

# Approach to E2E service assurance on the mobile Internet

Ratko Vukovic and Jan Rådemar

The use of the mobile Internet is expected to escalate. Businesses will rely more heavily on communication infrastructure and services provided by public carriers. Network and information systems will mean the difference between business success and failure. That is why the industry is setting standards for business-critical applications through user-driven service-level agreements that establish committed levels of network and application availability and responsiveness.

End-to-end (E2E) service performance and quality of service will be two of the major parameters that differentiate service providers in the mobile Internet marketplace. To understand the problems and to be in a position to propose a concept for developing a management solution, Ericsson has analyzed the most important aspects of end-to-end service-performance assurance.

## Challenges in service management

Telecom network and information systems play an increasingly important role as strategic components in today's successful businesses—a role that is transforming traditional operations into e-commerce and m-commerce operations. To bring services to communities, the main requirements put on the telecommunications network and service provider (SP) information systems are:

- multimedia applications;
- capabilities to manage large volumes of data;
- complex client-server architectures; and
- large user communities.

Businesses need a service-management strategy that incorporates a good understanding of information system and network environments and of user behavior.

Because of the large effect that network and information systems have on business success, the industry is setting standards—called service-level agreements (SLA)—that establish committed levels of network and application availability and responsiveness.

Although the individual costs of computing technology continue to fall, system costs are continuing to rise in terms of absolute price and percentage of business revenue. Providing excessive network and information technology resources as a way of guaranteeing service levels cannot be accepted as a long-term solution. The performance objectives can be met efficiently in a cost-effective way by implementing comprehensive network- and service-management solutions. These solutions also reduce the time and cost of providing the service.

The main goal of the third-generation network is to establish a multiservice network that enables rapid implementation of end-user services. Technology changes in the new telecom world increase competition and cooperation among operators by enabling a flora of services and convergence between them. The market-consolidation and business-optimization process will encourage service providers to develop new business models that focus on sharing and outsourcing network and IT resources.

Many mobile operators are now building a new GPRS-UMTS network as they make the transition from being operators of a traditional telephone company to that of providing Internet, application, and m-commerce services. To provide a wide range of services, operators are sourcing some service and network components from external providers or partners. Regardless of the origin of the service or network component, the service provider who “owns” the customer maintains responsibility for the quality of the total service portfolio. Therefore, to manage QoS-related business relationships, service providers are establishing an SLA interface to external providers (Figure 1).

Because delivered service performance is becoming a major aspect of differentiation for service providers, the service-management solution has begun to play an increasingly important role in the process of implementing new networks. As to new

### BOX A, TERMS AND ABBREVIATIONS

3GPP	Third-generation Partnership Project	PDP	Packet data protocol
ATM	Asynchronous transfer mode	QoS	Quality of service
CCS7	Common channel signaling no. 7	RANOS	Radio access network operating system
CN-OSS	Core network OSS	RNC	Radio network controller
DNS	Domain name server	SGSN	Serving GSN
FTP	File transfer protocol	SLA	Service level agreement
GGSN	Gateway GSN	SLO	Service level objective
GPRS	General packet radio service	SNMP	Simple network management protocol
GSN	GPRS support node	SP	Service provider
HLR	Home location register	SPI	Service performance indicator
HTTP	Hypertext transfer protocol	STM	Synthetic transaction monitor
IETF	Internet Engineering Task Force	TE	Terminal equipment
ISDN	Integrated services digital network	UMTS	Universal mobile telecommunications system
IT	Information technology	VPN	Virtual private network
MO	Management object	WAP	Wireless application protocol
MT	Mobile terminal		
OSS	Operations support system		
PBN	Packet backbone network		

