface (GUI).

The user may then select and invoke a service from the list, issuing a signed service request that provides a non-repudiation service.

The broker authenticates the message and grants a ticket for executing the service. Tickets are only granted when a service is available. The ticket is sent as an input parameter when the selected service is launched.

An enhanced CORBA-naming service gives the broker service high availability. Consequently, several broker servers are registered for availability in the naming service. The naming service abstracts the location of the network service. The same mechanism is used when a broker selects an application server.

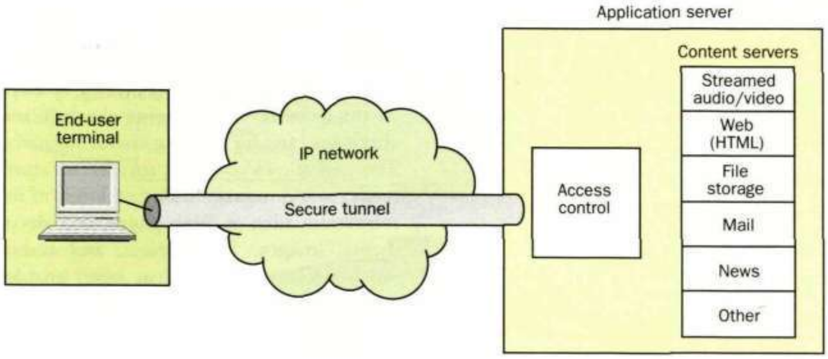


Figure 4
Content hotels offer secure access to information stored on content servers. Content is encrypted and transported via a secure tunnel from the application server to the end-user terminal. Access content is controlled using mechanisms in the Public Intranet service platform.

Application servers

Applications that use the Public Intranet service platform run on dedicated application servers. These servers contain parts of the platform and provide APIs to the applications.

Application servers may be added independently in order to increase capacity. The brokers balance load and handle redundancy between the servers. Thus, in addition to being fully scaleable, the service network fulfills requirements for high availability.

Public Intranet services

The system includes several end-user applications designed to take advantage of the functionality provided by different system components. Network operators may easily add new services using the application development kit.

Content hotel

The Public Intranet content hotel provides functions for distributing and retrieving content. End-users may visit a content hotel to search for, browse, or retrieve data. When the end-user selects a content hotel service, a secure-tunnel SSL is established between the end-user terminal and the application server, guaranteeing that the selected content may be transmitted securely (Figure 4).

With support of user authentication and user-based access control for each selected page or file, the content hotel facilitates the implementation of content rooms (data repositories), thereby allowing content to be distributed to selected groups of end-users. Service providers register new content providers, change privileges, and perform other administrative tasks via a management interface.

A tool is provided for registering and updating content, for managing check-in and check-out, and for viewing visitor statistics. The tool allows content providers to store, retrieve, and manage different kinds of information, such as Web pages, text documents, images, Java applets, and audio-video streams. The content hotel enables content providers to publish information on the Public Intranet.

The content hotel consists of a main application server and several content servers, which handle distribution. The actual content is stored on disks. All content servers are controlled by the application server, which handles core services – those services that charge data, invoke other services, and handle faults and events. The application server also hosts Java applets, which may be downloaded and executed on the client. Various content-charging schemes are supported, based on connection time, bandwidth usage, and kinds of service invoked.

Among the content servers are:

- a Web server – which manages content and the distribution of HTML files, with or without Java applets, as well as the distribution of HTML- or Java-based user interfaces to the client;
- mail and news servers;
- streamed audio and video servers.

A catalog server is used for searching content stored in the content hotel.

Audio and video servers handle the streaming of audio and video content to the client.

Virtual private intranet

The virtual private intranet (VPI) service allows a set of protected user domains to share common service and network resources without compromising the integrity and privacy of each domain. Resource sharing makes it possible to model highly cost-effective intranet and extranet solutions for large and small communities of users. The VPI service relies on the access control and privacy functions provided by the security framework. These functions ensure that each VPI remains a closed user community.

Directory service

A search function, which is supplied by a directory service, provides information from the database on PISN end-users. The directory service may be used directly by an end-user via a graphical user interface, or through an application that uses a lightweight directory access protocol (LDAP) interface.

External databases with an LDAP interface may also be searched using the directory service.

Internet access

Internet access is provided using the access control in the content hotel. This makes it possible to customize access privileges for different user groups and to obtain detailed user statistics.

Good characteristics are achieved by adding extensive caching, by replicating recent Web data, and by replicating or mirroring popular Web sites on local servers.

A high-performance super-Internet service may be offered by implementing broad bandwidth access; for example, via asymmetric digital subscriber lines (ADSL) for residential users.

Multimedia telephony

The Ericsson multimedia telephone is more than an IP telephone application. It is a complete IP-based video, data, and voice conferencing system incorporated into integrated services digital network (ISDN) and the public telephone network. It adheres to the ITU-T H.323 standard and consists of clients, a gatekeeper, a multipoint control unit (MCU), and gateways to the public switched telephone network (PSTN) and to ISDN.

The multimedia telephony system enables interpersonal communication between desktop users over a high-capacity IP network using integrated video, audio, and data. It provides the same benefits as a full-fledged public telephone system, including reliability, charging, and interoperability with PSTN telephony and ISDN video telephony (Figure 5).

The H.323/T.120 client terminal, which provides end-users with an interface to the multimedia telephone network, serves the same purpose as an ordinary telephone in the PSTN augmented by the functionality of a multimedia telephone. Clients equipped with a camera may enjoy high-quality video conferencing capabilities.

The H.323/T.120 multipoint control unit controls multiparty sessions and mixes data, audio, and video streams.

All multimedia telephone endpoints (terminals, gateways, and MCUs) use a gatekeeper as an intermediate point for call and control signaling. By piping all computer telephony call and control signals through a gatekeeper, the network can provide the requisite access control, security, and charging mechanisms offered by Public Intranet.

Gateways – which enable interoperability with existing PSTN/ISDN networks – manage the conversion of call, audio, video, and control signals, making it possible for a Public Intranet client to connect with an ordinary PSTN telephone or an ISDN video conference. Using the multimedia telephone system, one party can call or be called by parties in the PSTN who use regular telephones.

The multimedia telephone system allows users to exchange information found on their respective multimedia telephone clients. Data conferencing applications, such as character-based communication tools, a shared whiteboard, and shared applications, are integrated into the multimedia telephony service.

The system also transfers live video of participants in a call – thus, people talking on the multimedia telephone may see each other. Using the multimedia telephone, parties can initiate multiparty conferences, which comprise H.323 endpoints or a mix of H.323, H.320 and PSTN endpoints.

As with any application developed for or integrated into the Public Intranet, the MMTS application generates charging records based on the user's ID, duration of the call, and the use of service network resources. The MMTS service provider uses these charging records as a basis for billing end-users.

The gatekeeper introduces end-user mobility to the system by registering all end-users who are logged onto the Public Intranet, including their current locations. A directory service facilitates the lookup of end-user addresses, including E.164, a universal personal telecommunications (UPT) number, a GSM number, or a mail alias. Service-specific QoS parameters, such as the actual load on the MCU and gateway, are monitored by a built-in resource handler.

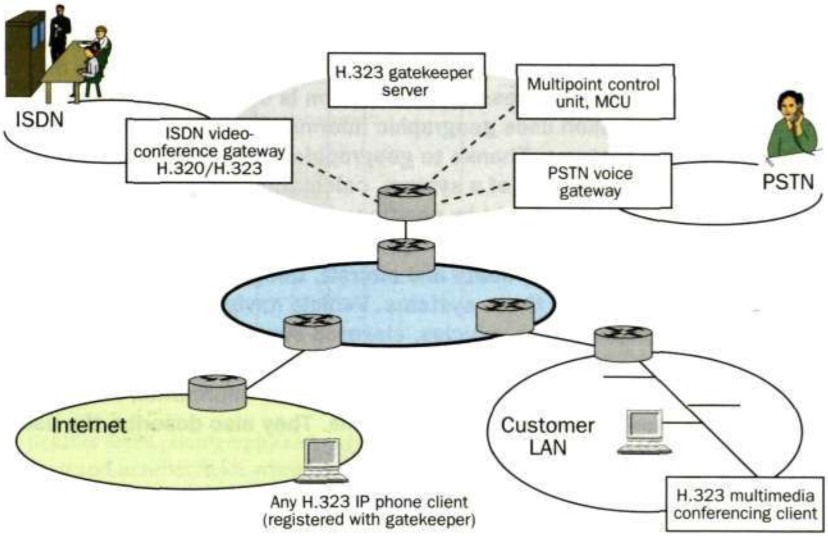

ISDN
ISDN video-conference gateway H.320/H.323
H.323 gatekeeper server
Multipoint control unit, MCU
PSTN voice gateway
PSTN
Internet
Customer LAN
Any H.323 IP phone client (registered with gatekeeper)
H.323 multimedia conferencing client

Figure 5
The multimedia telephony application allows users in the Public Intranet to connect to users in the PSTN and ISDN via gateways. The multimedia telephony application offers IP telephony, video conferencing and shared application functionality.

Conclusion

The Public Intranet defines a business concept that provides a framework for developing and deploying commercial IP-based services. The framework includes support services, applications, and an IP network solution. The Public Intranet effectively combines new services with techniques for supplying secure, high-quality service in large networks, giving service providers a foundation on which to develop their businesses. With it, service providers can move up the value chain by offering new, more valuable services for which end-users are willing to pay a premium. Examples of Public Intranet services are a content hotel, virtual private intranet, virtual private network, Internet access, and multimedia telephony.

The system architecture is divided into two parts: the Public Intranet service platform and service network applications. The service platform consists of a management server node with common resources and interfaces to external systems, brokers and application servers. The IP network consists of a set of customer premises networks that are interconnected by an IP network. The service center comprises a set of servers that provide services to end-users on customer premises networks. The Public Intranet contains necessary tools for implementing security, including firewalls, packet-filter functions, and Internet gateways.

A service resource-handling function of the service network fulfills two objectives: it guarantees that services are supplied in accordance with end-user requirements, and it gives service providers a means of monitoring and controlling the use of resources.

Geographic information systems

Hans Brandtberg, Johan Frössling, Lena Lüning and Mats Åkerlund

Geographic information systems are growing in prevalence and their use can be found at many of Ericsson's locations. Digital geographic information is used in various calculations and for presenting maps on electronic displays. The quality of all wireless communication is dependent on terrain, and that is why Ericsson uses geographic information systems for planning mobile telephone systems. Thanks to geographic data, network engineers can optimize the dimensions of a system, calculating the exact location at which each base station should be positioned.

Ericsson develops display screens for command and control systems that are integrated into vehicles, boats and aircraft. Geographic information plays an important role in these systems. Vehicle navigation is improved with maps that display other vehicles, elevated obstacles, and a preferred route of travel.

The authors describe how Ericsson uses geographic information for planning radio networks and mobile telephone systems. They also describe the use of geographic data for displaying maps in vehicles.

Geographic information systems (GIS) are becoming more and more widespread. Today, Ericsson includes them in several different areas and applications. Two such areas, which are highlighted in this article, are radio network planning and digital maps in aircraft.

In the 1970s and early 1980s, geographic information systems mainly consisted of presenting maps on a display to support various types of geographically related information. Early systems could not be used for analyses and calculations, since access to information was limited and computer-processing power was inadequate.

More recently, however, considerable gains have been made in technology. Geographic information systems are now being called upon to perform analyses and calculations using digital geographic data. The data is still being presented on displays – usually simulating the layout of a paper map – although other presentation methods exist as well; for example, relief structures for displaying height relationships.

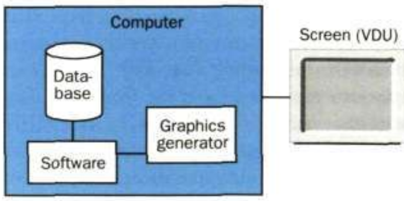


Figure 1
Block diagram of a geographic information system.

Compared with other computer applications, systems with geographic information usually require larger volumes of memory and a high-performance computer. However, again, thanks to gains in computer technology and improved data access, the number of applications with geographic data is now quite large. In the future they will be used even more widely.

Geographic information applications fall into three main categories:

- Background maps for different types of information – data on an object and its position are superimposed on a map, sometimes in real time, to create a good overview. Users may also use a background map for drawing route plans and for inputting other map-related information in layers on top of the map. Background maps are found in planning and vehicle-guidance applications.
- Computational aids – geographic data can be used for calculating the optimum position of base stations in a mobile telephone network; for the deployment of radar stations; and for planning the best route for building a road. The results are often displayed in the form of a map. However, some applications use geographic information systems for computation and analysis without displaying the results. For example, geographic data may be used to improve the performance of a radar sensor. Drawing on this data, sensors can more easily distinguish between the reflected signals of geographic and non-geographic objects, such as a boat.
- Navigational aids – information is presented in real time on a map display that is linked to a navigational sensor – for example, to an inertial- or satellite-navigation system (global positioning system, GPS) – that gives a steady read-out of the current position. Planned and actual routes can be presented on the map. If a radio communication system for digital information is included, the position of other vehicles and other geographic information may also be received and superimposed on the map.

A geographic information system consists of several parts (Figure 1), including (a) a database with geographic data and (b) a computer with software, which fetches data for calculations. The computer is equipped with (c) a graphics generator for presenting images on (d) a display screen. The systems may be configured in many different ways. A simple configuration consists of application software and an ordinary personal computer. A more complex configuration might involve a fully integrated system that combines the GIS with other functions in an avionics module onboard an aircraft.

Maps and geographic information

Opportunities for using a geographic information system largely depend on availability and the quality of digital geographic information. Happily, significant improvement has been made in each of these areas in recent years. Nonetheless, in terms of the scope, quality and characteristics of geographic data, much improvement is yet needed to meet the growing collection of GIS applications. In Sweden, the main suppliers of geographic data are the National Land Survey Administration and the National Maritime Administration which, together with other organizations, such as the Geological Survey of Sweden, the Swedish Space Corporation, and units of the national armed forces, answer for the development and supply of geographic data.

Geographic data must be kept up to date. If this requirement is not taken into consideration, data will age and become useless. Thus, developers and engineers of geographic databases must ensure that their databases can be updated.

Geographic data falls into two main categories, depending on how it is created and used:

- Cartographic data that mainly consists of geometric information is used for presenting maps on a display.
- Geographic data that represents reality.

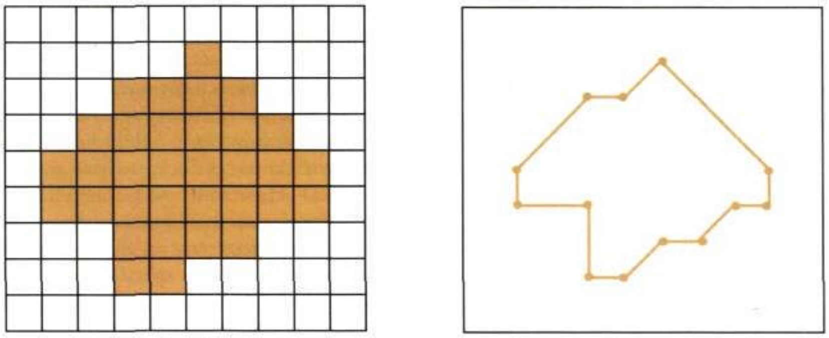
Geographic data is usually described by means of raster data or vector data (Figure 2).

With raster data, information is built up by a regular right-angled grid, where information in each point is stated in code. Some raster data is relatively easy to create, by scanning maps or, preferably, their print originals. Other data, such as digital terrain models, requires a more complex process. Raster data occupies considerable memory, but with the help of various compression algorithms the data can be reduced in size. Since both data and displays use raster points, it is easy to present maps on raster display screens. Raster data is required in the fields of aerial photography and satellite imaging and in other fields where it is necessary to obtain digital data of a new area very quickly.

With vector data, information is built up by defining the location and range of a geographic object. Point objects are defined by their position and attributes; linear objects by associated range and attributes; and surface objects by their enveloping edge (perimeter) and attributes. Vector data provides true geographic information that, compared with raster data, is not related to how the presentation appears on a display. Depending on how the information is to be used, different characteristics and attributes must be added. For example, for a roadway network to calculate the best route, it must be linked to real points of intersection. For other decisions, the geographic data must also contain topology, so that the system can calculate relationships and connections; for example, to determine whether or not a road enters a forest.

In the foreseeable future, raster data and vector data will continue to be used in the applications to which they are best suited. Notwithstanding, vector data is expected to grow in importance in systems whose diverse analyses and decision-making functions rely heavily on attributes and topology.

When geographic data is supplied or exchanged, it is formatted for transmission; today several different formatting standards exist. It is relatively easy to translate from one format to another, if the formats have been specified and use the same terminology. A formal standardization work is currently under way – for example, the STANLI project in Sweden works together with other European and international standardization efforts.



Raster data in a grid

Vector data with connected points

Figure 2
The structure of raster and vector data.

