

# Real-time performance monitoring and optimization of cellular systems

Per Gustås, Per Magnusson, Jan Oom and Niclas Storm

As cellular networks grow in size and complexity, the process of managing them becomes increasingly costly for mobile operators. For this reason, automated network management is very appealing.

The authors describe an architecture that uses event-based statistics for real-time performance monitoring and optimization of cellular networks. They present the results of a field trial together with SmarTone in Hong Kong, and conclude with an overview of current product developments in this field.

## Introduction

Cellular networks are becoming increasingly complex. In coming years the introduction of third-generation networks will compound this situation. As a consequence, there is a need for simple or automated operation and maintenance (O&M), to decrease operators' costs and to use hardware investments in the most effective way. Likewise, operators need customized functionality—a requirement that puts great demands on system flexibility. This kind of flexibility can be achieved by allowing the network to produce real-time events instead of counter-based statistics. But doing so requires an effective and well-structured network architecture.

This article describes an architecture that employs event-based statistics for building effective and flexible performance monitoring and optimization applications. It also points to Ericsson system properties, activities, and experiments, showing that Ericsson is well on its way toward creating a common platform for the monitoring and optimization of second- (GSM) and third-generation (WCDMA) mobile systems.

## Optimization in a radio access network

The radio access network provides connection links between mobile stations (MS) and the core telecommunications network.

In GSM systems, the radio communication with each mobile station is maintained by base station controllers (BSC), which control several base transceiver stations (BTS). A BTS covers a limited geographical area known as a cell. When a mobile station moves between cells, it is up to the BSC to move the signaling to the corresponding BTS. This process is called handover. In WCDMA systems the principle is the same, but the terminology is somewhat different.

The operator uses a management system to manage the network. In Figure 1, the emphasis is on the management signaling within the radio network. Along with a variety of other applications—for example, fault management and customer services—the typical management system has tools for helping the operator to optimize, or tune, the radio network.

Networks have traditionally been optimized by a radio engineer who, using planning tools, performance analysis data, and experience, sends configuration commands to the network.

The term optimization algorithm should be interpreted to mean software that can assist the radio engineer and handle a wide range of optimization situations, from slow variations in the radio network, such as deploying new cells, to fast variations, such as changing traffic load. Ideally, the optimization algorithms optimize system performance without continuous operator intervention. However, in complicated situations it is important to include the radio engineer in decisions.

An algorithm could be employed to decrease the coverage of a cell due to temporary traffic overload. An optimization application of this kind continuously configures the network in accordance with measurements emitted from it. If this application were removed or ceased to function, the network could continue to maintain its traffic services, albeit in a less optimized way.

## The architecture

### Distributed agents

Traditionally, management applications execute in a central management node. Pre-

### BOX A, TERMS AND ABBREVIATIONS

3GPP	Third-generation Partnership Project	MS	Mobile station
BSC	Base station controller	MT	Monitoring task
BTS	Base transceiver station	MTR	Mobile traffic recording
CORBA	Common object request broker architecture	NCS	Neighboring cell support
CT	Control task	NOX	Neighboring optimization expert
CTR	Cell traffic recording	O&M	Operation and maintenance
FAS	Frequency allocation support	OSS	Operations support system
FOX	Frequency optimization expert	PM	Performance monitoring
GPEH	General performance event handling	RANOS	Radio network operation support
GSM	Global system for mobile communication	RNC	Radio network controller
GUI	Graphical user interface	R-PMO	Real-time performance monitoring
HCS	Hierarchical cell structure	TCH	Traffic channel
IP	Internet protocol	TCP	Transport control protocol
IRP	Integration reference point	UETR	User equipment traffic recording
		WCDMA	Wideband code-division multiple access

defined statistics are periodically sent from network elements to the central management node. By contrast, real-time performance monitoring and optimization applications require a more flexible aggregation of measurement data. However, it is not feasible to send raw measurement data, since in large networks (having, for example, as many as 40,000 BTSs) this would put extreme performance requirements on the central management system and on the data capacity of the management network (Figure 2).

To create an effective environment for the performance monitoring and optimization functions that rely on large amounts of event-based measurement data, an architecture is needed that permits the distribution of functions. Raw measurement data must be handled efficiently and should be collected and processed near the source before it is forwarded to the next-higher level in the hierarchy (Figure 2).