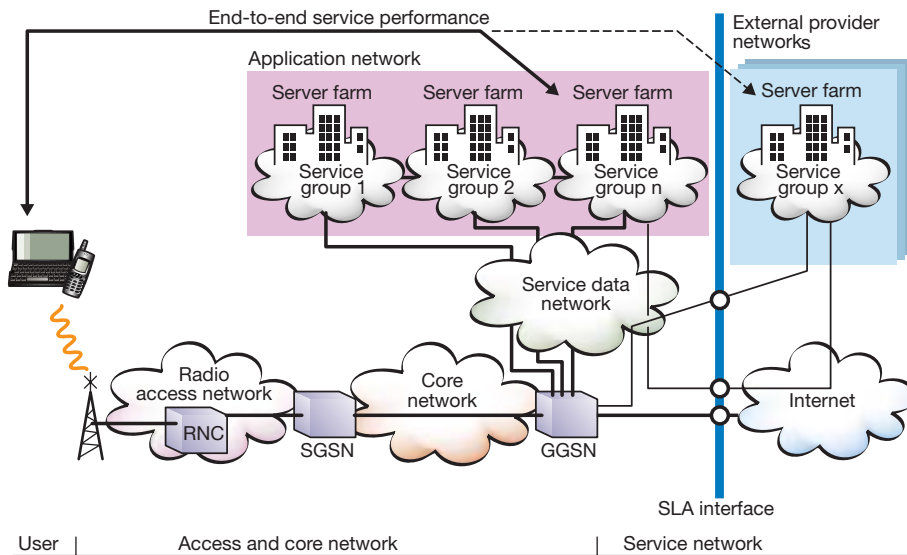


Figure 1  
The mobile Internet network.

business and technological realities, the task of providing a service-management solution can be defined as providing end-to-end service management across multiple network technologies and enterprises that contribute to the service-delivery process.

## The need for a new management approach

A good starting point when addressing the challenge of providing service management is to look at the operations model that successfully worked for circuit-switching and traditional IT networks. That model was based on managing servers and mainframe centers that fully controlled all aspects of service delivery to end-users in a single network. Servers had built-in management functions, so they needed only a relatively simple management system (Figure 2).

As an example, the service performance of voice and circuit-based data services was measured as the rate at which the switching center successfully provided circuits. The switching center also provided service-performance indicators (SPI) that were derived from out-band signaling protocol statistics (for example, CCS7). The service provider did not manage (and did not need to manage) what was happening inside the delivered circuit, and it did not need to ac-

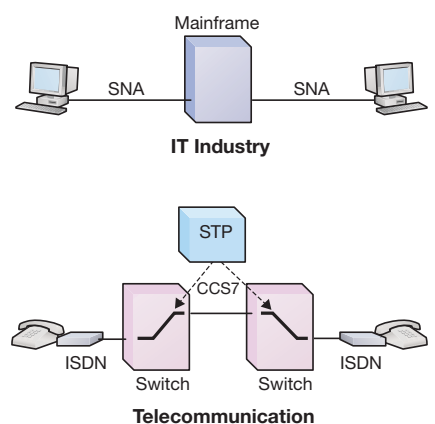
count for other systems and networks that made up the end-user service.

This delivery model, which is often referred to as the *server model*, was built on three rules that cannot be applied in the third-generation network service concept:

- a single network connected terminals with the server via a real-time protocol. At present, services are provided via multiple networks, and several application protocols are engaged between the server and terminal applications (for example, e-mail, HTTP and FTP);
- mainframe operating systems (OS) were previously used to prioritize different levels of real-time, batch, and query tasks, creating the network policy that controlled terminals and user behavior. By contrast, due to a lack of centralized task administration, a distributed network and service environment cannot guarantee consistent real-time service; and
- simple end-user terminals were used—for example, telephones, and character-oriented terminals. Modern terminals (such as mobile phones and PCs) are full-blown computers with their own management needs.

Internet services, enabled by a complex network and IT infrastructure, are becoming a common public utility, and as such, must be measured and managed in terms of con-

Figure 2  
The server management concept.



**TABLE 1, SERVICE PERFORMANCE PERSPECTIVES**

Perspective	Measurements
Individual customer	Use of network component to indicate performance of a specific service for an individual customer (or associated group of customers).
End-to-end service	Specific service delivered to all customers; the measurement provides different views of how the service is performing.
Service component	Contribution of a specific component of a service in relation to overall performance of the service.

**TABLE 2, ROLES OF THE NETWORK AND SERVICE TOPOLOGIES**

Topology	Role
Network	Represents the specific direct relationships between network components involved in delivering the service for the access, core, and service networks. The service component consists of a single or a group of network components and is a distinct measure of the performance of a service.
Service	Defines the relationship between the service components involved in delivering the service. It is used to aggregate information to generate the service-performance indicators for the service, and it must be defined in a way that isolates it from changes in the network topology.