Box A
Abbreviations

DECT	Digital enhanced cordless telecommunications
EET	Ericsson engineering tool
ENPT	Ericsson network planning tool
GDT_EET	Geodata transmission_Ericsson engineering tool
GIS	Geographic information system
GPS	Global positioning system
GSM	Global system for mobile communication
GST	GeoBox support tool
OSS	Operations support system
OTW	Out the window
RAPS	Radio network planning system
RWO	Real-world objects
TIFF	Tagged image file format

Geographic information systems

As has already been mentioned, the storage format with associated attributes and topology plays an important role in how well a GIS can or cannot perform analyses and calculations. Algorithms and functions are being developed that simulate how geographic data affects various phenomena. At Ericsson, this applies specifically to the propagation of radio waves and radar energy. System results are improving, thanks to access to better, more detailed data, and to the ongoing development of algorithms.

Results are usually presented on a color display unit. Since the resolution of such displays is inferior to a paper map, digital maps must usually be magnified (zoomed). Symbols and text must also be added, providing links to the computerized functions of the application.

In order to present an image from geographic data, the system must know the current display attributes – these attributes, which state how the object is to be presented, are separate from the geographic attributes that define real characteristics. For instance, for an object called Highway, a geographic attribute would be used to define the width of a road. By contrast, a presentation attribute defines how to display an object of that width. For ease of use, the presentation usually simulates a paper map, although other presentation methods are also used – for special system characteristics or demands.

The level of detail that can be derived from data determines to what extent it can be used for calculations and presentations. In terms of presentation, the level of detail must be adapted to the size and resolution of the display. Excessive detail produces a map image that has too many small contours and is therefore difficult to read. Also, detailed images take longer to draw. Thus, it is sometimes necessary to generalize data, in order to present it on a smaller scale. For example, let us assume that the detail of data in our system database corresponds to a topographic map drawn to a scale of 1:50,000, which is adequate for presenting scales between 1:10,000 and 1:50,000. However, to use this database for presenting smaller scales – say, between 1:100,000 and 1:200,000 – the data must first be generalized. Generalization, which entails straightening contours and eliminating minor objects, considerably reduces information per geographic area. Ordinarily, the generalization process cannot be performed in real time. Instead, many systems store databases of the same geographic area on different scales. This is relatively easy to do, since the largest scale contains the primary base of information.

Users very often want to show or hide specific object classes, displaying only those features that truly interest them. By hiding a portion of regular map information, other information may be superimposed more legibly. This is one of the prime advantages of presenting maps on an electronic display – the information can be adapted to show what a user wants to see at any given moment.

Radio network planning

Radio communication has grown considerably in the past decade. Frequency ranges are constantly being extended higher and higher, yielding greater transmission capacities. At the same time, radio systems are being developed to better exploit existing frequency ranges. Thanks to significant cost and time reductions, radio communication can often complement or replace transmission over copper or fiber-optic cable.

A radio network may consist of

- point-to-point connections with radio links for large-capacity transmission – for example, using the Ericsson MINI-LINK;
- surface-coverage systems for personal communication – for example, by means of cellular telephony (GSM);
- point-to-multipoint systems
  - – for fixed, installed personal communication – for example, using DRA 1900 (DECT);
  - – for business communication – for example, AIRLINE.

A common feature of all types of radio communication is that the transmission quality is heavily influenced by the characteristics of the terrain, obstacles, and reflection between the transmitter and receiver. Thus, given the risk of interference between each transmitter and receiver in a radio network, the larger the network grows the greater and more complex the risk becomes.

The secure projection of a radio network requires modern, computerized aids that can manage and analyze each input connection. The radio network planning system (RAPS) is a Windows-based application developed by Ericsson Business Systems for planning any type of radio network.

RAPS contains functions for analyzing radio-wave propagation and interference, and support functions for building radio networks. It dynamically handles geographic and organization-dependent information as well as presentation functions for maps, graphs and diagrams. The RAPS geographic system is made up of three separate parts: a geographic database, analysis functions, and map presentations.

The geographic database, called GeoBox, was developed by Ericsson for the Defense Material Administration of Sweden. It is a model for storing and accessing basic geographic data, and was developed especially for operational run-time use and for exchanging geographic data. Core geographic data consists of geometric data represented in raster, vector and text format together with associated attributes and descriptive data (metadata).

A GeoBox database contains all necessary information for reusing data in different applications without having to manually describe its contents for each individual case – interpretation, geographic reference system, data quality, etc. The information, which is accessed via function calls and simple data types, is chiefly intended for application programmers. By using filter functions – such as resampling raster data or filtered attributes of vector data – the application can determine how data is meant to be displayed. GeoBox may even be used for dynamic, operation-specific information, such as data on relative humidity and field-strength measurements. Several tools have been developed for supporting and maintaining information in GeoBox:

- The GeoBox support tool (GST) imports and controls the quality of data from many different standard and non-standard formats and coordinate systems.
- The GDT_EET supports the automatic transmission of data from the standard Ericsson engineering tool (EET) format.

The majority of map presentations are based on stored geographic information; that is, the system does not store ready-to-use map images. The data that the system uses to present a map is also used in the geographic operations that make up analysis functions. For instance, an elevation database, with height values in a 50-meter grid structure, is used for diverse map presentations, including relief maps and elevation curves (Figure 3), and for calculating line-of-site coverage (Figure 4) and obstacle attenuation (path profiles) (Figure 5).

Each presentation makes use of a dynamic presentation filter, which enables the same database to be used for a broad range of scales. In general, the maps that can be displayed are a combination of one or more basic maps, such as vector maps, text, raster images, elevation curves, elevation layers, relief maps, and thematic maps. Color and symbols are managed dynamically within different ranges of scale, and can be stored for reuse.

Examples of geographically based analysis functions are different types of coverage

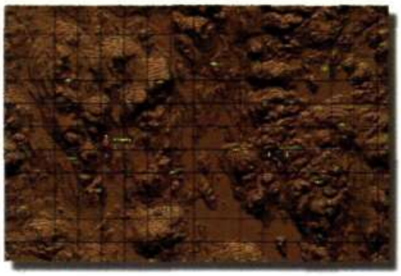


Figure 3
Topographic information from a 50 m raster database shown as elevation contours and with shading from an imaginary illumination from out of the northwest. Two point-to-multipoint networks are shown with radio terminals connected to the sectors in the base stations.

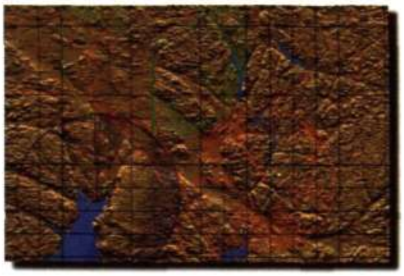


Figure 4
Topographic information from a 50 m raster database; roads and lakes from a vector database; map text. The radio coverage of three bases stations is superimposed.

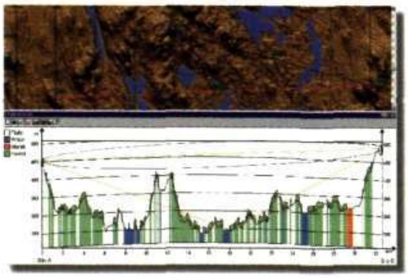


Figure 5
Topographic information shown on a map and as a profile between two radio link stations. Land usage information is identified by colors in the profile.

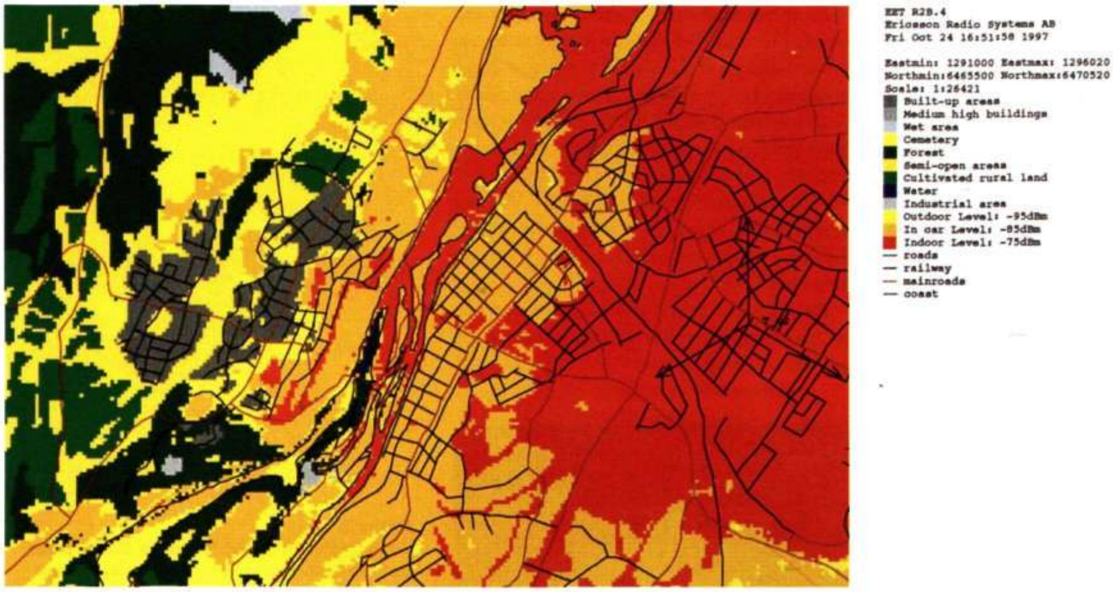


Figure 6
Graphic image of radio cells presented on a map of land-usage classes; background data is obtained from the EET.

area; that is, the symbolization of areas, in the form of an optical layer, that are covered from an observation point, or the field strength of one or more base stations. As a basis for this kind of analysis, one or more information layers, such as ground elevation and other ground characteristics, are generally used.

Planning mobile telephone systems

When planning a mobile telephone network, engineers must know the shape and appearance of the terrain. Anything that can affect radio signals – height, open surfaces, vegetation, buildings – must be accounted for when calculating radio-wave propagation and interference. Obviously, this requires the modeling to be based on good, up-to-date geographic information. Ericsson Radio Systems has developed the EET for planning radio networks. The EET is based on PlanNet, which operates in UNIX environments.

The EET, which may be used for planning the networks of several different analog and digital radio systems, manages geographic data in three layers. The *elevation database* and the *land-usage classes* – layers that are used for calculating the radio-propagation models – are stored as raster data with a resolution of between 20 and 100 meters, whereas all *background data* (roads, railroads, coastlines, etc.) is stored in vectors. Text may be added to improve orientation. Other supplemental background information, such as a TIFF image, a scanned map, or a satellite image may also be added. Thus, the EET handles a mixture of vector and raster data. It may be used to present a graphic image of radio cells (Figure 6), their coverage areas, and internal interference relationships. This enables cell planners to test the location and configuration of different antennas quickly, which is especially useful when planning the expansion of a system in a fast-growing environment, such as a large city.

Working outside of Sweden to the extent that they do, it is often difficult for Ericsson Radio Systems to acquire data. In many countries, maps and geographic information are still entangled in many restrictions. Access to existing digital geographic data – if it exists at all – is usually poor. Moreover, the data is seldom processed for use in radio network planning. Therefore, Ericsson Radio Systems has established its own map unit to serve the entire Mobile Systems Business Area. The unit has been assigned the task of developing

- geographic databases – for planning radio networks, mainly using the EET;
- less-detailed map data
  - – for other planning;
  - – for supervisory systems, such as the Ericsson network planning tool (ENPT) and operations support systems (OSS).

Basic input data is generally taken from topographic maps whose scales are between 1:20,000 and 1:100,000. This data is sometimes supplemented with satellite images. Translating a map from analog to digital form requires manual digitizing. For this task, the unit uses the GIS product Smallworld, which is an object-oriented geographic information system that makes use of real-world objects (RWO) complemented with Ericsson applications. Afterwards the data is rasterized and converted to its current form. Positional data, which is a necessary ingredient of geographic databases, is based on three parameters: the map projection, the reference ellipsoid, and the coordinate system; that is, the geodetic datum. In terms of map production, these parameters differentiate themselves in different parts of the world and in different map systems. The introduction of GPS technology has amplified the importance of handling geodetic data. It has also put stringent requirements on the GIS applications.

The increased use of mobile telephones has made it necessary to plan radio networks down to the level of microcells. Doing so requires minutely detailed, large-scale map data that includes parks, the width of streets, and building heights. The need for more detailed data, which introduces completely new requirements into the production of map data, has set in motion intensive development efforts at Ericsson.

Digital map systems for vehicles, vessels and aircraft

The role of display screens in vehicles, vessels and aircraft is growing. Modern computers can adapt and manage information for various needs; information can be processed, compiled and presented in a suitable fashion. One of the most crucial areas of information concerns a vehicle's surroundings – that is, the terrain or environment in which it travels. Today digital geographic databases can be obtained from map manufacturers in several countries. Moreover, access to digital geographic information is on the rise. Positioning and navigational systems that offer adequate performance at a low price have also begun to appear on the market. These systems enable a map image showing a vehicle's current position to be presented on a display, which helps operators (drivers and pilots.) to navigate and survey their present situation. Additional information that relates to the geography or to a location on a map may also be presented on the display. Examples of this include planned and actual routes of travel and the position of other vehicles.

Advanced displays are usually only found in airplanes, helicopters, boats and military vehicles. Modern jet fighters contain several such displays, which are used to present flight, navigational and tactical information. The presentation of information, which is available for immediate viewing at any time, may be altered manually or automatically. The layout of the display is designed so that crew members can assimilate information in a variety of conditions, such as total darkness or direct exposure to sunlight, or while experiencing heavy vibration, extreme g-load and stress.

Many vehicles interact with other vehicles and with the terrain. Their maps must be presented in real time (live motion), showing the direction of travel straight up on the display. A good presentation reduces the time operators need in order to orient themselves when they view the display. One example is Ericsson Saab Avionic's display system for the new JAS 39 Gripen jet fighter. Figure 7 shows the tactical display, which presents a background map with tactical information superimposed on it.

The system contains a digital geographic database in several scales. A computer reads the database continuously and draws a moving map image on the display. Since the aircraft travels at high velocities, a new image must be drawn approximately 30 times a second. The presentation can show different information content on different scales. The pilot can choose, manually or automatically, to have information displayed or hidden. The map symbols, which indicate roads, railroads, towers, and so forth, differ slightly from the symbols used on paper maps. This is because they must be adapted to the display unit; they must also be extra legible under a range of different flight conditions. A great deal of positional information is superimposed on the map. The plane's position is obtained from the navigation system. The intended flight plan and assignment are also shown. Position and objects of different kinds – provided by onboard sensors and databases, and by the plane's data links to other aircraft and to the command and control center – are superimposed on the map in real time.

A special system, called MDX, has been

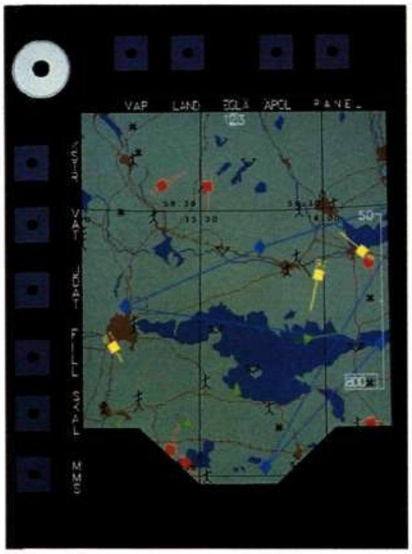


Figure 7
Presentation of a digital map with superimposed information on a horizontal situation display in the JAS 39 Gripen jet fighter.

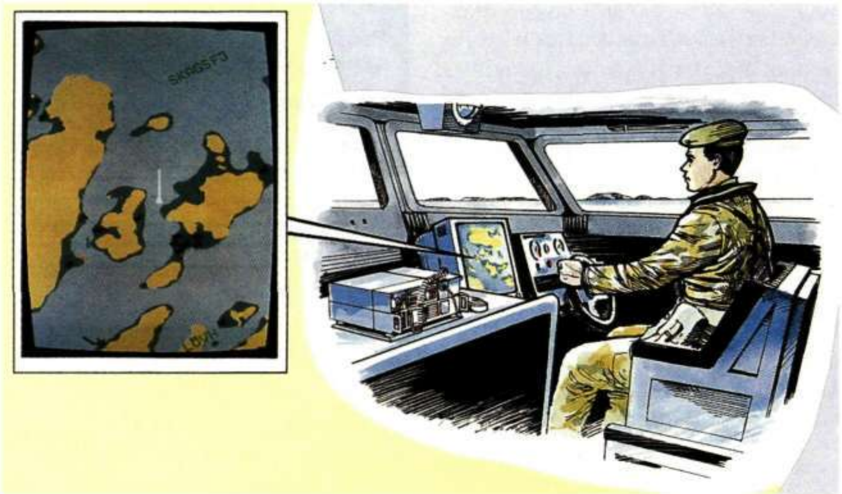


Figure 8
Presentation of a digital marine chart for navigation by high-speed boats.

developed for processing and adapting digital map information. Core data is based on available digital information from the National Land Survey Administration and from the Defense Material Administration of Sweden. The information to be stored onboard the aircraft is selected in MDX. Generalizations are added to create several databases in different scales from the same data. Having multiple databases facilitates searches when the map is presented. MDX uses a special technique for compressing data.