The topic of distributed management systems has been dealt with extensively throughout the years. The challenge here has been to apply the theories to create an architecture that is suited for the real-time performance monitoring and optimization of cellular networks.

Accordingly, the applications have been split up into parts and subdivided throughout the management nodes and traffic nodes of the network. In this article, we call a module that constitutes a distributable part of the management application an agent. Note: The word agent is used simply to refer to distribution and an open interface. It does not try to quantify any degree of intelligence. Agents interact via open interfaces at different levels of the system. The concept of integration reference points (IRP) is being used within the Third-generation Partnership Project (3GPP) as the means of achieving a management standard for interoperable third-generation cellular systems.<sup>1</sup> At present, standards exist within the fault- and configuration management areas and Ericsson has submitted a proposal for performance management.

The use of interacting, yet autonomous, measurement and control functions calls for robust and scalable management systems that can be distributed across the network. Ideally, the distribution should be independent of the borders constituted by physical nodes. This is achieved by using protocols that expose the same interface to the individual agents regardless of whether the data

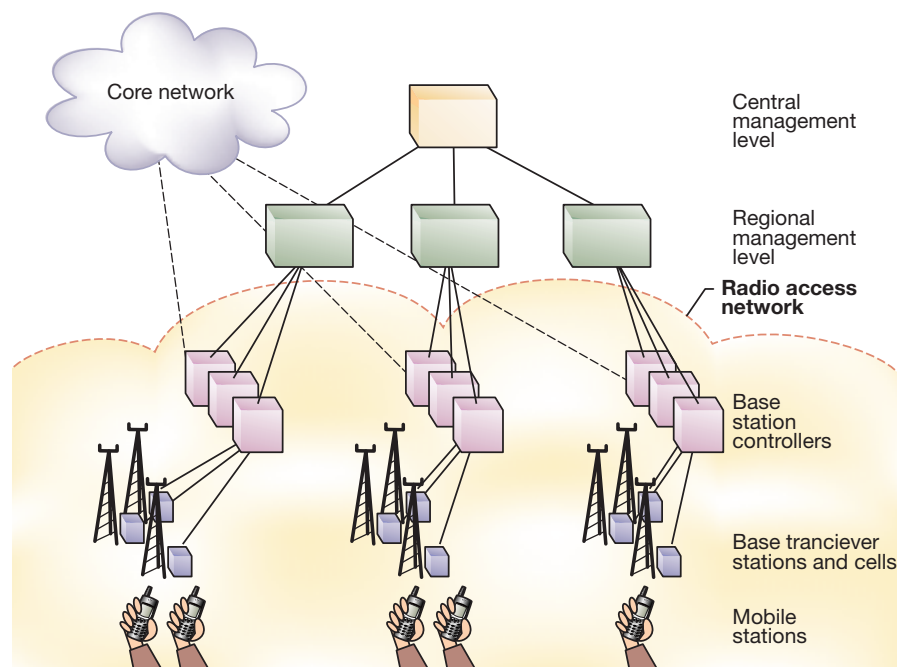
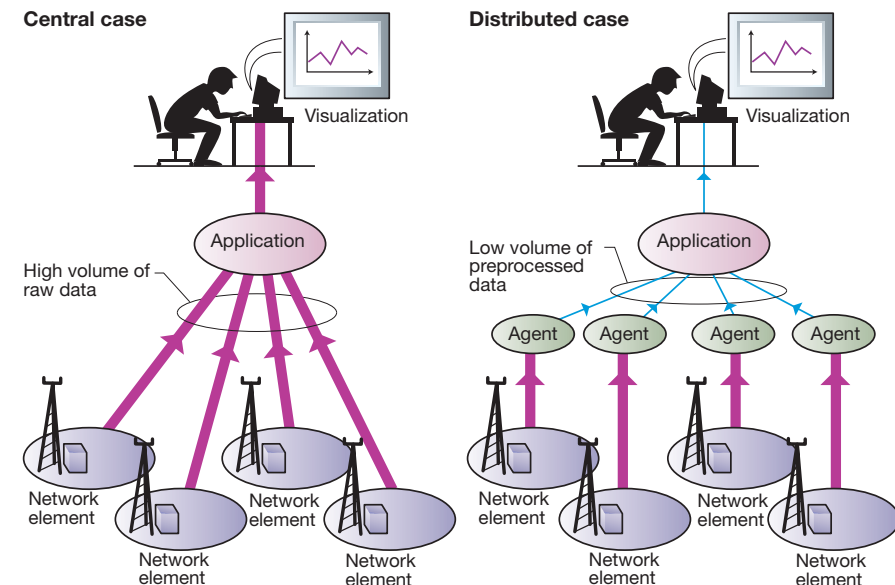
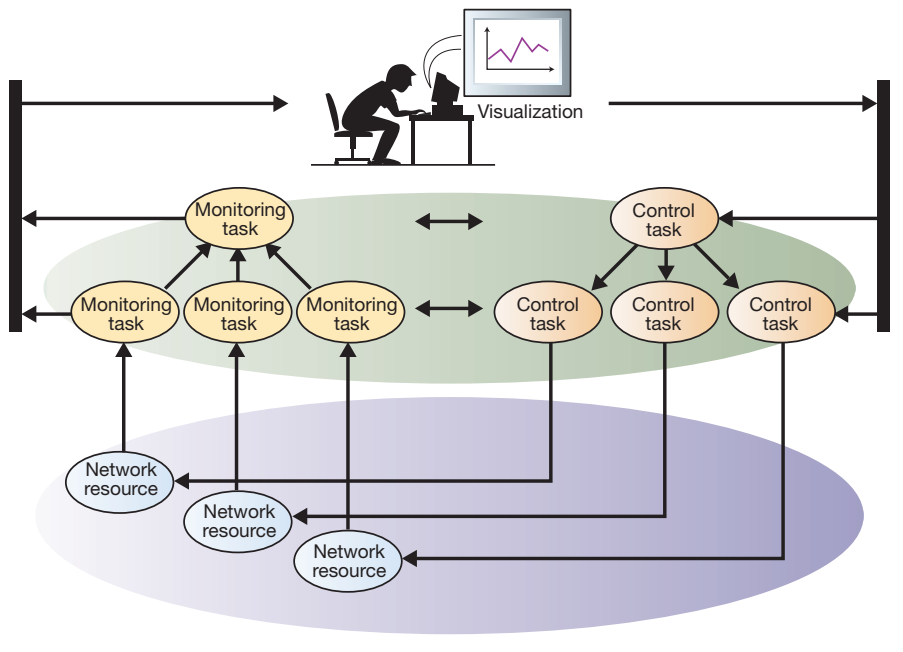