The measurement made for each of these perspectives must be considered in terms of

- time of measurement;
- where the measurement is made;
- the value of the measurement to service assurance; and
- the cost (people and infrastructure) of providing the measurement.

To achieve service assurance for a specified service, it is necessary to analyze the service in question. A step-by-step methodology is used to define the service:

- the service is described in terms of what customers see;
- customer service performance expectations are defined—the definition covers a measure of service performance that customers would use to describe how the service is performing;
- a service-delivery life-cycle is outlined—this includes a model of the sequences of events that can occur during the delivery of the service to customers;
- the service topology is defined—this involves mapping network components into service components and thus onto the service itself (Table 2). The topology provides a model for aggregating network-based measurements into service-performance indicators (SPI); and
- SPIs are defined—the process maps specific network measurements that identify the QoS from the viewpoint of an individual customer, service component and end-to-end service.

sumption and not in terms of the elements (nodes) involved in their delivery. From a management point of view, this can be achieved by implementing a successful end-to-end (E2E) service assurance strategy.

### Development of a management strategy

Three domains have been identified on the service delivery path in the third-generation network:

- the end-user domain, which represents customer expectations and experience regarding the quality and performance of a service;
- the UMTS infrastructure, which includes the access and core networks and provides transport capability for various services but has no awareness of the service layer; and
- the service network, which includes service data and application networks and represents service-aware functions in the network and the service's content.

By combining measurements taken from all three domains, it should be possible to obtain different perspectives on service-performance assurance (Table 1).

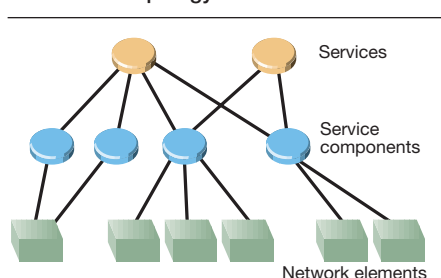
### Changing the mode of operation

Competition, customer requirements and the complexity of new networks and information systems are pushing service providers to move from the current delivery management model toward the service management consumption model. This requires a transition process that introduces new methods and tools gradually (in phases), to allow sufficient time for changes in business practices and in the organization. Process phases must be implemented proactively so that the new methods and tools can enable business growth. Figure 4 shows a simple service management maturity model.

#### Monitoring service components

As a first step, it is necessary to introduce the monitoring of service components. We must bring all service components into one common management console that is or-

**Figure 3**  
The service topology model.



ganized according to services and not according to technology or vendor (Figure 5, Phase 1). The most challenging task for system designers is to define the service topology and SPIs that are to be collected from networks or network management systems.

**Managing end-user experience**

Simply monitoring all service components will not give reliable information on end-user perception of service. A synthetic check of the service components and the service in total (from strategic locations and within carefully selected intervals) is a powerful method of testing and measuring end-user perception of service (Figure 5, Phase 2). The integration of the status of service components and reports on the synthetic transaction enables system professionals to create a consolidated view of service status.

**Controlling SLAs**

A service level agreement is a contract between a service provider and a customer. The contract guarantees a specific level of performance and reliability at a certain cost. A complete SLA can be a very complex document that describes the legal, technical, and operational aspects of the service and specifies the parties involved. From a network operation perspective, guarantees in an SLA are defined as a set of service level objectives (SLO) that comprise the set of measurements of service components to which constraints are applied.

The consolidated service view and performance reports obtained through service management in the first two phases can now be used to discover, recover, and control factors that affect the SLO. Because SLOs usually have a defined compliant percentage and an operating period during which the SLO must be compliant, it is necessary to implement mechanisms for discovering trends in service-level violation to prevent violation of an SLA. An alarm should initiate operator action or fully automatic recovery.