Small-scale maps are used more frequently in aircraft than in other applications, since aircraft usually travel at higher velocities and operate in larger areas than other vehicles. Thus, airborne geographic information systems generally require less data storage capacity than maritime vessels, which are highly dependent on detailed marine charts.

Geographic information systems for maritime navigation (Figure 8) are used to present digital marine charts, which are extremely valuable for high-speed boats and ships that operate near the coastline or in an archipelago. The systems may also combine output from a global positioning system to indicate a vessel's current position. Since the map layout must be displayed as an ordinary marine chart, the design of the display has been defined in a standard.

Besides the regular map layout, there are numerous other ways of presenting geographic data. For example, geography can be displayed in perspective; that is, as a synthetic image of the terrain as viewed from a fixed point (Figure 9). Perspectives are commonly used as an out-the-window (OTW) system in different types of simulator, but they are also found onboard vehicles, enabling crew members to survey and interact with the terrain.

Conclusion

The growing significance of geographic information technology has lead to an increase in the number of players in this field. Several publicly funded centers of competence with links to universities have been established. In some cases, professorships have been established. Various interest organizations and networks have been built up between organizations and within companies. To date, a great deal of literature has been written on display-based maps and geographic information systems.

Geographic information systems are used at Ericsson in several different applications. They comprise analytical and computational functions for planning radio networks, in general; and more particularly for planning mobile telephone systems. Further, they are found in the display systems of various vehicles and aircraft, such as the Swedish JAS 39 Gripen jet fighter.

Being able to exploit geographic features well has always been important. The development of better tools and access to better geographic data will greatly improve and raise the level of efficiency in all types of system. In the future, the use of computer-based maps and geographic information systems is expected to grow, gaining importance in a number of different systems.

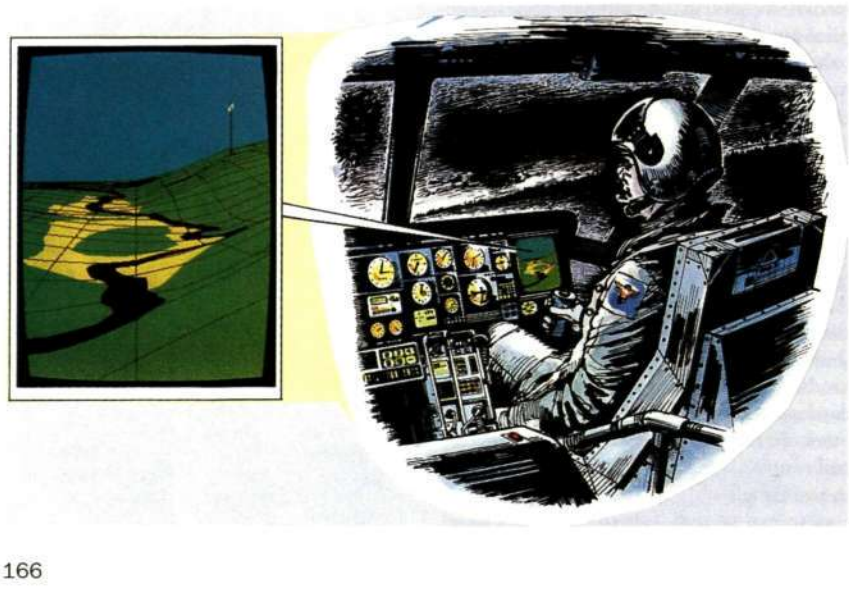


Figure 9
Presentation of a perspective map of the terrain in a helicopter application.

The NDB Cluster – A parallel data server for telecommunications applications

Mikael Ronström

The number of services in telecommunications networks is increasing. Most of these services require the network to store information that must be made available in real time. Such information is critical: failure to reach it often means that a service cannot be provided.

The number of information services available through the Internet is also increasing. The very essence of these services is to give users and applications access to information. Databases that support services of this kind must offer exceptional performance, very high availability, and real-time response.

The author describes the NDB Cluster (a parallel database server currently under prototype development at Ericsson), illustrating its use through two applications: the number-portability application and a Web cache-server application. The NDB Cluster – whose platform is based on AXE techniques – contains several novel solutions to high-availability-related problems while fulfilling requirements for high performance and real-time response.

Box A
Abbreviations

ASIC	Application-specific integrated circuit
DBMS	Database management system
DNS	Domain name service
HLR	Home location register
HTTP	Hypertext transfer protocol
INAP	Intelligent network application part
NDB	Network database
ODBC	Open database connectivity
PC	Personal computer
SCI	Scaleable coherent interface
SQL	Structured query language
TCP/IP	Transmission control protocol/Internet protocol
VM	Virtual machine

The NDB Cluster (Box B) is a database for storing all types of semi-permanent data for telecommunications services, including data that would otherwise be stored in application memory and in file systems. It has been engineered to fulfill current and future telecom application requirements for safe storage, high performance and short response time.

Applications and requirements

Databases in telecommunications networks, here called network databases (NDB), fall into three main network categories: service networks, management networks and information networks. Some network databases also make up part of communication networks (Figure 1).

Service network databases, which help operate telecommunications networks, include mobile databases, number-portability databases, service control nodes, name servers and route servers. These databases work with high rate-of-message applications. Most requests are small, however. Some applications merely read data; others also write data. The data in service network databases may be stored in main memory. Future ap-

Box B
NDB Cluster

NDB Cluster is the name of the prototype. NDB stands for network database, where network means telecommunications network. The term cluster indicates that it executes on a cluster of workstations or personal computers.

A system is a node in the telecommunications network. The NDB Cluster is part of a system. Thus, a system generally consists of the NDB Cluster plus several application and management servers.

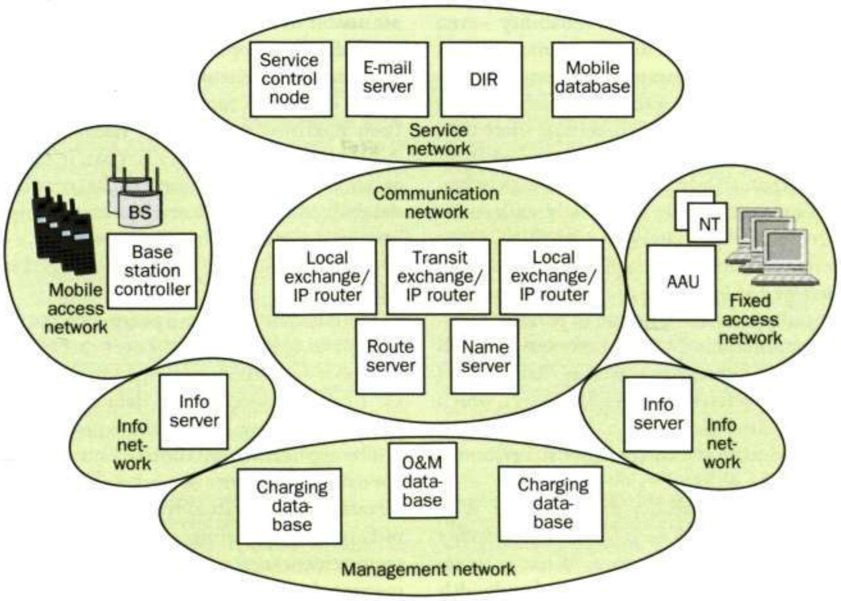


Figure 1
A database-centric view of the telecommunications network. Network databases are needed to support communication setup and services. They are also essential for operation and maintenance and for supporting applications that deliver information to users.

Box C, PLEX

The PLEX design environment offers many benefits:

- PLEX code is divided into blocks that may only communicate through signals. This makes PLEX suitable for handling large development projects.
- PLEX provides unique support for on-line testing. Testing and debugging a distributed database is a difficult task that requires powerful tools. The PLEX environment provides such tools, many of which are unavailable in the C++ development environment.
- Being a real-time language, PLEX enables engineers to design a database that fulfills real-time requirements.
- PLEX allows live software upgrades without requiring support of old software. This feature can be exploited for fixing bugs and for adding new functionality.
- Because only asynchronous signals are allowed between blocks (in the NDB Cluster), each block's residence is completely transparent. Therefore, a signal's destination can easily change from within a server to between servers.

plications with complex services might be required to interact several times with the database service. Thus, response times must be below 10 ms; for some applications, as low as 1 to 5 ms.

Management network databases help manage the operations of telecommunications networks. A typical example is a charging server from which toll-ticket records are distributed to various management systems. Real-time charging requirements imply that these servers must be highly reliable, real-time databases. Charging-related servers support network operations and the services provided in telecommunications networks (for example, information services). Databases of this kind receive roughly the same amount of messages as network databases in service networks. However, they must also store very large amounts of information and support more complex queries.

Information network databases are an integral element of the applications found in telecommunications networks. For the most part, they provide information services through Web servers, e-mail servers, directories, and news-on-demand services. These databases must have large storage capacity and support a high rate of messages per second. Information services must be able to send very large information streams – typically at the rate of thousands of bits per second. When there are many users, the system must send several gigabytes of data per second.

Each application must also fulfill very stringent requirements for reliability – even more stringent than for telecommunications exchanges. For instance, the annual allowable downtime for a network database is less than one minute. This is because more than 10 times as many users rely on network databases than on telecommunications switches.

This article deals specifically with number-portability databases and Web cache servers. Number-portability databases are involved at least once or twice in every call. Initially, they are called on to perform number translations. In time, however, they will also provide security as well as charging and directory services. Web cache servers, which are accessed each time a Web page is requested from the Internet, must frequently send large objects.

Number-portability databases and Web cache servers must each fulfill requirements for very high performance. What is more, Web cache servers require large bandwidth for query responses, whereas number-portability databases must guarantee very high reliability.

Studies suggest that future applications will be required to handle as many as 10,000 requests per second for the up to one million users connected to the network database[1]. Depending on the application, requests will be made to read, to write, or to read and write data. Some applications will also be required to send large objects.

Design considerations

Different application characteristics put heavy demands on databases and their platforms. Traditional telecommunications platforms lack information capacity and support of large information streams from a server application. In meeting demands for message capacity, information capacity, large information streams and very high reliability, Ericsson's engineers have approached the task from a distributed platform.

Given the requirements for efficient handling of data distribution and real-time response, the distributed database (to which many messages arrive simultaneously) must be able to trace performance and to handle asynchronous messages easily. The model found in the programming environment for PLEX is well suited to these design requirements (Box C). Other programming languages, such as C, may be used for special subroutines.

Current AXE control system platforms are unable to meet all the needs of every network database application. For this reason, the so-called APZ emulator, which was originally developed for testing purposes, has been modified extensively. Today, as the AXE virtual machine (AXE VM), it serves as an execution platform of the network database. Aside from some blocks of the operating system, which were written in C++ (Box D), the database is being developed in PLEX.

The database runs on separate processors. This gives it better control over caches and processor resources and improves reliability. In small systems, the data server and other servers may need to share processors.

The applications are isolated into special application servers. Several management servers may also exist in the system. To simplify recovery handling, the application and management servers should not contain permanent data. That way, no information is

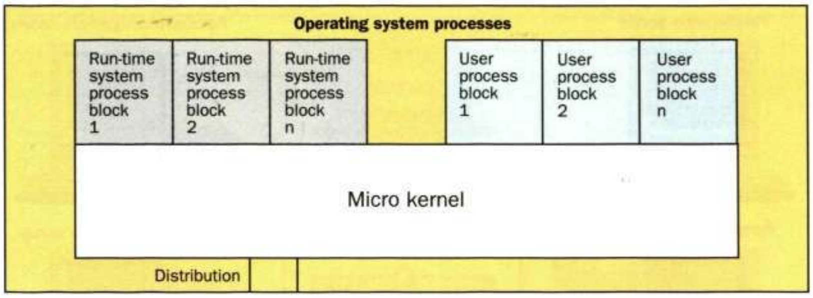

Operating system processes
Run-time system process block 1
Run-time system process block 2
Run-time system process block n
User process block 1
User process block 2
User process block n
Micro kernel
Distribution

Figure D1
The AXE VM architecture is based on a kernel that schedules job and manages distribution and timers. User applications and the run-time system may be run on top of the kernel.

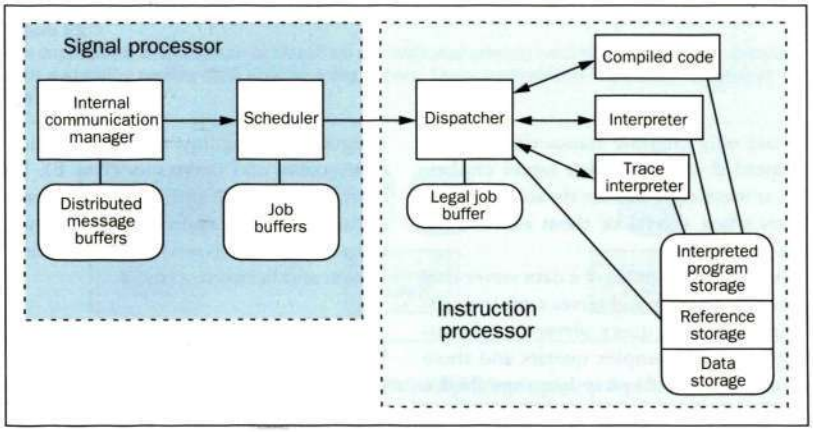


Figure D2
The kernel is divided into a signal processor and instruction processor. The instruction processor can alternate between compiled C++ code, compiled PLEX code and different interpreters of ASA. PLEX blocks can alternate dynamically between compiled code and interpreted code.

Box D AXE VM

The AXE VM emulates the software architecture of AXE systems. The basic parts of the architecture are blocks and signals (that is, messages). Blocks are modules that can only communicate through signals. They contain their own data and can only share data with other blocks by means of signals. The software is assigned the same properties as application-specific integrated circuits (ASIC), hardware boards and systems. Aside from one well-defined interface, there is no way to communicate with the blocks.

The AXE VM contains many characteristics that would normally be found in an operating system (Figure D1). Its design closely resembles a microkernel-based operating system, in which the microkernel executes signals, handles messages between blocks, and manages timers and servers.

Operating-system-like features – such as access to the file system, restart processing, dump handling, external interfaces, man-machine interfaces, and memory management – are added to the system as blocks. Since some blocks require access to the real operating system, they are programmed in C++. The microkernel treats these blocks the same as it treats user blocks.

For the most part, AXE switches are programmed in PLEX, which is rich in real-time features and very powerful in building real-time, high-performance systems with requirements for high reliability. The AXE VM provides extensive support for executing PLEX blocks.

Open-system environments must also be able to communicate with other environments. Thus, the AXE VM contains a set of schemes for communicating with other operating system processes. These include sockets, shared memory, distributed-shared memory (currentiy, SCI is supported), TCP/IP, and Ethernet.

The microkernel receives signals from other AXE VMs within the same configuration domain. As it executes these signals it generates new ones. Signals may be executed (Figure D2) in one of three different ways (patent pending).

- For common signals, the fastest way is to execute compiled code – code that was compiled by a PLEX or C++ compiler.
- PLEX blocks may also be executed in interpreted mode. This is the case for debugging and for uncommon signals – to avoid thrashing the instruction cache.
- Blocks may also be executed with a trace interpreter, which is used for any kind of on-line tracing and debugging – even in real customer systems.

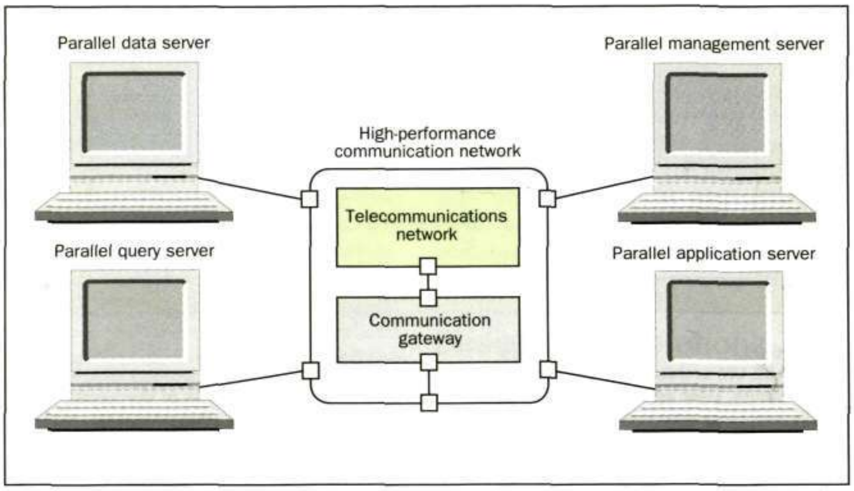

Parallel data server
Parallel management server
High-performance communication network
Telecommunications network
Communication gateway
Parallel query server
Parallel application server

Figure 2
System architecture in which the NDB Cluster reliably stores data for the system. This is achieved by providing a very efficient communication network.

lost and only ongoing transactions are interrupted if an application server crashes. Also, it is easier to balance the load between servers when several of them execute the same application.