Figure 1  
A radio access network.

Figure 2  
Distributed functions are a solution for handling large amounts of data.





**Figure 3**  
Schematic view of task allocation in a system (not considering deployment). The number of layers can be extended.

nected to other monitoring tasks in any way, a strict hierarchy (Figure 3) is most often suitable for cellular systems. The lowest level monitoring tasks collect low-level information, such as bit error rate, carrier-to-interference ratio, handover events, call drop events, and so on. It uses this information to calculate statistics, such as call drop rate. Higher-level agents subscribe to lower-level information in order to compute more abstract quantities, such as capacity and quality for a region or for the entire system. These quantities convey more meaningful information to an operator or a control system.

Control tasks also operate in real-time and are used for controlling network resources, either directly or via other (typically lower-layer) control tasks. A control action can be based on performance data from one or more monitoring tasks, external control actions received from other control tasks, or both. The control rules can be of almost any type suggested in the control literature. The control tasks can also be interconnected in any way. As with monitoring tasks, however, a strict hierarchy is often preferable for cellular systems. This arrangement solves the control problems at the lowest possible level, where the problems can be handled with maximum speed. But if this is not adequate, higher-level and more cautious controllers can be invoked to resolve the control matters.

Monitoring agents are built up from one or more monitoring tasks. Optimization agents are built up from a combination of monitoring and control tasks.

### Tasks versus agents

Agents provide a means of building applications that can be scaled and distributed. If different tasks should execute on different physical nodes, then they must reside within different agents. Different tasks should also reside within different agents in order to achieve full flexibility and scalability. However, to achieve high performance, the control and monitoring tasks of the same physical node could be configured to constitute an agent:

- Tasks can communicate more efficiently than agents, since they do not require an open communication interface that supports distribution.
- If several tasks within an agent subscribe to the same information, the information needs only be sent once to the agent, which can then distribute it internally to the appropriate tasks.

is transported locally or via a connecting network.