**Planning and predicting service performance**

As a rule, initial network implementations provide excessive network and IT resources as a way of guaranteeing the service level for the anticipated customer base. But in today's Internet environments, a successful service launch can dramatically change the demand for IT and network resources. Service performance planning is to become a

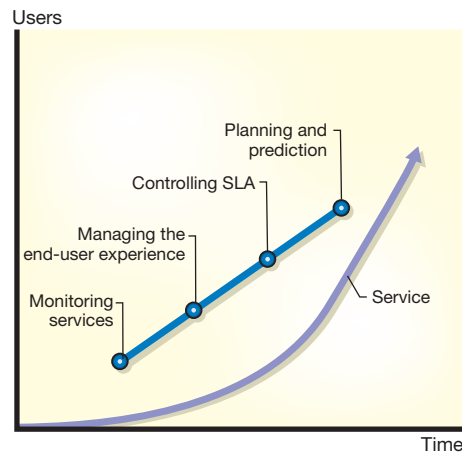


Figure 4 Evolution of the service management process.

key element of service management. A predictive service-performance model is needed to determine the resource requirements in response to *what-if* questions that reflect anticipated growth or service performance when demand is greater than predicted. To date, the industry has not come up with such a model or tools for the mobile Internet. Consequently, analysis and planning is provided for each of the networks that constitute the mobile Internet. Several powerful planning tools with simulation capabilities are available for radio access, packet backbone network (PBN) and application networks.

Figure 5 Monitoring service components and end-user experience.

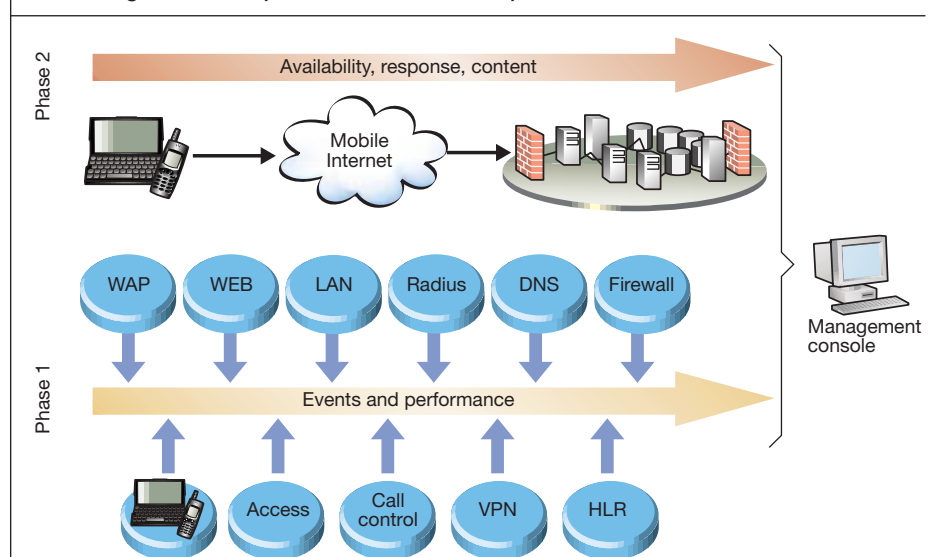
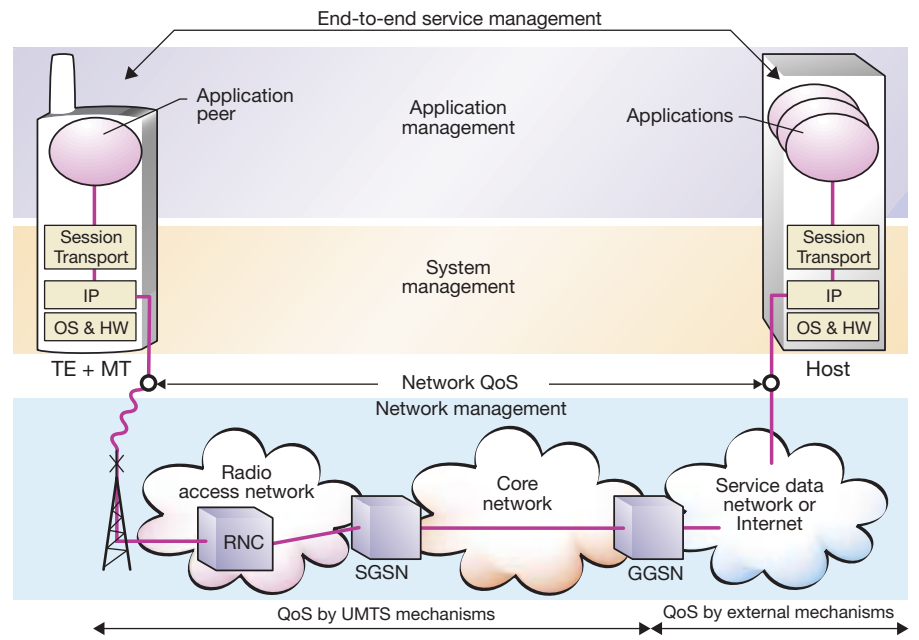


Figure 6  
End-to-end service-management for the mobile Internet.