The database consists of a data server that stores data reliably and serves simple application queries. A query server may be installed to serve complex queries and those that use structured query language (SQL), open database connectivity (ODBC), and other standards. Query servers request data from data servers and respond to applications that have requested data (Figure 2).

The communication procedure between the servers of a system strongly influences database performance. Communication media that facilitate a shared-memory model facilitate very efficient communication[2,3]. To send a message, the sending party (machine) writes the message into a message buffer in the receiving machine – which means that the operating system is not involved in communication. Checksums are used to guarantee communication. Communication failure is reported via UNIX signals to the AXE VM, which retransmits the message on a different communication path (that is, it writes the message in another message buffer). In terms of time, the cost of sending a short message is less than 10 µs.

Some applications require data servers to store data redundantly, within the system and between systems (network redundancy). Replication, which is handled by the data server, is transparent to the application and safeguards data against catastrophes such as earthquakes and hurricanes (Box E). Data servers support the replication of data within the system or between systems. They also support simultaneous replication within the system and between systems.

Functionality in the NDB Cluster

NDB Cluster usage model

A user defines a set of tables, where each table consists of a set of attributes. In this sense, the model resembles a relational database. However, compared with an ordinary relational database, other features have also been added. For instance, the tables always have a unique key, which is either a primary key supplied by the user, or a tuple key that is generated when the tuple is inserted. Secondary keys may also be defined. Table attributes are fixed in size; consist of an array of attributes that are fixed in size; or consist of an array of attributes that are variable in size. Thus, files may be stored in the table as attributes.

The fixed attribute sizes (in bits) are: 1, 2, 4, 8, 16, 32, 64 and 128.

Research has shown that dynamic attributes can be supported and that they can be used to support unstructured data[4,5]. Moreover, dynamic attributes may be stored.

The NDB Cluster supports main-memory data and disk-based data. The choice of data storage is determined on a per-attribute basis when the tables are created. This means that

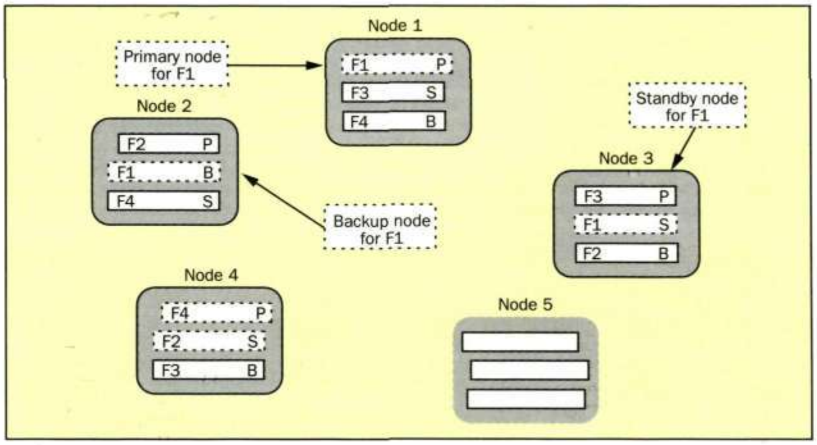


Figure E1
The replication architecture is based on a traditional primary and backup replica combined with a standby replica that acts as a log server. Thus, even double faults are handled at low cost.

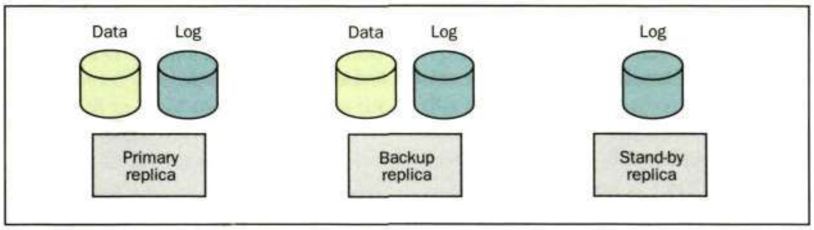


Figure E2
Example of distributed replicas. Node 1 contains the primary replica for F1; Node F2 contains the backup replica and Node 3 contains the standby replica. Node 5 is prepared to create new replicas in the event that any node crashes.

Box E
Replication structure

Merely safeguarding against a single point of failure will not achieve very high reliability in multinode systems – given requirements which specify that annual downtime may not exceed 30 seconds per year, the risk of losing another node during repair is far too great. Writing data to disk may not be part of a transaction; instead, data must be written to main memory in at least two processor nodes. If one node fails, then several other nodes run with only one replica. Ordinarily, to maintain real-time and reliability characteristics during repairs, another replica must be added. However, in terms of memory and processor load, most customers feel that the cost of having three complete replicas outweighs the benefit.

In the NDB Cluster, the solution is to add a special replica type called the standby replica. The standby replica solely contains a log of the transactions that changed the contents of the database. It does not contain a replica of any data and is therefore not useful for reading (Figure E1).

If the primary replica and all backup replicas fail, the log of transactions in the standby replica can be used for creating a new primary replica – the log is independent (a) of the location in which it was created and (b) of the replica that created it. Therefore, unless all replicas fail within a few milliseconds of each other, no data is lost.

Ordinarily, when the primary and backup replicas fail, the system must be shut down and restarted. With standby replicas, this is seldom necessary. Fragments of the database may be temporarily inaccessible, but the database system will continue to operate, quickly restoring them.

Figure E2 shows a table with one primary replica, one backup replica, and one standby replica. The table is split into four fragments, which have been evenly distributed among four nodes. A fifth node, which is used as a hot spare, is used for creating new replicas when another node fails. Having a spare node greatly limits the effects that a failed node may have on a system.

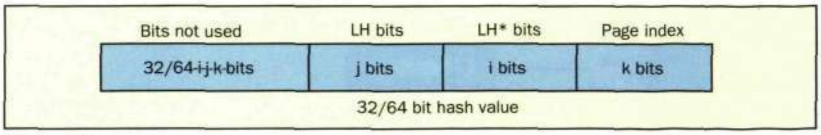


Figure F1
The LH^3 contains three parts that are used in the hash value. The LH*bits are used for finding nodes that contain replicas; the LH bits are used for finding the correct index page; and the page index is used for finding the container in which to start the search.

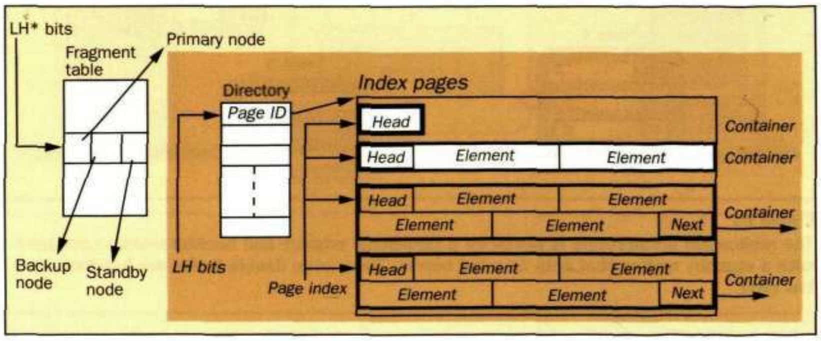


Figure F2
Visualization of how hash value bits are used to find the index element.

Box F LH^3

A scaleable database must include an index that distributes tuples effectively. In addition, if an index system expands over a long period, the distribution scheme should permit dynamic changes. Indexes of this kind have been proposed. The NDB Cluster, which uses a patent-pending extension of the proposed scheme, performs hashing in three steps[1].

A hash function is applied to the key. This function creates a hash value that is usually 32 bits in length. The bits are split into four parts:

a)The most significant bits are not used (Figure F1).

b)Several LH bits are used locally in the node as an index of the directory (Figures F1 and F2). The linear hashing algorithm is applied when j bits are used for the first time. If the value of these bits falls below a given threshold, then j+1 bits are used instead.

c)Using the same linear hashing algorithm, the LH* bits are used to find the fragment identity of the tuple (Figures F1 and F2).

d)The remaining bits are used to find a container in the page referred to by the directory entry that the LH bits found (Figures F1 and F2).

The size of the LH* bits is not fixed. Fragments may be split one at a time, provided more hash-value bits exist – each time a fragment is split, one LH bit is lost. However, the number of page-index bits is fixed when the table is defined.

The fragment identity is used for looking up node references to fragment replicas. Thus, an update can easily find every fragment replica that needs to be updated.

The LH^3 algorithm has many benefits. It provides a distributed index that is integrated into a highly efficient local index.

a table may consist in part of attributes resident in main memory and of attributes on disk.

Since the NDB Cluster is distributed, table fragmentation must be defined either by the user or automatically (Box F).

Automatic reconfiguration

When a processor node fails, several fragment replicas are lost. However, once the failure is discovered by the heartbeat function, the database management system (DBMS) automatically creates new replicas in working

Box G
Recovery protocols

The NDB Cluster contains two recovery protocols. The first is used when a replica is available in the system. The second is used for creating the replica of a fragment when no primary or backup replicas are available; it is also used for system recovery.

When a primary replica is available, the protocol uses it for creating new fragment replicas. As it does so, the protocol must also handle the concurrent execution of transactions. Basically, each transaction will attempt to write into the new replica – insertions may be made into the new replica at any time. However, the replica may not be updated, and nothing may be deleted from it, until it has received a tuple.

The copy process performs a mini-transaction each time a tuple is moved: the tuple is locked and moved to the new replica; when the tuple has been fully replicated the lock is released.

The recovery protocol for restoring a fragment in the absence of a primary replica consists of three steps that are based on information produced by a checkpoint:

- First, pages are read into main memory (this step does not apply to disk-based data).
- Next, the local undo log is executed, which restores to pages the contents they held when the checkpoint was added.
- Finally, every transaction that may have been lost is reenacted, starting from some point in time before the checkpoint.

processor nodes. Thus, the system is said to have failed gracefully. As long as the system contains live processor nodes, it continues to operate (Box G). If data consistency is in any way threatened – because too many processor nodes fail in rapid succession – then the system is restarted from an archive copy.

If a processor node is added, the tables can be refragmented using the LH^3 algorithm. Therefore, the system continues to operate optimally as it grows. Refragmentation may be performed on-line.

Local replication

Local replication ensures that only a series of failed processor nodes can cause the system to restart. Numerous processor-node failures may make some data inaccessible until the system has recovered the processor nodes.

When an application creates a table in the database, it must decide which and how many replicas the table should contain. The system then tries to maintain this number of copies. The number of replicas may be changed on-line.

Since every transaction is written in several processor nodes, it is not necessary to flush the log to disk during the commit phase. However, if errors cause a processor node to fail, the logs will be written to disk during the commit phase. For the system to crash, every node that contains the replica of a given fragment must fail within the space of about 100 ms.

Three types of replica may be used. A primary replica always exists, along with any number of backup replicas. There is also a special replica type called the standby replica. This replica, which is a log server, logs every update. The standby replica does not contain actual data from the table, but it is useful when the primary and other backup replicas fail. The standby replica ensures that no committed transactions are lost, and that the failure of primary and backup replicas does not cause a system to restart. Since it only writes a log, the standby replica provides extra reliability at a modest cost.

Network redundancy

If local replication is insufficient, the NDB Cluster can also support global replication – which may be achieved in parallel with local replication. A transaction is first executed by the primary system. It is then sent to the backup system, which also executes it. If response to the application is sent after the backup system has finished, the transaction is called 2-safe. If the response is sent before the backup system has finished, the transaction is called 1-safe. 1-safe transactions might be lost when a primary system fails. In either case, the transactions must be executed in the same order in each system. Transactions that merely read data are executed in the primary system only.

The NDB Cluster supports 1-safe and 2-safe transactions. Applications set this status at the attribute level. For example, location updates in a home location register (HLR) probably do not require 2-safe transactions; therefore, they are assigned the 1-safe attribute. However, when a new service is activated, a 2-safe transaction should probably be executed. If any 2-safe attribute is declared for a write transaction, then the entire transaction is classified 2-safe; otherwise it is executed as a 1-safe transaction.

Transaction support

A new, efficient two-phase commit protocol has been developed that is especially suited for replicated databases (Box H). Replication ensures that transactions need not use disks during the commit phase, thereby fulfilling real-time and high-performance requirements.

Dirty read, dirty write

Many telecommunications databases have introduced dirty-read and dirty-write transactions, in order to cut the cost of transactions – at the expense of consistency. In this implementation, nothing is gained from a dirty read, which is implemented as a simple read transaction. In simple read transactions, only one record is read and the commit phase may be eliminated.

A dirty write means that the commit phase is removed from updates. This does not increase performance significantly, since the greatest expense occurs in the phase during which objects are written. Removing the commit phase from updates does shorten delay, however. It also increases concurrent processing, since the locks can be released immediately after an object has been writ-

Box H
Two-phase commit protocol

One of the most crucial protocols in a parallel-data server is the two-phase commit protocol. This protocol is used to ensure that updates are atomic (that is, every update is either totally complete or it is not carried out at all) and durable (that is, the updates can survive any type of crash). This also applies to updates of several tables and several replicas of each table. Some of the features of a newly designed two-phase commit protocol (patent pending) are listed below:

- Optimized for several replicas of each updated data item.
- Low communication overhead – even when a transaction contains several updates.
- Durability is preserved by writing to the main memory of several nodes instead of writing to disk.
- Short response time — even when a transaction contains several updates.
- Easy integration with two levels of replication – to support network redundancy.

The basic protocol combines a normal two-phase-commit protocol with a linear-commit protocol for each fragment involved (Figure H1).

During the prepare phase, the protocol starts in the primary replica. Lock contention is handled in the primary replica. Nonetheless, to guarantee that simple read transactions (transactions that solely read a single tuple) may be sent to the backup node, the backup replicas also lock the data.

The commit phase ends in the primary replica, since the transaction is not safely committed until every replica in a fragment has been informed of the commit decision. Thereafter the primary replica releases its locks. The backup replicas release their locks in the complete phase, or when a new transaction is started on the same data item.

The real power of the protocol comes into play when the transaction involves many updates in several fragments. Each fragment executes the linear-commit protocol in parallel, which greatly reduces response time. For large transactions, performing the commit phase in series increases response time beyond acceptable limits.

The power of the protocol is seen even more clearly when two levels of replication are added to support network redundancy. The backup system is synchronized with the primary system. Communication between the systems is performed over a long-distance line, which rules out using a normal two-phase commit protocol between the systems. Instead, an optimized version is used with the backup system, which must either accept the transaction or die.

To achieve scaleable network redundancy, more than one channel must exist for sending transactions between systems. Primary replicas are assigned a channel to the master replica in the backup system (usually a primary replica but sometimes a standby replica). During the commit phase, the primary replica in the primary system contacts the master replica in the backup system, which means the primary system has already decided to commit. Thus, the backup system must agree. Otherwise it is restarted, since the information it contains is inconsistent with the primary system.

When a transaction involves several fragments, a transaction coordinator is used in the backup system. To achieve scaleability, the transaction coordinator is logically chosen in the primary system. The logical identity of the node is mapped onto a physical node identity through a table that is available in every node of the backup system (Figure H2).

The protocol may be optimized for a simple write transaction by transferring the role of the commit coordinator from the transaction coordinator to the last replica in the chain. This arrangement does away with communication between the last replica and the transaction coordinator – which is a reduction by two messages.

ten. In summary, the dirty-write method may be advantageous to hot-spots where the loss of an update is not costly. A typical example is statistical information.

To ensure that locking conflicts do not occur, a data model may be applied to simple read and simple write transactions. Similarly, if the data is current (has been updated), then simple read queries could be used to read old committed data. This yields desired real-time behavior without introducing unusual recovery situations. These methods should be avoided if inconsistent data can disrupt the system (which it very often does).

Prepare-to-commit function

The support of a transaction coordinator outside the NDB Cluster requires the support of a prepare-to-commit function. Should the transaction coordinator fail, the data is locked or frozen until the transaction coordinator has recovered; therefore, it needs to be very reliable.

On-line schema changes

A very reliable database must be able to handle on-line schema changes. Schema changes may be integrated into network redundancy and crash recovery. They may also be in-

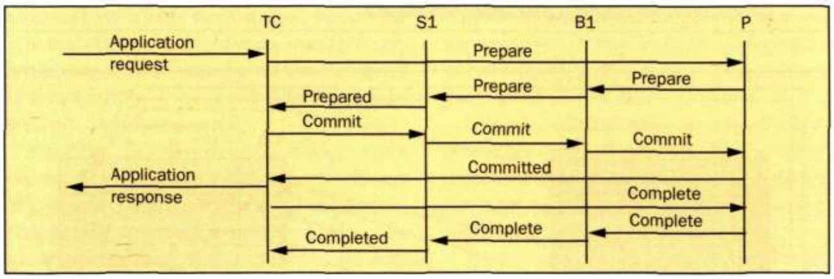


Figure H1
Message flow in the NDB Cluster for a simple write transaction. The commit action is not safe until the primary replica has the commit message.