### Monitoring and control tasks

A real-time performance monitoring or optimization application is functionally built up by means of monitoring tasks (MT) and control tasks (CT). Figure 3 shows a schematic and ideal distribution of tasks throughout a complete system. Note, however, that consideration has not been given to the physical deployment into different nodes or agents.

Measurement information flows upward in the system, whereas control information flows downward. Control tasks typically act on measurement information in the same layer and control information from the next-higher (superordinate) layer. Monitoring tasks are usually configured by control tasks at the superordinate layer.

Monitoring tasks monitor the performance of related resources (ordinarily subordinate agents or real network resources) by subscribing to real-time performance data. Once a subscription has been set up, events will be delivered to the subscriber until it cancels the subscription.

Although a monitoring task can be con-

- Different scheduling mechanisms can be used at the agent and task levels—that is, operating system-based scheduling can be employed at the agent level, whereas a simpler scheduling mechanism can be employed at the task level.

## A monitoring and optimization prototype

A prototype project was launched in 1999 with the primary goal of demonstrating that real-time performance monitoring and optimization is feasible using this architecture in a real environment.

### The optimization problem

Hierarchical cell structures (HCS) are used to push traffic down to a lower layer (smaller cells). For instance, by pushing the traffic down to layer-1 cells, operators make better use of the hardware in these cells and reduce the load in surrounding layer-2 and layer-3 cells. The result is an increase in the total network capacity.

The mobile station uses the received signal strength from a BTS to make handover decisions. The HCS feature defines a signal strength threshold (LEVTHR) for each layer-1 cell. If a mobile station that is using a layer-2 or layer-3 cell measures a layer-1 cell signal strength that exceeds this threshold it will attempt to hand the call over to the layer-1 cell. Likewise, if the signal strength falls below this threshold, the mobile station will abandon the layer-1 cell. Accordingly, there is a close relationship between the LEVTHR threshold and the geographical size of the cell. Notwithstanding, there were some drawbacks to this feature that needed to be resolved through the introduction of an optimization algorithm:

- The use of a threshold setting that is too aggressive can cause congestion in the layer-1 cell. When this is the case, it is difficult to provide service to the mobile stations that are close to the micro base station.
- A setting that is too aggressive can push traffic down causing it to suffer from interference due to frequency reuse in the network.

Note: In recent releases of GSM from Ericsson these drawbacks have been resolved.

### Optimizing algorithm

An algorithm was developed to resolve these drawbacks. The goal was to improve capacity while maintaining quality in a network



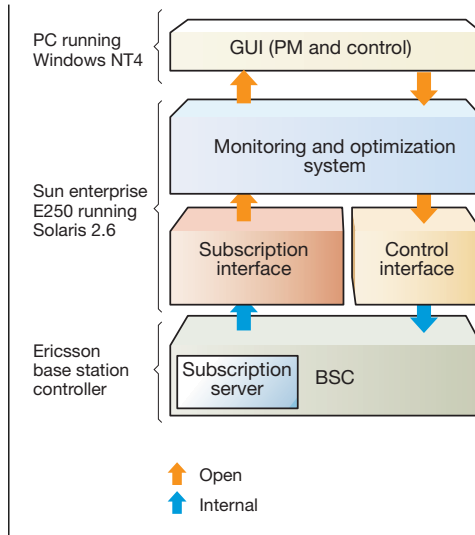
**Figure 4**  
Hierarchical cell structures.

with hierarchical cell structures. The algorithm continuously measures the capacity and quality of all regulated cells, and periodically modifies the size (LEVTHR threshold) of layer-1 cells according to the following guidelines:

- To improve the capacity (grade of service), the traffic load should be balanced between layer-1 cells and upper layer neighbors. This is done by adjusting the size of the layer-1 cell.
- If the quality is inadequate in the layer-1 (micro) cells, the size of the layer-1 cell should be decreased. To determine quality, the algorithm measures
  - drop rate (percentage of abnormally terminated calls) in the layer-1 cells; and
  - failed handovers from neighboring layer-2 and layer-3 cells to layer-1 cells.

The criteria mentioned above are periodically combined into a new cell size proposal for each layer-1 cell, which is sent to the BSC.

**Figure 5**  
Architecture of the prototype system.



- a monitoring and optimization system (also implemented in Java);
- a subscription and control interface that converts from a proprietary Ericsson format to an open agent interface (using CORBA); and
- an Ericsson BSC (Release 7) which had been extended with a subscription server that can extract real-time performance data.