## Managing model

Quality of service is defined as “the collective effect of service performances that determine the degree of satisfaction of a user of a service” (3G TR 21.905). Figure 6 shows a simplified mobile Internet service model that is used to identify the main contributors to *degree of satisfaction* and their respective management domains. The model depicts three major management domains (Table 3).

The third-generation network service-provisioning concept greatly helps the service provider to implement the new service-

management concept. It provides a clear separation between the bearer and service plane; and it implements a strict QoS policy mechanism that controls and manages the way the bearer service is delivered to other layers or networks that constitute the service.

For the UMTS domain, SPIs provide information on user success in connecting to a service node and information on the QoS parameters being measured during the lifecycle of the service session. SPIs at a higher network layer aggregate the SPIs at a lower network layer. This aggregation signifi-

TABLE 3, MANAGEMENT DOMAINS

Domain	This domain includes
Network management	All networks (radio access, core, and service-data networks) which constitute the bearer service that connects the terminal to an application. From a management perspective, the major characteristic of the network layer is the support for QoS throughout all segments of the bearer plane. Inside the mobile network environment (UMTS), the QoS is governed by UMTS standard mechanisms that are mapped to the Internet Engineering Task Force (IETF) QoS standards supported by the Internet or service-data network. This means that the network layer provides end-to-end QoS assurance functions between the terminal (MT) and the service node.
System	Hardware, operating system, and communication protocols of the management service node and terminal (MT+TE). The QoS classes that are supported in some of the operating systems must be mapped to Internet or service-data network QoS classes.
Application	The application that provides service to end-users and the corresponding application peers in the terminal equipment.

cantly reduces the number of SPIs needed during service management.

There are many interdependencies among systems within UMTS for ensuring end-to-end QoS, including mapping functions that the Third-generation Partnership Project (3GPP) did not standardize. These mapping functions can be defined by system vendors or by an operator and include mappings between UMTS classes of service and the underlying connectivity services based on the Internet protocol and asynchronous transfer mode (ATM). QoS provisioning at the network boundary, where the core network is connected to external PDNs (for example, Internet SPs) is not within the scope of the 3GPP specification, although it is key to providing users and applications with end-to-end QoS. Logical entities located at the gateway GPRS support node (GGSN) support the interconnection QoS management functions. The mapping of the UMTS QoS classes depends on the QoS capability of the external network and the service provider's strategy.

Monitoring and managing the QoS mapping and interconnection functions are the key tasks of the management system—for maintaining the bearer-service SLO.

## Architecture of the service-performance management system

The main objective of the service-performance management system is to manage end-to-end service performance proactively and promptly. The system architecture and management tools should be capable of demonstrating and verifying service performance compliance with SLOs and SLAs.

The main architectural principles of Ericsson's approach to implementing end-to-end service performance management are

- the monitoring of all networks involved in connecting the terminal to the application under the network-performance-monitoring function, which creates the bearer service view;
- the monitoring of all application components and their environments under the application-monitoring function, which creates the application service view; and
- the integration and correlation of the bearer and application-service views to create an end-to-end service-level view and to verify compliance with SLOs and SLAs.

Figure 7 shows the architecture of the solu-

tion. Note, however, that the service management functions in the figure (and described in this article) do not necessarily correspond to the management systems. Ericsson, together with partners (leaders) in the IT industry, implements solutions that best reflect the service provider's business practice, service offering, and current assets of the network/service management systems. To simplify the discussion of service management in this article, the architecture reflects a business model in which the service provider owns all the networks that constitute the mobile Internet. If a network segment (for instance, the packet backbone network) is supplied by a third party, the interface from the SP service management components to the corresponding OSS (PBN-OSS, which in this case, is located in the other enterprise) represents a communication channel for SLO reports and events. The type of channel (e-mail, OSS integration, and so on) and reporting rules are defined in the service level agreement. The service provider can fully rely on the reports provided by the supplier, or it can deploy simple management tools that measure delivered service in terms of consumption experience. These tools would thus replace the corresponding OSS in the architecture (for example, the PBN-OSS).