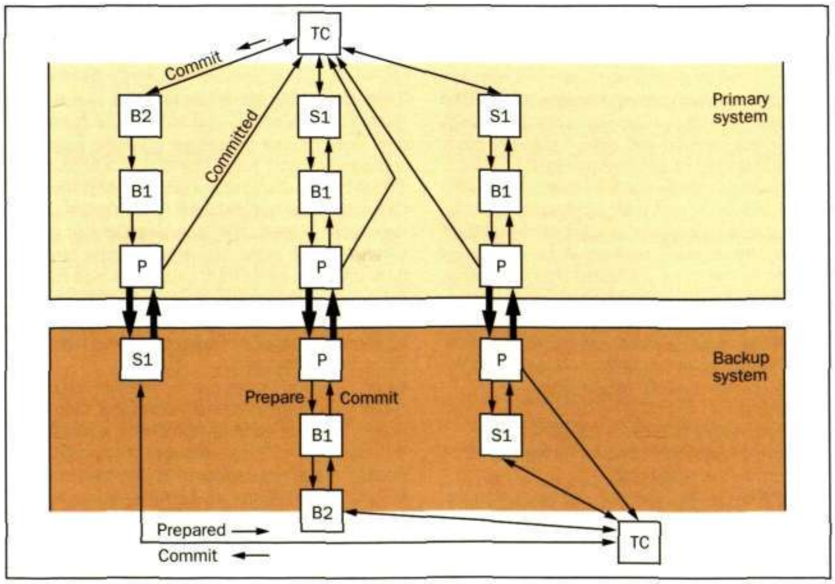


Figure H2
Each write action uses an internal linear-commit scheme between replicas. The TC performs a normal two-phase commit. When the primary node hears the commit message it contacts the backup system.

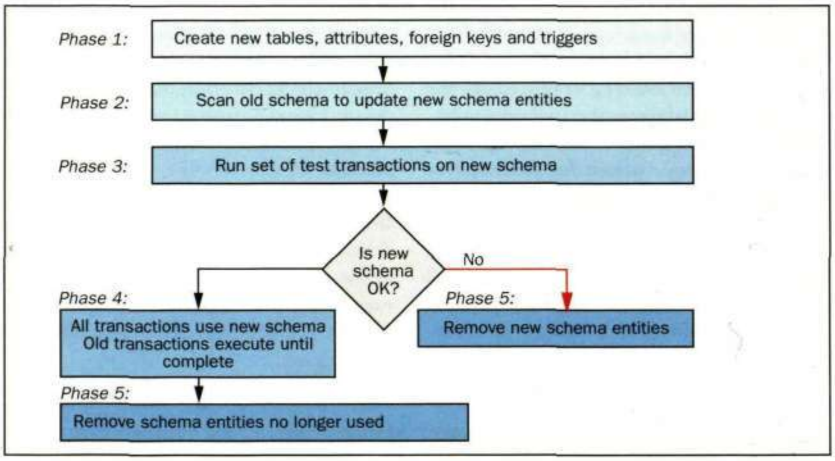
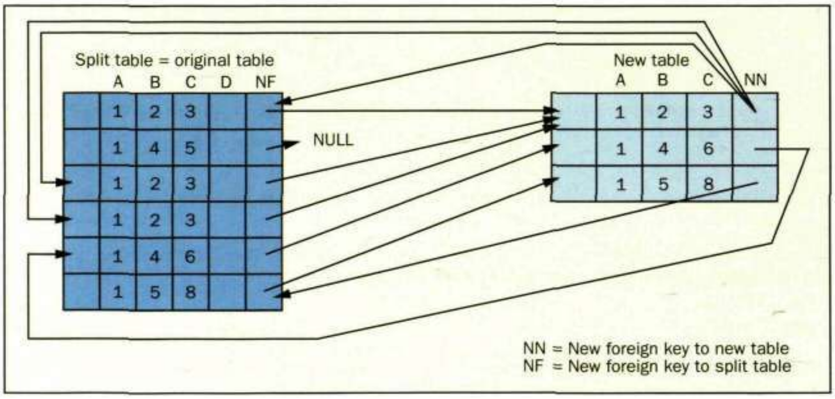

Phase 1:
Create new tables, attributes, foreign keys and triggers
Phase 2:
Scan old schema to update new schema entities
Phase 3:
Run set of test transactions on new schema
Is new schema OK?
No
Phase 4:
All transactions use new schema
Old transactions execute until complete
Phase 5:
Remove new schema entities
Phase 5:
Remove schema entities no longer used

Figure I1
A new table is created from parts of the original table. During the schema change, foreign keys ensure that both tables are up-to-date. After the schema change, unnecessary parts may be discarded.


Split table = original table
NULL
New table
NN = New foreign key to new table
NF = New foreign key to split table

	A	B	C	D	NF
	1	2	3		
	1	4	5		
	1	2	3		
	1	2	3		
	1	4	6		
	1	5	8		

A	B	C	NN
1	2	3	
1	4	6	
1	5	8	

Figure I2
Complex schema changes require an algorithm of simple schema change, normal transaction and test transactions. It might even be necessary to change application code.

Box I
Example of an on-line schema change

A highly reliable telecommunications database must be able to support changes in the data structure. That is, it must be able to add and drop attributes, tables, and triggers. It must also support more complicated actions, including various ways of splitting and merging tables. A complex schema change that involves copying data between old and new formats is called a long-lived transaction. Transactions of this kind are performed in several phases. A reliable database can handle changes even when nodes and systems crash. For handling long-lived transactions, the NDB Cluster introduces a SAGA table, which stores system information on how to undo and/or redo changes that are part of a complex schema change. The table is also replicated onto the backup system, which means that complex schema changes are integrated into network redundancy. Slight differences may exist in primary and backup system schemata, since they do not always require the same types of indexes and triggers.

A simple schema change solely affects the schema, not data. A complex schema change affects the schema and data. Splitting a table is one example of a complex schema change. For instance, let us assume that a table with four attributes is split into two tables (Figure I1). The schema change was brought on by the result of an optimization, performed to give quick access to the three attributes in a new table. The original table now has two keys: its own, and a foreign key – the key to the new table. The new table contains the three remaining attributes. When the optimization is no longer needed, the tables may be merged again.

Summary:

- Create a new table with its attributes.
- In the original table, add a foreign key that refers to the new table.
- Add a key in the new table that refers to the original table.
- Add a trigger that updates the new table when the original table is updated.
- Introduce a trigger that updates the original table when the new table is updated.

The foreign keys are initially set to null. If the tuple is a copy, the foreign keys are also set. Thus, by looking at the foreign key it is easy to determine whether or not to execute the trigger. The process of copying from the original table to the new table starts as soon as the first schema changes are properly installed. Once the copying process is complete, the new table is ready for immediate use. If possible, execute some test transactions on the new schema. Aside from cases in which there is no mapping from the new format to the old format, the test transactions may be executed concurrently with transactions that use the old schema (Figure I2).

Example: Total wages are recorded in the new table. The original table contained monthly wages and the number of months worked. In this case there is no backward mapping.

Finally, if the new schema is deemed ready for use, the triggers may be removed as soon as every transaction associated with the old schema has been completed. The two attributes from the original table are dropped. Whether the foreign keys are dropped or not depends on the user and which keys are needed. If the foreign keys are not needed they may be dropped.

tegrated along with software changes, allowing application software to keep pace with schema changes.

Examples of schema changes that may be performed on-line are (Box I):

- add/drop tables;
- add/drop attributes;
- add/drop views;
- add/drop secondary indexes;
- change attributes (new attributes are a function of the old schema);
- split table horizontally and vertically;
- merge tables horizontally and vertically;
- add/drop referential constraints;
- add/drop attribute constraints.

If it is logically possible, each change is performed as a soft schema change. That is, transactions which were started before the schema change may execute concurrently with transactions started after the change. If it is not safe to do so, a soft schema change will not be implemented.

Example: Before a schema change takes place, two attributes are stored: Hours Worked and Hourly Wage. After the schema change, the attributes are merged into a single attribute: Total Wage. However, because there is no way of deriving the number of hours worked from this new attribute, the old and new transactions cannot execute concurrently, which means that a soft schema change cannot be implemented.

In a hard schema change, every transaction that was started before the change must be completed before any new transactions affected by the schema change may be started.

NDB Cluster interface

The interface to the NDB Cluster is designed from a set of C++ classes. The methods in these classes enable queries to be defined for tuples. Initially, only queries to a single table are supported, since the need for joining queries in real-time telecom applications is very limited. These queries can read and write attributes or parts of attributes. Using an interpreted language, future releases will allow general programs to execute in the database kernel.

Other database products

Many different kinds of database have been developed, most of which are standard products with a focus on business applications and serving a very wide range of applications. They normally use disks as safe media. Thus every transaction must be flushed to disk before committing, which limits the number of write transactions that can be served by these products. Moreover, they have not been designed for use in real time, which means they cannot achieve response times between 1 and 5 ms. They also give a fairly long delay in write transactions. For this reason, the databases do not meet every requirement of very demanding telecommunications applications.

A large set of databases has been developed for real-time applications – predominantly main-memory databases. These usually come bundled with a real-time system. In contrast, the NDB Cluster can be used with main memory and disk data; it may be used in an open environment with other real-time systems and with other non-real-time systems; it may even be configured for use as a desktop database in a personal computer (PC).

ClustRa, which is developed by TeleNor, is another database product for telecommunications applications. The NDB Cluster contains many innovations that extend the support of such applications.

Some commercial database products support network redundancy, but few of them do so in a scaleable manner. They have only one channel over which all log messages must pass in the order in which they were executed. To enable scaleability, the NDB Cluster uses parallel log channels between database systems. In addition, it is flexible in its support of executing 1-safe and 2-safe transactions with network redundancy.

Many database products support only 1-safe replicas, whereas the NDB Cluster supports 2-safe replicas within the system as well as 1-safe and 2-safe replicas between systems.

The NDB Cluster introduces several innovations, including

- transaction support (patent pending);
- replication support (patent pending);
- new data structure for indexes of tuple keys (patent pending);
- new data structures for tuple storage – which combine efficient support of small, fixed tuples while providing efficient support of very large tuples;
- support of network redundancy;
- new index for hypertext transfer protocol (HTTP) addresses (patent pending);
- support of advanced schema changes – which has been integrated into crash recovery and network redundancy (patent pending);
- overload handling and load regulation.

Box J
HTTP tree

Names that are organized alphabetically in a list do not differ significantly from their neighbors. Thus, data structures can be compressed to less than one-fifth their original size, which greatly speeds up processing – the data structures occupy less memory and yield a better hit rate in cache memory. The HTTP tree capitalizes on this technique.

The HTTP tree structure is based on a normal B+tree structure. However, instead of storing the full name of each entry, only differences from the previous entry are stored. The index maps names to a tuple key (or to a local key, if one is used as a tuple-key index), which acts as a file handler when the data server is used to implement scaleable file systems. Example:

START=http://www.ericsson.se/ndb/compbtree/description.html

2)http://www.ericsson.se/ndb/compbtree/nodedescription.html

3)http://www.ericsson.se/ndb/compbtree/system_arch.pdf

4)http://www.ericsson.se/ndb/compbtree/system_arch.ps

To describe this with the compressed algorithm we need a start value (START). Next 2), 3), and 4) are described, as follows:

2) Equality at end (1 bit), 4 bit minus and plus (1 bit), difference information exist (1 bit), extension = .html (5 bits), -0 (4 bits), +4 (4 bits), Last 11 equal (8 bits), "node" (4 bytes)

3) Not equal at end (1 bit), 4 bit minus and plus (1 bit), difference information exist (1 bit), extension = .pdf (5 bits), -15 (4 bits), +11 (4 bits), "system_arch" (11 bytes)

4) Not equal at end (1 bit), 4 bit minus and plus (1 bit), difference information do not exist, extension = .ps (5 bits)

Thanks to compression algorithms, 2), 3), and 4) require only 21 bytes instead of the original 158 bytes. Obviously, compression-related gains depend on several factors, such as naming style. Still, in most cases, the compression techniques yield substantial gains.

The basic structure is a B+tree, some of whose pages are organized into a hierarchical structure. An additional B+tree structure is used within the pages to help find a key within its pages. Here, nodes are used, since they are smaller (128 bytes) than pages. Finally, using the compression algorithm, a sequential search is performed. The index is also very effective for indexing telephone numbers.

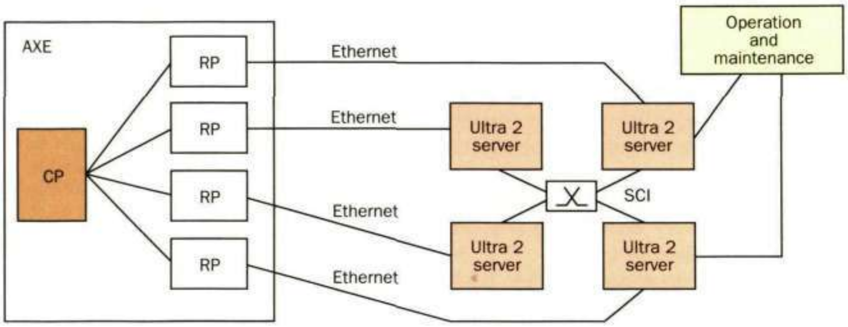


Figure 3
Architecture of a number-portability prototype. The central processor asks the NDB Cluster for number translation through any regional processor connected to the NDB Cluster.

Number portability

The first application to use the NDB Cluster is a prototype application that supports number portability. The implementation, which may be integrated into existing telephone exchanges, may also be used in networks that use INAP, TCP/IP, or both.

In its first prototype application, the NDB Cluster functions as an external database to AXE. It is accessed as part of the call setup and from operation and maintenance systems using the C++ interface. The main benefit of the NDB Cluster is that it provides data to AXE and to the operation and maintenance system in real time.

Many different network and server configurations may provide a number-portability service. The NDB Cluster is flexible in meeting demands for very high reliability and high availability at low cost (Figure 3). To reduce cost, two servers may be used instead of four. To further reduce cost, less expensive servers may be used, although this lowers performance and increases response time.

Predefined messages in the call-setup application access the database for the translation of a telephone number. The messages are incorporated into the call-setup application. Prototypes have shown that this may be achieved at a low cost to the CP and database servers. The operation and maintenance system have general, real-time access to all data in the database through the C++ interface.

The short response time is achieved through two-processor servers, which convert Solaris processes into real-time processes. One processor handles operating system activities and non-real-time activities; the other processor simply waits for input, which means it can respond immediately. In this way, a standard operating system can be used to guarantee real-time response.

Single-processor servers may also be used, but the real-time properties are not as good. The NDB Cluster and AXE VM may be configured to adjust automatically to the present configuration, whether for single-processor servers, multiprocessor servers, single-server systems or for multi-server systems. The NDB Cluster automatically adapts its configuration to accommodate available servers – even when new servers are added while the system is operating.

Web cache server

Some sets of services on the Internet require scaleable access to data. These include route servers, domain name service (DNS) servers, Web servers, e-mail servers, and file servers.

Web cache servers store a large volume of frequently requested Web pages. In large networks, several Web cache servers work together. If one server fails to locate a Web page, another server might. Web cache servers should also function as replication servers; that is, they should be able to mirror the contents of Web pages at other servers. This feature is useful when Web pages contain dynamic data, since Web pages of this kind cannot be cached.

An example network architecture is shown in Figure 4. Since all data resides in the NDB Cluster, the scheduling servers may call upon any cache server to perform work. To avoid the costly setup of TCP/IP connections, the scheduling servers might use a permanent connection.

Cache servers must contain an index of the Web pages they store. They should also contain an index of the Web pages stored in cache servers at a lower level or at neighboring servers in the hierarchy. Furthermore, they should keep statistics on the Web pages they store and on Web pages that might be stored – pages that are accessed frequently.

Since cache servers at higher levels in the hierarchy are accessed by many users, the storage must be scaleable and quickly accessible. The NDB Cluster fulfills these requirements thanks to its automatic table-fragmenting functionality. Also, the transport of data over scaleable coherent interfaces (SCI) with very high bandwidth enables efficient communication between server processors.

The NDB Cluster contains a very efficient index for HTTP addresses. Thus, an index of literally millions of Web pages may be stored in main memory on one computer together with brief statistics and a reference to where the Web pages are stored. For fast access, the statistical attributes and the index reside in main memory, whereas the actual Web pages may reside on disk. Since a Web page may reside as an attribute in the same record as its descriptive attribute, the reference to disk is direct and need not pass through extra directory structures.

To avoid unnecessary moves within the Web cache server, the NDB Cluster will be equipped to send Web pages directly to requesting entities over a TCP/IP link.