The monitoring and optimization system (Figure 6) included the drop rate measurement, failed handover rate measurement, traffic load measurement, and HCS control tasks, which implement the optimizing algorithm.

The drop rate measurement task subscribes to the call disconnect and handover performed events from the BSC, and calculates the drop rate of each cell.

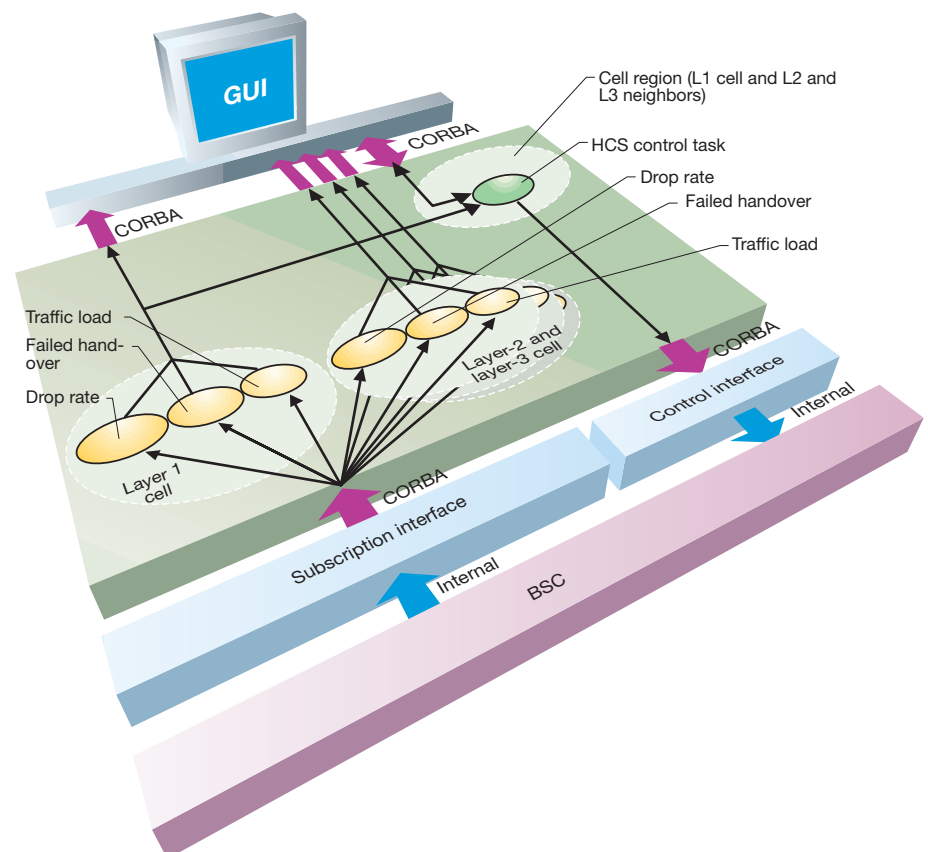
The failed handover rate measurement task subscribes to the handover attempt and handover performed events from the BSC, and calculates the handover failure rate from the layer-2 and layer-3 neighbors of a specific layer-1 cell.

### Prototype implementation

The prototype (Figure 5) consisted of

- a graphical user interface (GUI) implemented in Java—to present real-time performance statistics and to control the adaptive tuning algorithms;

**Figure 6**  
Task implementation showing the control of one layer-1 cell.



The traffic load measurement task subscribes to the traffic load event from the BSC and calculates traffic load for each cell. Traffic load is received every 10 seconds from the BSC.

The HCS control task subscribes to events from the drop rate, failed handover and traffic load measurement tasks of its own layer-1 cell and to traffic load measurement tasks of all neighboring layer-2 and layer-3 cells. It also regularly calculates a new LEVTHR threshold value.

A single host (Sun E250 workgroup server) was able to run the entire adaptive tuning system for the limited field trial area, which included a total of 105 cells. All the tasks were contained within one agent. Open interfaces (CORBA) were used between the GUI, optimization system, and the BSC, but not internally between tasks.

### Field trial

The field trial took place in Hong Kong in cooperation with SmarTone. Five cells were chosen as layer-1 test cells. All were GSM 900 cells that could be categorized as street cells rather than microcells. Their neighbors, a total of 100 cells, were also included as test cells, giving a total of 105 cells.