### Network performance monitor

The network performance monitor aggregates service-performance indicators from access, core, and service-data networks to provide the bearer-service view, which shows the association between users and their applications. The network performance monitor collects network-performance data from the respective sub-network managers and produces the bearer-service status view, QoS performance reports, and QoS alarms.

The most significant sources of performance indicators are the call-control network elements (GSN). As described earlier, the call-control layer aggregates performance indicators of the underlying connectivity (PBN) layer and access network. The call-control layer provides the QoS view from the user terminal to the service network via the packet data protocol (PDP) context-activation performance indicators. The data defined as the source for SPIs on this layer includes

- QoS performance reports and QoS alarms for tracking network-generic service performance; and

- the call charging record that provides the service provider with the ability to track changes on the PDP context activation at an individual level and at the aggregated (network-generic) level.

The performance-monitoring function in the radio access subnetwork manager allows operators to define a threshold related to a statistical counter or gage. Important aspects of performance degradation that should be considered in the air access network are

- accessibility, which focuses on how successful the network is in
  - establishing connections over the air interface and the load on the links;
  - pinpointing link bottlenecks;
- retainability, which represents how well the connections are maintained and under what circumstances they are dropped; and
- service integrity, which deals with the quality of connections and links, such as carrier-to-interference ratios, bit error ratios, delays, and throughput.

The service-data network, which connects the UMTS-GPRS network to application nodes that provide service to end-users, is considered as a service-specific management object (MO), as an MO-specific to a group of services, or—in the case of a VPN customer—as a user-specific MO. The service-data network can be seen as a set of VPNs provided by the PBN.

The policy-based IP VPN is emerging as the preeminent connectivity strategy. In terms of service-management, the PBN subnetwork manager performance function is required to provide reports on virtual association between two terminating points of the VPN traffic flow. With this requirement, we are isolating the service-performance function of the underlying network technology and topology that is changing continuously in fast-growing IP networks.

### Application monitor

The application monitor creates the application-service view by correlating the measurements of application availability and performance with measurement of end-user response time. To create the service view, the components of the application service-monitoring function cooperate with IT system management tools in the service network and with end-user response-time management tools that are located in the customer environment. The components used in managing the performance of the ap-

plication network fall into the following five categories:

- Agents—*in-the-box* passive software listeners that identify and collect non-simple network management protocol (non-SNMP) events and data, such as traps and system transaction logs.
- Probes—*out-of-the-box* passive software listeners whose functions are similar to those of agents.
- Monitors—active probes or management components that provide synthetic checks of a service component or an entire service function at periodic intervals across the networked environment.
- End-user agents—these are permanently or temporarily inserted in the end-user terminal to collect specific data on end-user performance.
- Application (site) monitor server—a management server collects, aggregates, categorizes, correlates, and reports management information from agents, probes, monitors, and end-user agents.

These management components can be grouped in a flexible way to constitute the management component block for a particular application or site. The set of selected components depends on the nature of the application, operating system environment, required key performance indicators, and so on. The application monitor collects, aggregates, categorizes, correlates, and reports management information in a complex application environment.

As mentioned above, merely monitoring all service components will not give reliable information on the end-user service experience, especially application response time. A synthetic transaction monitor (STM) provides a synthetic service check from the end-user environment. The STM can be programmed by recording typical user transactions and the service responses and set up to play these transactions back at defined intervals to determine the current conditions of the service. During transaction playback, the STM collects information, such as response time, domain name server (DNS) name resolution time, network connection, or application server errors.

The service provider is interested in measuring the critical response time of a service, such as for credit-card transactions within an e-commerce service. The performance of critical transactions of all users can be measured by inserting special cookies in the application server. The browser only loads the cookies during a particular service session,