This solution delivers scaleability in the application and the database parts of the server, which is critical to the design of future Internet servers. The NDB Cluster also contains functionality for balancing the load between database servers – thanks to replicas of data and the NDB Cluster's ability to reconfigure data distribution automatically.

When used as a Web cache server, the NDB Cluster functions more or less as a distributed file system. Indeed, the automatic fragmentation of data and plug-and-play scaleability features are nicely suited to this type of application. The NDB Cluster is designed for high-performance systems. By adding specific parts for handling large flows of data in these applications, it performs equally well as specialized file servers. The patent-pending new-index feature enables very rapid lookup while reserving a large part of main memory as a buffer for the most frequently requested Web pages. High-performance communication between servers is another essential part of the system characteristics.

Conclusion

The NDB Cluster is a scaleable solution to database applications in telecommunications networks with requirements for very high performance, very high reliability, and real-time performance.

The NDB Cluster introduces several innovations, including transaction support (patent pending); replication support (patent pending); new data structure for indexes of tuple keys (patent pending); new data structures for tuple storage; support of network redundancy; new index for HTTP addresses (patent pending); support of advanced schema changes (patent pending); and overload handling and load regulation.

It is the first distributed database designed on the assumption that distribution and the support of replication are cheap. Many other network databases are specialized in that they support main-memory data. The NDB Cluster supports main-memory and disk data; and it is a general-purpose data server for telecommunications applications that can be used by applications written in other languages, provided they use any of the communication mechanisms that the NDB Cluster supports.

The NDB Cluster is based on true distribution transparency, which greatly facilitates the writing of applications. In its prototype application, the NDB Cluster functions as an external database to AXE.

References

1 RONM97a. M. Ronström, Design and Modeling of a Parallel Data Server for Telecommunications Applications, Ph.D. dissertation to be published in 1998

2 SCI. The Dolphin SCI Interconnect, White Paper, Feb. 1996

3 Memory Channel. R.B. Gillet, Memory Channel Network for PCI, IEEE Micro Vol 16, no. 1, p. 12-19, Feb. 1996

4 RONM97. M. Ronström, Report on the AXE Virtual Machine, UAB/B/U-97:097

5 Lorel S. Abiteboul, D. Quass, J. McHugh, J. Widom, J. Wiener, The LOREL Query Language for Semi-structured Data, Technical Report from Stanford University

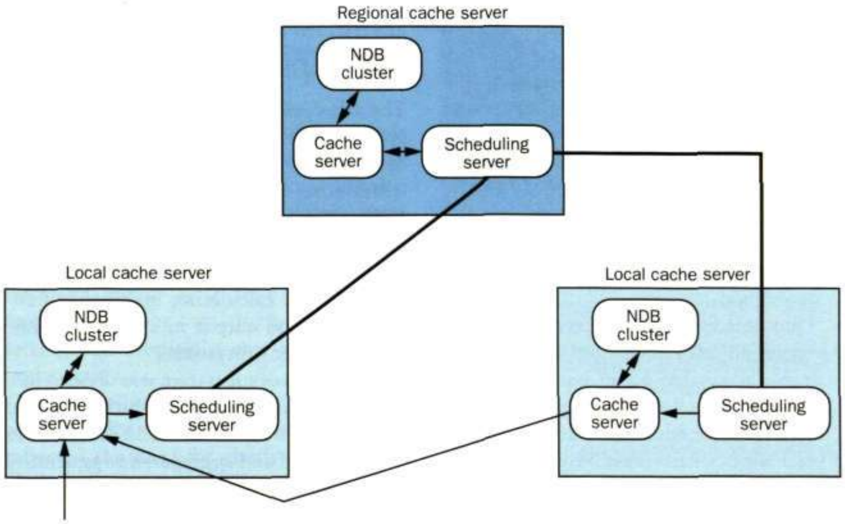


Figure 4
Network architecture of a Web cache server. Most communication between systems is handled by the scheduling server, which uses permanent connections.

Environment, for better or worse (Part 2)

Mats-Olov Hedblom

People and industries the world over are acknowledging that we can no longer take the environment for granted. The resolutions that were passed at the Rio Conference in 1992 charge every world society with the task of attaining a proper balance between their social, economic and environmental responsibilities. Partly in response to that conference, world industries have agreed to develop a suite of tools for environmental management – the ISO 14000 series of standards.

Corporations who have identified the environmental dimension as the most important aspect of good global citizenship are currently seeking tools that will help them to condense this dimension into strategic scenarios of the future. One such tool is the life-cycle assessment.

The author describes how Ericsson, as a member of the IT industry, is in the process of establishing a sound environmental platform on which to base their operations, now and in the future.

Part 1 of this three-part series of articles deals with various international perspectives of the environmental issue. Part 2 describes the life-cycle assessment in depth. Part 3 demonstrates how Ericsson can apply the findings from the life-cycle assessment, designing for the environment and labeling products according to the emerging ISO 14025.

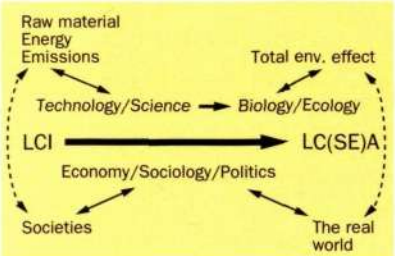
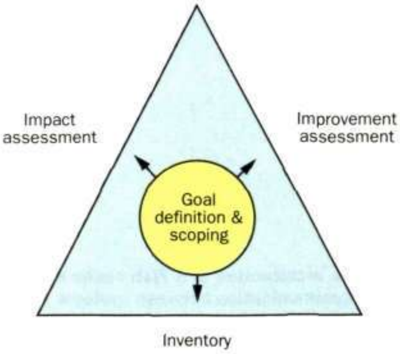


Figure 1
Cooperation between the engineering and ecological sciences can bridge the gap in knowledge of the cause and effect of environmental threats.

Figure 2
SETAC technical framework for the life-cycle assessment.



Life-cycle assessment – industry's strategic tool for survival

The purpose of this article is to give insight into emerging methods that compose the life-cycle assessment (LCA), which will be used for describing interactions between industrial and societal activities and the immensely complex environment that forms the basis of sustained wealth and human prosperity.

An article recently published in Nature[1] on the general role of research and science stated that:

- "The scientific enterprise is full of experts on specialist areas but woefully short of people with a unified world view."
- "Scientists are trapped in their own specialisms, leaving others, often poorly qualified, to represent to the public the larger architecture and interconnections of modern scientific theories."
- "Scientific education has become so specialized that scientific literacy is little more advanced among scientists than it is among non-scientists."
- "Undergraduates who have completed courses on cell biology and evolution are unable to discuss broad issues in evolutionary theory, let alone Earth history or cosmogony, in any greater depth than can their non-scientist peers. Physics students don't know how a protein differs from a nucleic acid; chemistry students don't know the age of Earth; geology students cannot give a simple account of metabolism or say why the sky is blue."

With tools developed for the LCA, we are finally able to obtain scientifically defensible descriptions of good and bad interrelationships between industrial product systems, the human society, and the external environment. The most important part of this development is the link between mass- and energy-engineering sciences and sciences (biology, ecology) that deal with the environmental indicators that most often reflect the negative impact of industry or society's activities (Figure 1).

The absence of a life-cycle assessment has enabled anyone to speak out on subjective grounds as "environmental specialists." While stakeholders in branches of industry that heavily burden the environment (automobile manufacturers, forestry, oil companies) fear, express concern over, or even fight ISO 14000 standardization, the information technology (IT) industry, with its inherently low-impact profile, only stands to gain from it.

This article presents the dynamics of LCA-related developments within the International Organization for Standardization (ISO). It also introduces elements of the most advanced form of the LCA, called the life-cycle stressor-effects assessment (LCSEA), which is certain to be practiced by environmentally progressive companies throughout the world. Ericsson is the first telecommunications corporation to apply the LCSEA.

Overview of the life-cycle assessment – historic development

The life-cycle assessment was first developed as a system-oriented tool for tracking material and energy flows in industrial systems during the energy shortage of the early 1970s. Having emerged from the science of throughput analysis, the life-cycle assessment was essentially an inventory exercise that involved calculating material and energy input and output in a defined industrial system or subsystem.

The system or subsystem was divided into discrete unit operations for which input and output data were collected. The data was then gathered and aggregated to generate a sum total of resources used, energy consumed and of environmental releases by species. Next, the data was normalized to a specified functional unit of measure. Aggregation, allocation and normalization were typically based on standard assumptions about mass and energy balance. Procedures and data were reported in mass and energy units.

Several computer models were developed in Europe and in the US to perform the iterative calculations needed to conduct life-cycle inventories (LCI). In time, these models were expanded, integrating data on a broad range of discharges into the air, water and ground, and accommodating the growing volume of data for basic upstream processes. By the late 1980s, a variety of life-cycle models were in use around the world.

Life-cycle methods expanded as government agencies, industry, institutions and non-governmental organizations began to use them for more than simply tracking inventories. For instance, the life-cycle assessment was thought capable of facilitating comparisons of the environmental impact associated with alternative production technologies and competing materials, as well as of communicating the environmental performance profile of products sold on the market.

The SETAC LCA framework

In an effort to harmonize methods and the resulting databases of various models, the Society for Environmental Toxicology and Chemistry (SETAC) set out in the late 1980s to spearhead the development of a common conceptual technical framework for the life-cycle assessment. The framework was to be hammered out in a series of technical workshops based on the contributions of developers, practitioners, industry users, and stakeholders of the LCA model.

In 1990, SETAC published the proceedings of their first workshop, which was a proposal for a technical framework. Under this framework, the LCA was described as having three separate components: inventory, impact assessment, and improvement assessment (Figure 2). In later workshops, a fourth component – which consisted of goal-setting and scoping – was added to the framework.

The inclusion within the technical framework of a distinct impact-assessment stage reflected the need for clarifying the environmental significance of data collected at the inventory stage and recognized the in-

Box A Terminology

BOD Biological oxygen demand.

Category endpoint Representation of the natural environment, human health, or resources used to designate an impact category.

Category indicator Modeled inventory results that represent a given life-cycle impact within an impact category.

CD Committee draft.

Characterization factor A factor that converts a life-cycle inventory result into a common numerical scale within an impact category for aggregation into indicator results.

Completeness check Process of verifying that sufficient information exists from each phase (inventory analysis, life-cycle impact assessment) to reach a conclusion from interpretation.

Consistency check Process of verifying that interpretation complies with the defined goal and scope before conclusions are reached.

DfE Design for the environment.

DIS Draft international standard.

Environmental issue Input, output (results from the LCI) and environmental indicators (results from the LCIA), that are important to the definition of the goal and scope.

EOL End of life.

EPA Environmental Protection Agency.

EPS Environmental priority strategy – A Swedish LCA tool for designers and engineers.

Evaluation The second step of the life-cycle interpretation, including completeness check, sensitivity check, consistency check, etc.

GIS Geographic information system.

GWP Global warming potential.

HC Hydrocarbon.

Impact category Group of life-cycle impacts that illustrate the connection between certain inventory results and a specific indicator and endpoint.

Indicator A simplification and distillation of complex information intended as a summary description of conditions or trends.

ISO International Organization for Standardization.

IT Information technology.

LCA Life-cycle assessment.

LCI Life-cycle inventory.

LCIA Life-cycle impact assessment.

LCSEA Life-cycle stressor-effects assessment.

Life-cycle impact Representation of environmental change caused by a product system. Note: The life-cycle impact does not indicate actual environmental effects.

Life-cycle interpretation Phase of the life-cycle assessment in which findings of the inventory analysis, of the impact assessment, or of both, are combined in a manner that complies with the defined goal and scope, in order to reach conclusions and recommendations.

Life-cycle inventory result Outcome of a life-cycle inventory analysis which, in crossing the system boundary, represents interaction of the system with the environment.

Measurement endpoints A change in flora, fauna, human health or resources that represents a significant measurable deviation from a defined baseline.

NO_x Nitrous oxide.

POCP Photochemical ozone creation potential.

SAGE Strategic Advisory Group on the Environment.

Sensitivity analysis Systematic procedure for estimating the effects on the outcome of a study of the chosen methods and data.

SETAC Society for Environmental Toxicology and Chemistry.

SO_x Sulfur oxide.

Stressor A specific system input, output or activity that is linked to an observed effect or related group of effects.

Stressor-effects network The interlocking physical, biological and chemical events that connect a specific system input, output or activity (that is, the stressor) to an observed effect or related group of effects.

Uncertainty analysis A systematic procedure for ascertaining and quantifying the uncertainty introduced into the results of a life-cycle inventory, due to the cumulative effects of input uncertainty and data variability. Uncertainty analysis uses either ranges or probability distributions to determine uncertainty in the results.

VOC Volatile organic compound.

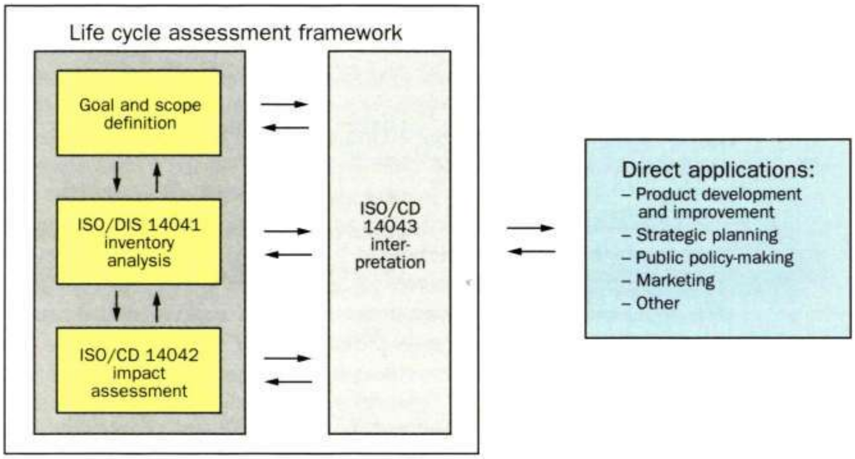

Life cycle assessment framework
Goal and scope definition
ISO/DIS 14041 inventory analysis
ISO/CD 14042 impact assessment
ISO/CD 14043 inter-pretation
Direct applications:
- Product development and improvement
- Strategic planning
- Public policy-making
- Marketing
- Other

Figure 3
ISO/DIS 14040 – phases of the life-cycle assessment

herent difficulties in deriving environmental significance from input and output data that had been aggregated, allocated, and normalized.

Besides SETAC's achievements, several national and regional efforts were launched to describe the techniques, application and limitations of the life-cycle assessment. These include the Nordic Guidelines, which were developed by a consortium of Nordic research institutions, and draft guidelines by the Environmental Protection Agency (EPA) in the US. These initiatives lent further support to the basic SETAC framework, including clear recognition of the need for an impact-assessment phase.

ISO standardization of the 14040 series on life-cycle assessment

Following the United Nations conference on environment and development in 1992, and the subsequent formation of the Strategic Advisory Group on the Environment (SAGE), pressure grew to make the life-cycle assessment the standard tool for environmental management. After the ISO 14000 standardization effort was formally initialized in 1994, the SETAC technical framework was adopted as the starting point for standardizing the life-cycle assessment (Table 1).

Based on existing practice, the work of formulating general life-cycle assessment principles (14040) and of writing the mass and energy inventory standards (14041) proceeded rapidly to draft international standard (DIS). However, efforts to standardize the life-cycle impact assessment (LCIA, 14042), which is based on 14041, ran into severe roadblocks because the life-cycle inventory was never designed to link input and output data to actual environmental effects.

ISO 14040 – principles and framework of the LCA

The framework standard ISO 14040, which defines the content of any investigation that claims to be a life-cycle assessment, clearly states its limitations as being one of several environmental management techniques (risk assessment, environmental performance evaluation, environmental auditing). Moreover, it does not address the economic or social aspects of a product or product system. A life-cycle assessment must define its goal and scope, inventory analysis, impact assessment and interpretation of results (Figure 3).

It is no secret that by fudging with the scope and system boundaries of an investigation it has been possible to manipulate results. Therefore, to be credible, investigations must have well-defined goals and scope.

Table 1
Documents in the ISO series of standards

Document	Title	Status (Fall, 1997)
14040	LCA Principles and Framework	Draft international standard
14041	LCA Life-Cycle Inventory Analysis (mass and energy)	Draft international standard
14042	LCA Life-Cycle Impact Assessment	Committee draft
14043	LCA Interpretation	Committee draft

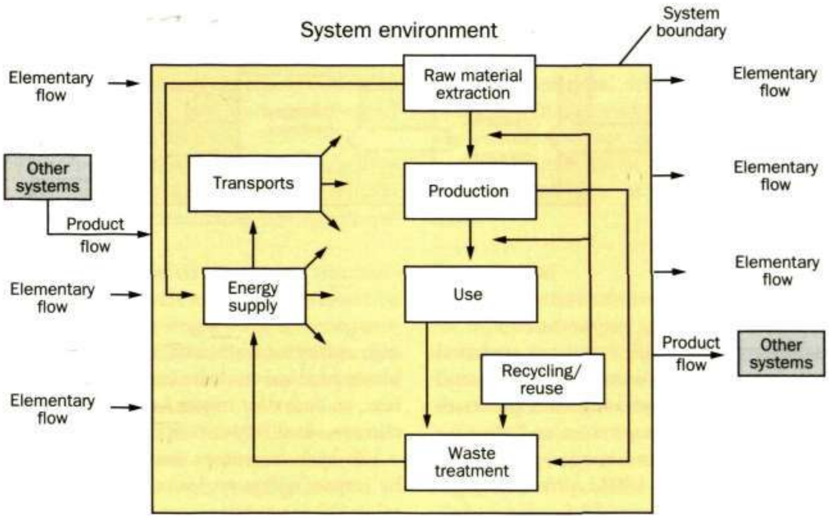


Figure 4
Example of a product system, including unit processes, elementary flows, and product flows that cross the system boundary (either into or out of the system), and intermediate product flows within the system.

In the past, the life-cycle assessment could knowingly be used to give false marketing messages. It was also self-deceptive; that is, if not scientifically well founded, it could give false guidance.

The scope of an investigation must clearly define and describe the following items:

- Function – the scope of the life-cycle assessment must specify the function of the system being studied.
- Functional unit – a functional unit is a measure of the performance of the product system's functional output. A system may comprise several functions. The function being studied must be directly dependent on the goal and scope of the assessment; moreover, its related functional unit must be measurable.

A system's boundaries determine which unit processes are to be included in the life-cycle assessment. Several factors affect system boundaries, including the intended application of the study, assumptions, cut-off criteria, data and cost constraints, and the intended audience.

Data-quality requirements specify in general terms the characteristics of the data needed for the study. The data-quality requirements of an investigation which supports a comparative assertion that is to be disclosed to the public must stand up to scrutiny. The same goes for rules that stipulate how the critical review process may be performed.

ISO 14040 defines the requirements that must be fulfilled by the life-cycle inventory analysis, the life-cycle impact analysis, the life-cycle interpretation, and the final reports.

ISO/DIS 14041 – life-cycle inventory analysis

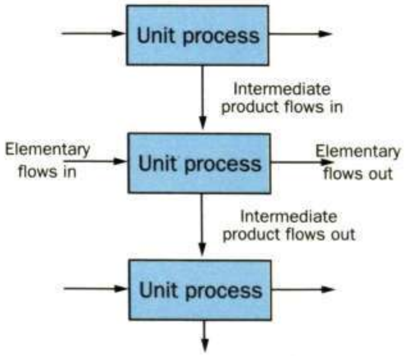
The most important part of the life-cycle inventory analysis (ISO/DIS 14041) deals with how data is collected and handled in order to give results of high integrity.

A product system is a collection of operations whose flow of intermediate products performs one or more defined functions (Figure 4). The essential property of a product system is characterized by its function; thus, a product system is not defined solely in terms of its final products.

Product systems are subdivided into unit processes, where each process encompasses the activities of a single operation or group of operations (Figure 5). Because the system is a physical system, each unit process obeys the laws of conservation of mass and energy. Therefore, mass and energy balances provide a useful check on the validity of any unit process description.

One problem with any life-cycle assessment method lies in the allocation or partitioning of the input and output flows of a unit process to the product system being studied. Life-cycle inventory analyses consist of linking unit processes within an overall system by simple material and energy flows. In practice, however, few industrial

Figure 5
Example of a unit process within a product system.



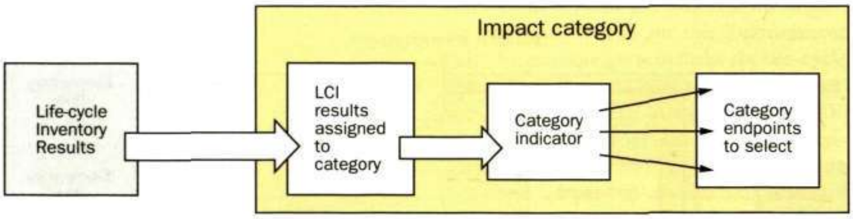

Impact category
Life-cycle Inventory Results
LCI results assigned to category
Category indicator
Category endpoints to select

Figure 6
Schematic of concepts for defining impact categories.

processes yield a single output or are based on a linearity of raw-material input and output. In fact, most industrial processes yield more than one product and they recycle intermediate or discarded products as raw materials. ISO 14041 provides guidance on the principles and procedures of allocation.

Companies and institutions that gather mass and energy unit process data into huge data banks for subsequent sale to third parties have driven the development of life-cycle inventory methods. Their intention is to create easy-to-use database tools for designers and construction engineers whose work influences the ultimate environmental profile of a product system (design for the environment, DfE). The life-cycle inventory will also play a major role in coping with upcoming legislation in much of Northern Europe on producer responsibility – which directly addresses the cradle-to-grave flow of mass in product systems.

ISO/CD 14042 – life-cycle impact assessment

The life-cycle impact assessment assists in interpreting and identifying the significance of life-cycle inventory results, thereby making them more understandable and manageable relative to the natural environment, human health, and resources. It also helps direct the focus of other environmental techniques for assessing in greater detail a particular environmental impact and for better ascertaining the significance of the impact.

The general procedure for conducting a life-cycle impact assessment begins with selecting and defining impact categories (Figure 6). An impact category illustrates the relationship between certain inventory results and a specific indicator and its endpoint. Typical impact categories are global warming, eutrophication, and the formation of photochemical oxidants (smog). In this context, an endpoint might be a change in the climate, dead fish or hospitalized people.

Life-cycle inventory results are classified by impact category. Inventory results that solely belong to one category are assigned to that category; for example, phosphate is exclusively assigned to eutrophication. Other results relate to several categories; for example, nitrogen oxides contribute to acidification and eutrophication.

By means of modeling categories (often referred to as characterization), each contribution assigned to an indicator is normalized – through equivalency factors – to the overall category effect. The scientific relevance of using characterization and normalization factors has been the subject of much debate. Several supporting techniques have been suggested for evaluating the significance of end results. The most important application of the life-cycle inventory assessment directly compares competing product systems whose major environmental impact categories are well recognized.

ISO/CD 14043 – life-cycle interpretation

Life-cycle interpretation, which is the last phase of a complete life-cycle assessment, condenses useful results for use by clients in their decision-making processes. Some parties advocate that it is possible to bypass the life-cycle impact assessment, jumping directly from life-cycle inventory (14041) to the interpretation phase.

At the same time, the need for assessing environmental impact has given rise to several impact-evaluation models. Owing to numerous shortcomings, however, these models have not been accepted as the basis of the impact-assessment standard. In general, they

- rely on aggregated and allocated mass and

energy inventory data;
- introduce subjective weighting factors, in order to quantify impact by category;
- introduce subjective approaches for ranking the overall environmental impact with a single score;
- introduce social and economic indicators that obscure the quantification and interpretation of measurable environmental impact.

DIS 14040 explicitly states that the findings of the life-cycle assessment may not be boiled down to a single score for comparative assertion. It also reinforces the fact that social and economic factors are not part of the life-cycle assessment.

According to DIS 14043, the interpretation phase of the life-cycle assessment must include the following steps (Figure 7):
- Identify significant environmental issues.
- Evaluate the issues by
  - completeness check;
  - sensitivity check;
  - consistency check;
  - other checks.
- Draft conclusions and recommendations for the final report.

Limitations of CD 14042, as based on ISO/DIS 14041

The introduction to the current draft of ISO/CD 14042 states that the life-cycle impact assessment "has limitations that are related to both the system-wide efforts and energy and mass functional unit approach. The inventory-accounting convention may omit or not provide spatial, temporal, threshold and linear/non-linear and other environmental information."[2]

The key limitations that result from the current technical framework are: data treatment; aggregation (spatial resolution, temporal resolution, threshold and linear/nonlinear modeling); allocation (general allocation procedures, recycling allocation procedures); normalization; reliance on modeled, averaged data; no link to receiving environment; and omission of critical impact.

Data treatment

Input and output data are converted into mass and energy unit values. However, when this is done, all connections to vital characterization data are lost, such as the concentration and rate of emission flows. Also, converting the data into mass and energy values builds a mass bias into subsequent calculations. This practice encourages the premature aggregation of certain output data. For instance, volatile organic compounds (VOC) are commonly aggregated at the outset, without consideration for the varying magnitude of impact that a compound or mixture of compounds might have.

Aggregation

The aggregation of mass and energy values for input and output across all unit operations removes most of the spatial, temporal, threshold and linear/non-linear information that is needed to characterize actual effects.
- *Spatial resolution.* Crucial information that could link input and output data to regional and localized effects is lost when data from different unit operations are combined.

  Example: Vital product components are obtained from many parts of the world. Some figure is used to account for the water that went into the process of manufacturing each component. However, this figure tells us nothing about the supply of water in the different parts of the world; neither does it suggest the fate of the water after its use.
- *Temporal resolution.* The aggregation of data is equally problematic from a temporal point of view, since environmental processes and their responses to emissions loading occur over different time scales. Data aggregation erases all temporal information, obscuring differences in production rates, emission rates and environmental persistence, which may vary from site to site. Example: Some releases of organic molecules into air and water disappear almost immediately, having only a negligible environmental effect. Other releases, which are environmentally potent, persist a long time in nature.

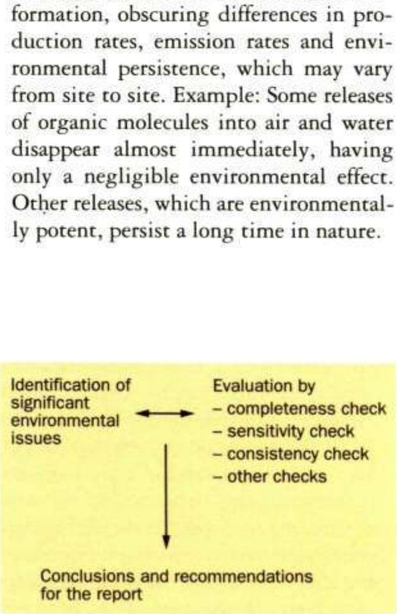

Identification of significant environmental issues
Evaluation by
– completeness check
– sensitivity check
– consistency check
– other checks
Conclusions and recommendations for the report

Figure 7
Steps in the interpretation phase of the life-cycle assessment.

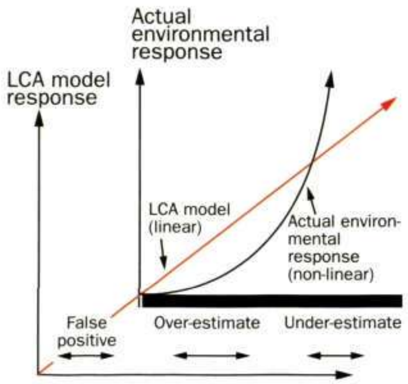


Figure 8
Presumptions about linear relationships can be very misleading.

- *Threshold and linear/non-linear modeling.* Aggregation obscures the linearity or non-linearity of responses to stressors and to the existence of thresholds. Impact modeling that is based on unit or cumulative mass and energy values creates false positives (Figure 8) when loads fall below a response threshold – the level at which adverse environmental responses may first be observed. Moreover, thresholds are not necessarily constant but may vary from site to site. Above a threshold, impact modeling based on mass and energy inventory loads cannot account for non-linear responses and may significantly underestimate the impact. Dioxins are a typical example.

Allocation

- *General allocation procedures.* The allocation of mass and energy values among the co-products of a given unit operation automatically assumes the existence of a linear relationship to environmental effects between the relative masses of co-products and their associated input and output. Although variations of the allocation approach, such as stochiometric allocation, are embedded into DIS 14041, they nonetheless retain the same subjective assumption. Example: Besides producing steel, steel-production systems also generate slag – a waste product that has found a secondary-use market and is now considered a co-product. A typical allocation of mass assigns 30% of all system input and output to slag, since that is its proportional mass relative to the total system. Nonetheless, the correlated reduction in input and output assigned to steel (100%-70%=30%) is purely arbitrary, given that the amount of iron ore or coke required to make steel has not changed.
- *Recycling allocation procedures.* The allocation of mass and energy values among virgin materials and corresponding recycled materials in an extended system assumes the existence of a linear relationship to effects between the original input and output of the virgin system and the input and output of subsequent recycling steps. However, this assumption is erroneous. With the exception of same-site closed-loop recycling systems, open-loop and closed-loop system processes do not occur in the same time period; they do not operate in the same receiving environments; and they seldom use the same technologies. Thus, identical emissions from different sites may result in a different magnitude of environmental effects.

Normalization

Mass and energy balances tend to dominate among the practice of life-cycle assessments that aim to study industrial efficiencies. Not surprisingly, then, normalization has also been geared towards this end, with the appropriate functional units of measure being per joule and per kilogram. However, the normalization of allocated and aggregated mass and energy values to corresponding functional units of mass and energy severs the links between input and output data and effects. Functional units have no relationship to the size of a manufacturing plant or to its total output. Therefore, they preclude environmental assessment, which clearly requires a different kind of normalization than what is used for studying efficiency. Example: Only a very small percentage of the output of a large refinery may be devoted to a particular solvent. In turn, only a small percentage of the solvent that is produced may be shipped for use in the system being studied. Thus, the inventory in no way represents the actual emission and environmental effects of the refinery.

Reliance on modeled, averaged data

Most life-cycle inventories rely on generic modeled-data sets for a portion of their data, often without reference to the origin of the data. Reliance on generic data can substantially increase the uncertainty of the study since the range of particular input and output data can vary greatly per functional unit. For example, the oil industry has been very good at providing generic information on their unit operations, which is an effective way of hiding low performers, since no one sees or even thinks to look for variability.

No link to the receiving environment

The only way to settle the environmental relevance of input and output data is to establish mechanistic links between input and output and actual environmental effects. Under the SETAC framework reflected in 14040, 14041 and 14042, the life-cycle assessment is disconnected from the receiving environment and cannot provide these links. Thus, it does not contain any procedures for characterizing measurement endpoints or appropriate nodal indicators.

Establishing equivalent impact potentials, such as global warming potentials (GWP) or photochemical ozone-creation potentials (POCP), is not synonymous with establishing system links to the receiving environment.

Omission of critical impact

The focus of the mass and energy framework on input and output data excludes the examination of other system activities that might cause environmental effects. For instance, although it is seldom tracked through input and output calculations, habitat depletion – as a result of land usage – is clearly a system impact. Man's use of automobiles claims vast areas of land for paved highways, which are completely devoid of biodiversity.

Together, these factors severely restrict the usefulness of the life-cycle inventory assessment in assessing the environment. Citing once again from CD 14042: "LCA based on mass and energy balances does not identify, measure or represent actual environmental impacts; does not predict potential environmental impacts, or estimate threshold exceedance; and is not a measure of safety margins or risks."[2]

Thus, while the mass- and energy-orientated life-cycle inventory is a useful engineering tool for tracking material and energy flows in industrial systems, it generally does not provide the framework for fulfilling the need of environmental management, which is to shed real guiding light on the relationship between a company's industrial activities and actual environmental effects.

Need of integrated impact assessment

Despite the shortcomings of mass- and energy-orientated life-cycle inventories for environmental management, the life-cycle assessment remains the only true scientific method being standardized within ISO 14000. Its inherent cradle-to-grave scope of assessment represents the only possible basis for comparative assessment, environmental performance evaluation and environmental labeling.

Today, CD 14042 describes the life-cycle impact assessment as being one of many tools, suggesting that LCIA emission loading and resource indicators may be used "to indicate where other environmental techniques may provide complementary data and information to decision-makers."[2]

However, it does not clarify how or when such techniques should be used in conjunction with life-cycle assessment results. The position that the life-cycle impact assessment cannot be used for assessing impact on the environment has been met with skepticism and a growing disenchantment with the entire life-cycle assessment process.

The general scientific fields of environmental assessment are well developed, ranging across a broad spectrum of environmental disciplines and techniques. Many assessment methods are sophisticated and widely practiced, collecting valuable spatial, temporal, threshold and linear/non-linear characterization information that can be applied directly for assessing the environmental significance of emissions and used resources. Moreover, the types of data these techniques generate are often readily available through government or other databases.