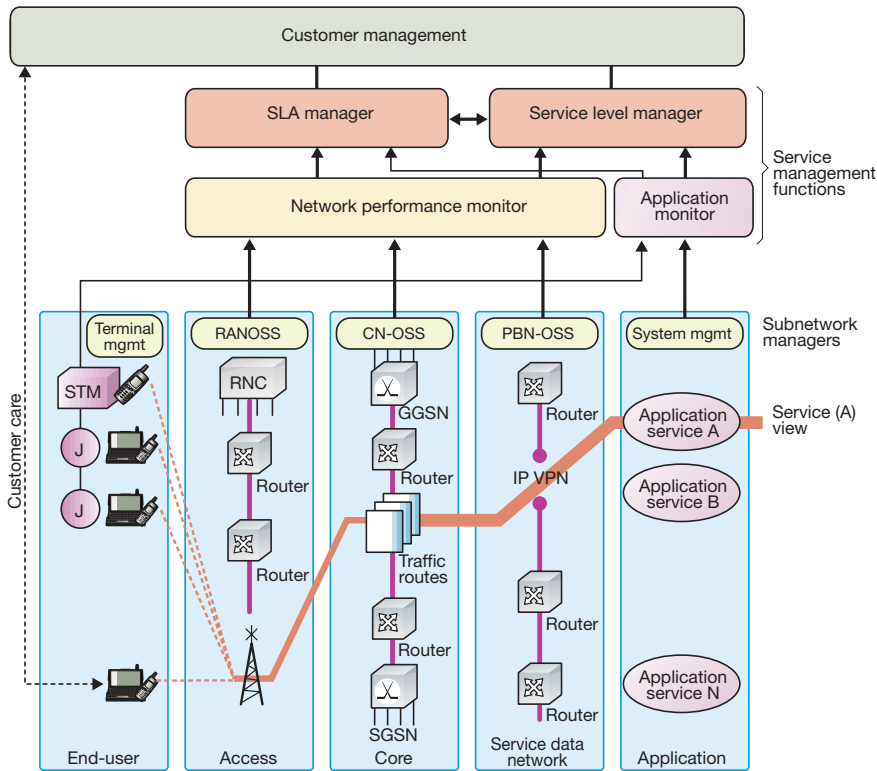


Figure 7  
Service management architecture.

and only specified measurements are made (“J” in Figure 7).

The application monitor receives measurements from all deployed remote and local agents to portray the overall user experience including the perceived status of applications.

### Service-level monitor

The service-level monitor provides an end-to-end service view and SLO surveillance that reports events received from network-performance and application-monitoring functions. Events can be notifications, such as alarms, alarm terminations, and violations of service performance.

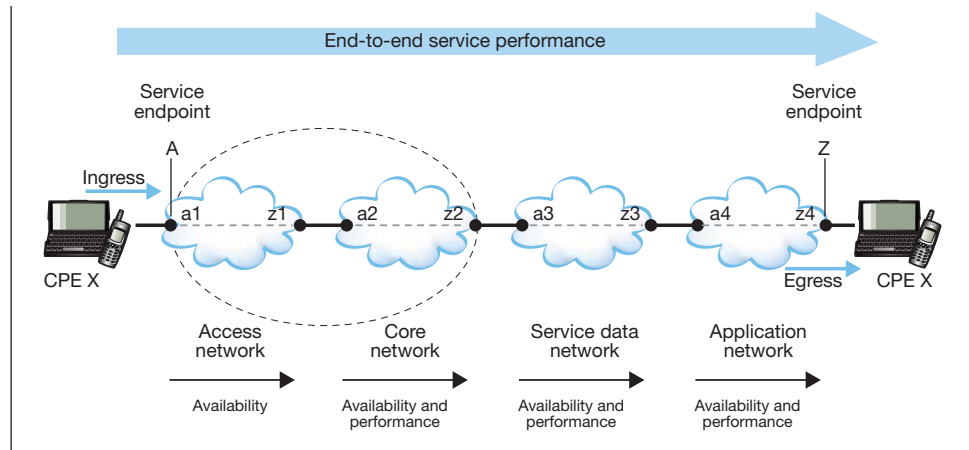
Based on the service topology model, the service-level monitor creates a network view of the service performance by correlating events from both domains. The network view includes attributes that display bearer-service, application-service, and the performance of end-user applications. This model is the basis for the process of resolving cross-domain service problems and provides the root-cause analysis of the end-to-

end service problem as well as impact analysis—for example, the impact that a deteriorated packet delay in the service-data network has on the response time of the end-user application.

The bearer and application networks form a complex environment in which the classic approach to root-cause analysis and service impact simply does not work. (This approach deals with collecting management information from all the network elements and correlating them according to a model that describes all mutual relationships.) The environment is too complex and is constantly changing.