The integration of these techniques into the cradle-to-grave scope of the life-cycle assessment, reorients the life-cycle assessment, enabling it more effectively to meet the needs of environmental performance evaluation, environment labeling and other environmental management activities. The need of new types of data, for evaluating en-

Table 2

Resource depletion
Water
Wood and wood fiber
Fossil fuels, biofuels, nuclear fuels
Oil and gas (non-fuel usage)
Metal ores (specific)
Minerals (specific)

Direct habitat depletion (from land usage)
Marine resources

Emission loading
Greenhouse gases
Acidification
Ground-level ozone

Stratospheric ozone layer
Aquatic oxygen depletion
Human health effects from hazardous chemicals (specific)

Eco-toxic (specific)

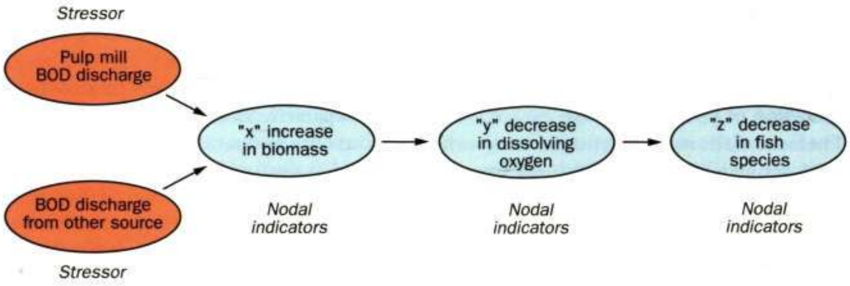


Figure 9
In this simplified stressor-effect network schematic, pulp-mill discharges combine with discharges from other sources, leading to a measurable decline in the fish population.

vironmental impact, has long been recognized within the LCA framework, as reflected, for instance, in the major findings of the SETAC workshop on life-cycle impact assessment:

"The potential for use of such data within the LCA framework can be illustrated by the example of releases that can potentially cause acidification effects. While acidification is largely attributable to SO_x and NO_x emissions, only a small fraction of given point-source emissions may actually result in a measurable acidification loading on the environment. Characterization data has been used to map out exceedances of acidity threshold by specific geographic information system (GIS) grids; such data have been compiled for much of Europe, the United States, Africa and parts of Asia. These data can be used to accurately partition out the portion of total SO_x emissions which actually results in acidification, rather than focus on the total amount released by the system.

"There are no inherent technical barriers to actually integrating useful environmental assessment techniques and their resulting environmental characterization data into the LCA framework. Equally important, by integrating these data in a formal manner, practitioners and users will be guided in a rational, consistent approach to their use."[5]

Emergence of the LCSEA – goals

The life-cycle stressor-effects assessment is the first cradle-to-grave assessment framework specifically developed for enhancing the role of the life-cycle assessment as an environmental performance evaluation and decision-making tool that complies with the objectives of ISO 14000. It is a multidisciplinary tool that explicitly integrates the life-cycle assessment into other environmental assessment techniques.

The objectives of the life-cycle stressor-effects assessment are:

- to enable users to determine the environmental significance of environmental effects caused by the input and output of industrial systems across every stage of a life cycle;
- to produce a cumulative resource-depletion and emission-loading profile of the system being studied, which can subsequently be normalized to functional units that reflect the environmental assessment objectives of the study;
- to provide necessary information for evaluating environmental performance, analyzing design and management options, and for making accurate environmental claims and product declarations;
- to satisfy the requirements of DIS 14040 and CD 14042 for comparative assertion and product declarations;
- to eliminate, at an early stage, all unnecessary life-cycle inventory work, concentrating instead on environmentally significant parts – which greatly increases cost effectiveness.

Stressor-effects networks

The stressor-effects assessment becomes the driving force of the LCA process in the LCSEA (Figures 9 and 10). Stressor-effects networks are the interlocking physical, biological and chemical events that connect a system's input, output or activity (that is, the stressor) to an observed effect on resources, the natural environment, or on human health. Stressor-effect networks may be simple or complex; often several intermediate effects, biological processes or chemical processes – called nodes – can be identified along the pathway between the initial stressors and the effect (for example, the measurement endpoint).

Stressor-effects networks may be triggered by any system-related activity. The most familiar stressor-effects networks are associated with environmental releases; for example, ground-level ozone, which is formed in the following manner:

$$VOC + NO_x + O_2 \xrightarrow{\text{UV (sunlight)}} VOC_{-1} + CO_2 + O_3 + NO_x$$

where NO_x acts as a catalyst in the presence of sunlight, oxidizing hydrocarbons into ozone. The reaction is non-linear, with a sharp spike of ozone formation when hydrocarbons (HC) or nitrous oxides (NO_x) are first introduced. Ground-level ozone has been linked to two major types of environmental effect: phytotoxicity and respiratory system effects, such as asthma in humans. In the above example, releases of NO_x and VOCs from an industrial system are stressors; the formation of ground-level ozone is a nodal indicator along the stressor-effects network; documented phytotoxicity or respiratory illness is the measurement endpoint.

A single stressor may trigger multiple effects in series or in parallel. For instance, the release of NO_x into the environment results in acidification (series effect). This in turn, increases nutrient loading in a receiving body of water, which leads to eutrophication (second series effect). Similarly, a portion of the original NO_x release lands directly on water, causing immediate eutrophication (parallel effect).

A group of stressors may also contribute to a single effect, as in the case of greenhouse gases.

Stressor-effects networks are also associated with resource extraction. Habitat depletion, for example, is caused by the bulldozing and refilling of roads that are used for harvesting timber resources. The digging and subsequent refilling of a mining pit is another example, which has significant effects on habitat and ground water. These effects, which are clearly linked to the system, would be overlooked entirely if we worked exclusively from input and output data. However, by identifying early on the stressor-effects networks associated with a given system, we can

- help ensure that the right data (data needed for drawing meaningful conclusions) is collected;
- avoid costly analysis of inventory data that has no environmental relevance.

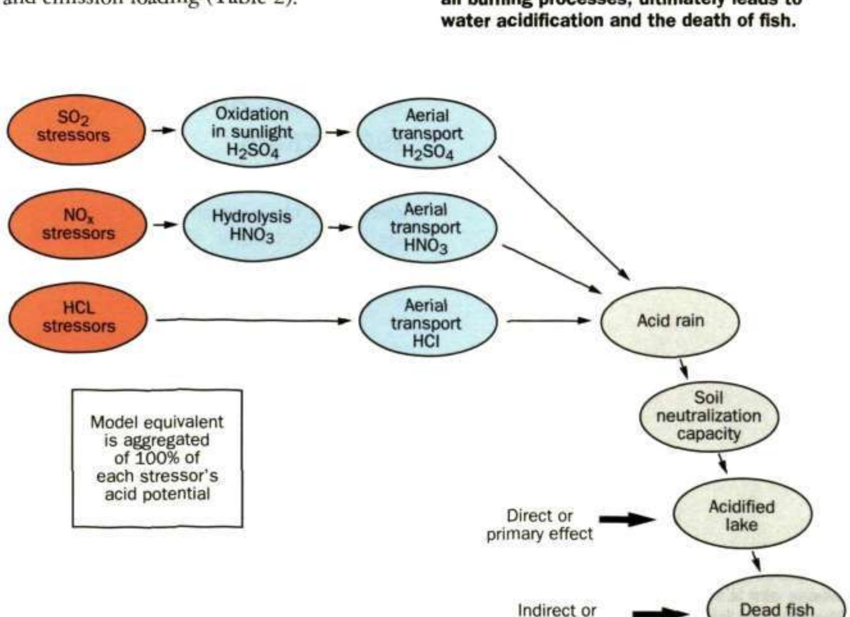
Not all system-related input and output exceed an actual threshold. When this is so, the input or output is not considered a stressor. Thanks to the stressor-effect framework, we have a systematic approach for settling this issue.

Integrating other assessment techniques into an expanded LCA framework

The life-cycle stressor-effect assessment benefits from the research and methods practice that have evolved in other areas of environmental assessment, by integrating them under the umbrella of a unified life-cycle assessment framework. Its cradle-to-grave scope ensures total-system assessments, while the integration of data generated by site-orientated impact and risk assessment techniques adds certainty to the relationship between calculated loading and effects. The extent of assessment tools and techniques that may be employed reflects the range of scientific disciplines involved in addressing environmental issues. These include resource management, toxicology, hydrology, field ecology, soil science, medical research, meteorology, climatology, and physics.

The life-cycle stressor-effects assessment condenses all information into several recognized categories of environmental impact, which are based on resource depletion and emission loading (Table 2).

Figure 10
Acid rain, which is the end result of virtually all burning processes, ultimately leads to water acidification and the death of fish.


SO_2 stressors
Oxidation in sunlight H_2SO_4
Aerial transport H_2SO_4
NO_x stressors
Hydrolysis HNO_3
Aerial transport HNO_3
HCL stressors
Aerial transport HCl
Acid rain
Soil neutralization capacity
Model equivalent is aggregated of 100% of each stressor's acid potential
Direct or primary effect
Acidified lake
Indirect or secondary effect
Dead fish

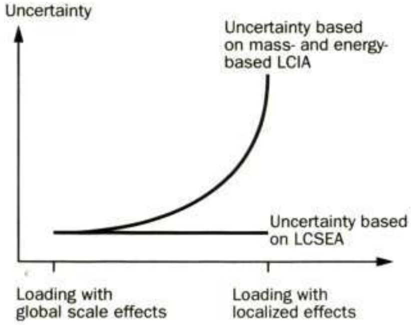


Figure 11
Using LCSEA instead of the LCIA dramatically improves the reliability of end results.

The uncertainty that plagued SETAC LCA development may largely be eliminated. Previously, the level of uncertainty was quite high for all but a few impact categories, owing to the lack of spatial, temporal, threshold and linear/non-linear resolution in traditional life-cycle impact assessments based on mass- and energy-oriented life-cycle inventories.

The only impact categories that surmount the inherent shortcomings (spatial, temporal, response shape, and threshold) of the inventory process are climate change and ozone depletion[4].

The purpose of the LCSEA framework is to reduce uncertainty by explicitly returning spatial, temporal, threshold and linear/non-linear characterization to the process (Figure 11).

Strategic significance and application of LCA techniques

The life-cycle stressor-effects assessment completes the necessary LCA toolbox, enabling stakeholders to ensure that their strategic direction into the next century is based on a sound environmental foundation. In principle, a telecommunications company needs only conduct the life-cycle stressor-effects assessment once. The results of the assessment identify – in qualitative as well as in quantitative terms – where and to what extent environmental problems exist. The assessment also identifies environmental benefits. Currently, Ericsson and AT&T are jointly conducting a life-cycle stressor-effects assessment.

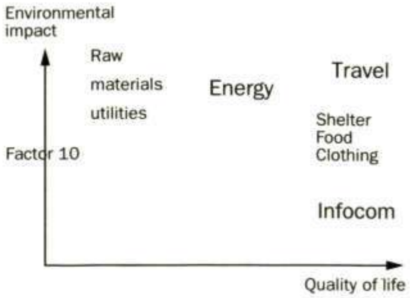


Figure 12
Compared with all other branches of industry whose aim is to satisfy basic human needs, the infocom industry has an enviable environmental position.

Another feature of the life-cycle stressor-effects assessment is that it facilitates true comparison of the relative environmental load associated with fulfilling basic human needs for transportation, lodging, food and clothing. This information will be valuable in promoting IT-based solutions for lowering the total environmental load of societal activities in developed parts of the world (Figure 12). The life-cycle stressor-effects assessment may be used to prove the environmental superiority of telecommunication solutions.

Given the findings of a comprehensive life-cycle stressor-effects assessment, the life-cycle impact assessment and the life-cycle inventory may be practiced within a very narrow scope, in principle concentrating only on known problem areas. Preferably, all product systems will use the life-cycle impact assessment.

In general, product systems undergo a product-development phase regardless of whether the product function is expanded or not. During this phase, it is interesting to compare, from an environmental point of view, the system functionality of old and new products (Figure 13). Ericsson and Telia are currently engaged in a study of this kind (green telecom services).

After the product system has been investigated, and necessary or desirable areas of improvement have been identified, designers may apply information from the life-cycle inventory, designing for the environment. Design and engineering departments are frequently called upon to improve a product under short lead times. To aid them, special software applications are being developed that contain world environmental legislation.

Future challenge

In general, normal technical development and improved environmental performance neither exclude nor contradict one another.

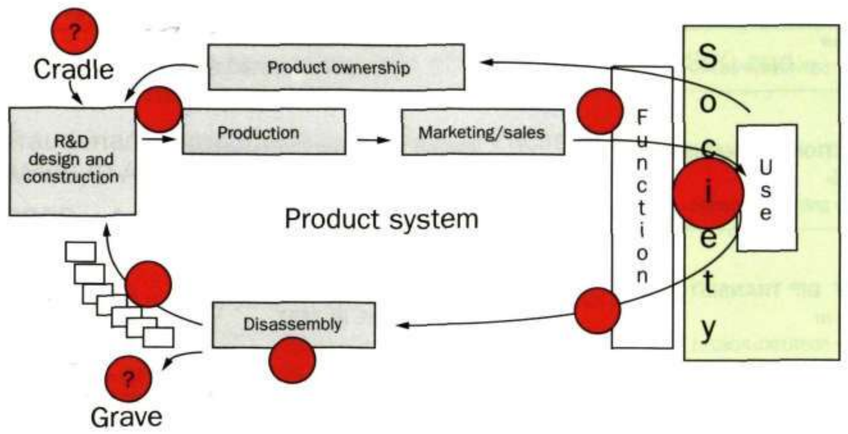


Figure 13
The contribution of environmental load from a typical product system of infocom products is represented by the areas colored in red.

Whereas the main yardstick of technical development is cost-benefit analysis, environmental improvement is measured in terms of dematerialization (services instead of mass), design for environment, and end-of-life (EOL) treatment. Offering customers more useful services that consume less mass and energy is a common denominator that also satisfies environmental requirements. However, environmental requirements restrict

- the use of certain kinds of mass (mercury, cadmium, etc.);
- the waste of energy;
- emissions of environmentally harmful waste materials.

The life-cycle assessment has not matured to the point that it can handle connections between industrial product systems and their environmental effects and economic implications. Where industry is concerned, the real value of the life-cycle assessment will manifest itself when the LCA produces a credible link between molecular and money-accounting principles.

End of Part 2. Part 1 dealt with various international perspectives of the environmental issue[5]. Part 3 describes how Ericsson can apply findings from the life-cycle assessment, designing for the environment and labeling products and services according to various developing ISO standards.

References

1 Nature, Vol. 388, 14 August 1997
2 Draft of ISO/CD 14042.2
3 Fava, J., Consoli, S., Denison, R., Dickson, K., Mohin, T., Vigon, B.: A Conceptual Framework for Life-Cycle Impact Assessment, Society of Environmental Toxicology and Chemistry and SETAC Foundation for Environmental Education, Inc., Pensacola, Florida, March, 1993.
4 Owens, J.W., ISO TC 207/SC5/WG4/N-30, Discussion paper on LCA Impact Assessment Category. May 1995.
5 Hedblom, M-O.: Environment, for better or worse (Part 1).Ericsson Review 74(1997):3, pp. 124-129.

New patents within Ericsson

OMT LOADING
Benny Boman
Patentnumber 5654901/P06340

GENERIC INTERACTION MECHANISM
Robert Khello
Patentnumber 5657451/P06062

ISOLATED RESURF BIP TRANSISTOR 2
Andrej Litwin
Patentnumber 5659190/P06231

ASIC MODE SELECTION
Dan Lindqvist
Patentnumber 505556/P06711

JOINT REGION CONTACT
Karl-Erik Leeb
patentnumber 505658/P05853

DEFORMED & DIFFUSED ATTENUATOR
Wenxin Zheng
Ola Hultén
Patentnumber 505591/P06404

DETECTING BURST SIZE ON DCC
Torbjörn Wård
Patentnumber 5663958/P06876

AUTH PAGER
Johan Falk
Björn Jonsson
Patentnumber 5668876/P06151

ROUTING VERIFICATION LOAD
Roch Glitho
Patentnumber 5544154/P06501

VERIFICATION TEST SCHEDULING
Roch Glitho
Richard Holm
Patentnumber 5638357/P06592

CALL BACK SERVICE
Howard Hsu
Patentnumber 5661790/P06739

TIMING RECOVERY TECHNIQUE
Christer Sölve
Antoni Fertner
Patentnumber 5675612/P06496

SOCKETFREE IC TEST
Åke Gustafson
Patentnumber 505869/P06264

PROSHARE IN CAS NETWORKS
Leif Isaksson
Patentnumber 505905/P06693

FUSION CURRENT CONTROL
Wenxin Zheng
Patentnumber 505782/P06403

TWIN-CORE TO SINGLE-CORE
Wenxin Zheng
Patentnumber 505771/P06268

WAVEGUIDE SYSTEM CONCEPT
Anders Qvist
Kennet Berntsson
Per Olof Glinder
Patentnumber 505504/P07011

MICRO-CELLS IN ATM-CELLS
Lars-Göran Petersen
Mats Olstedt
Patentnumber 505845/P06384

POLARIZATION SELECTIVE SIDEWALLS
Björn Johannisson
Peter Svedhem
Lars-Erland Torstensson
Patentnumber 505796/P06738

Contents

Previous issues

No. 4, 1996

No. 1, 1997

No. 2, 1997

No. 3, 1997

Telefonaktiebolaget L M Ericsson
S-126 25 Stockholm, Sweden
Tele: +46 8 7190000
Fax: +46 8 6812710

ISSN 0014-0171
Ljungföretagen, Örebro 1997