Instead of simply receiving events from every, often unrelated, element in separate management platforms, the solution must associate only service components and relevant network elements. The service components must be defined in a way that aggregates and maintains the status of the underlying subnetwork and network elements. According to the service topology model, the end-to-end service view aggregates and maintains the status of the entire set of ser-

**Figure 8**  
Series of service topologies.



vice components (objects) that constitutes the managed service. Figure 8 illustrates the method used to create a mobile Internet service view (represented by the horizontal line) to display the state of all service-relevant components across different network domains.

#### Service-level agreement manager

The SLA manager function provides service-performance reports for managing and monitoring SLAs by correlating customer and service information. The SLA function allows service providers to associate service-performance thresholds with the conditions of individual contracts (for example, compliant percentage and operating period). Using this feature, service providers can effectively manage and report on adherence to SLA performance commitments.

Threshold levels are also used as an early detection system to address service-performance issues before they become a critical problem for customers. The SLA

thresholds are based on summaries of

- statistics on service-component performance;
- service availability alarms or alarm-termination information; and
- information from trouble reports.

The SLA thresholds are forwarded to the service-level monitor to display them with other service alarms; the service-level monitor supplies the SLA manager with the service availability events (alarms). Notification of SLA violations must be provided to the user via the customer management system. Open communication is critical for enhancing the relationship between service providers and their customers. End-users want to know that the service provider is aware that service does not live up to expectations and that actions have been taken to rectify the problem.

End-to-end service performance requires the SLA function to configure service objects to have either independent or series topologies. The SLA function allows service

providers to correctly represent the configuration of service components and accurately track service problems for components connected in series, or for independent components grouped in a service.

The ability to summarize the service performance (or service availability) of a series service is the key feature in providing statistics on end-to-end service performance or service availability, and enables network and application-layer service-performance indicators to be brought into a common equation.

## Service-management capability—the critical factor

The separation of the bearer and service plane and implementation of QoS policy mechanisms in the third-generation network service-provisioning concept greatly helps service providers to implement a modern service-management concept.

The software industry has already developed several technologies and tools to build the service-performance management systems that can provide the end-to-end view. These have mainly been developed for wire-line Internet access. Some of them can be directly applied in the mobile Internet; the others need modification. It is expected that more of these will become available on the market in the months to come.

The service design process is a critical phase for successful service management. The measurements and data that will be required for service management must be planned at an early stage of the service. These measurements and data need to be supported by network equipment and management tools. Service-management capability will become a major factor of differentiation when service providers select a network

equipment supplier or system integrator that is to provide management solutions.

## Conclusion

Delivered service performance is becoming a major aspect for differentiation amongst service providers. Consequently, the service-management solution is beginning to play an increasingly important role in the network implementation process.

As to the new business and technological realities, the task of the service-management solution can be defined as providing end-to-end service management across multiple network technologies and enterprises that contribute to the service delivery process.

As with any other utility, teleservices must be managed from the end-user perspective. This requires a transition process that assists service providers in introducing new methods and tools gradually, to allow sufficient time for changes in business practices and in the organization.

Service assurance should be planned at an early stage of service design. It requires a methodology that includes a series of steps to determine the performance management strategy in terms of the individual customer, service components, and end-to-end service.

The separation of the bearer and service plane and the implementation of QoS policy mechanisms in third-generation networks greatly helps service management professionals to define and measure key service performance indicators. In this article we have provided a high level of SPI analysis per UMTS network segment, and have described the functional architecture of the end-to-end service assurance management solution that is capable of demonstrating and verifying service performance compliance with service level objects and service level agreements.

## REFERENCES

- 1 Nilsson, T.: Toward third-generation mobile multimedia communication. *Ericsson User Review* 76 (1999):3, pp.122-131.
- 2 Forslów J., Jarret J., Moran P., Szviatovski B.: Management solutions for IP networks. *Ericsson Review* 77 (2000):1, pp.42-52.
- 3 Pehrsson; S.: WAP—The catalyst of the mobile Internet. *Ericsson Review* 77(2000):1, pp. 14-19.
- 4 Taking System Management to the Next Level: Managing Business-Process Assurance From the End-User's Perspective, White Paper, BMC Software Inc.
- 5 NETCOOL®/Internet Service Monitors™, Version 2.0, White Paper, Micromuse Inc.
- 6 BGP/MPLS VPNs, draft-rosen-rfc 2547 bis-02.txt, IATF.