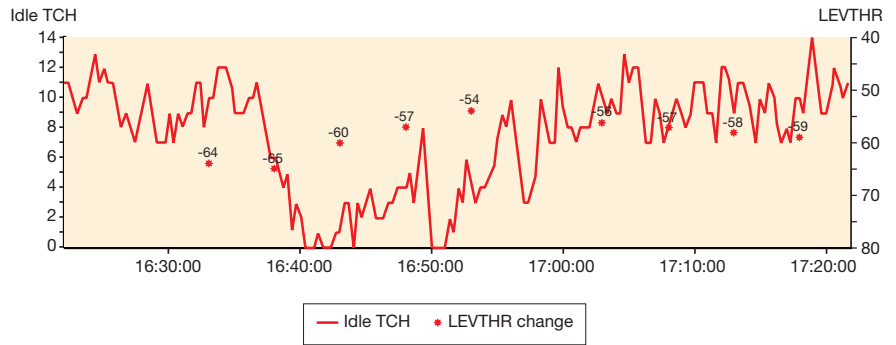
Drop rate monitoring, traffic load monitoring, and HCS control tasks were instantiated for all 105 cells. These tasks subscribed to BSC events and periodically generated the following results:

- Events from subscribed monitoring and control tasks were sent every 10 seconds to the GUI for real-time display.
- A new LEVTHR value from each HCS control task was sent every five minutes to the BSC.

### Algorithm evaluation

The adaptive tuning algorithm worked as expected according to simulations. Figure 7 shows how the number of idle channels in layer-1 cells can be increased due to changes in cell size (LEVTHR).

Ordinarily, the network worked fine. Therefore, we reduced the available capacity in the field trial area to evaluate the algorithm. When we activated the algorithm, we were able to measure enhancements to quality and capacity in the test area. Capacity improved in all five test cells—that is, congestion was reduced from 5-10% to 1%. We could also measure a significant improvement of the bit-error-rate probability distribution for two of the five test cells.<sup>2</sup> Figure 8 shows the improvement of one of



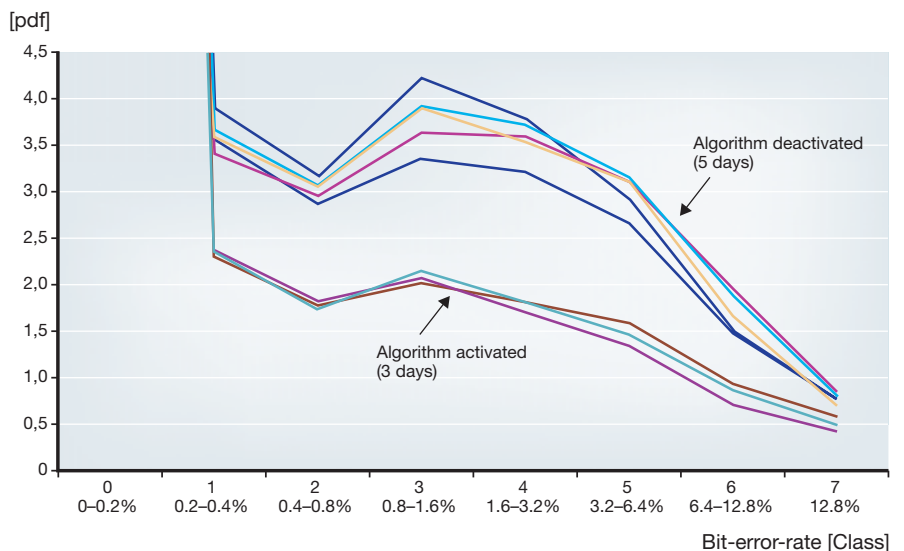
**Figure 7**  
The threshold (LEVTHR) is increased when the number of idle traffic channels decreases, forcing mobile stations to hand over to other cells. This, in turn, further increases the number of idle traffic channels.

the two cells (eight consecutive 24-hour measurements).

### Architecture evaluation

During the field trial, the prototype system was able to monitor 75 of the 105 cells included in the area. However, there was not time enough to optimize performance.

**Figure 8**  
Distribution of bit-error-rate probability for one of five test cells. Activation of the algorithm reduces the bit error rate.



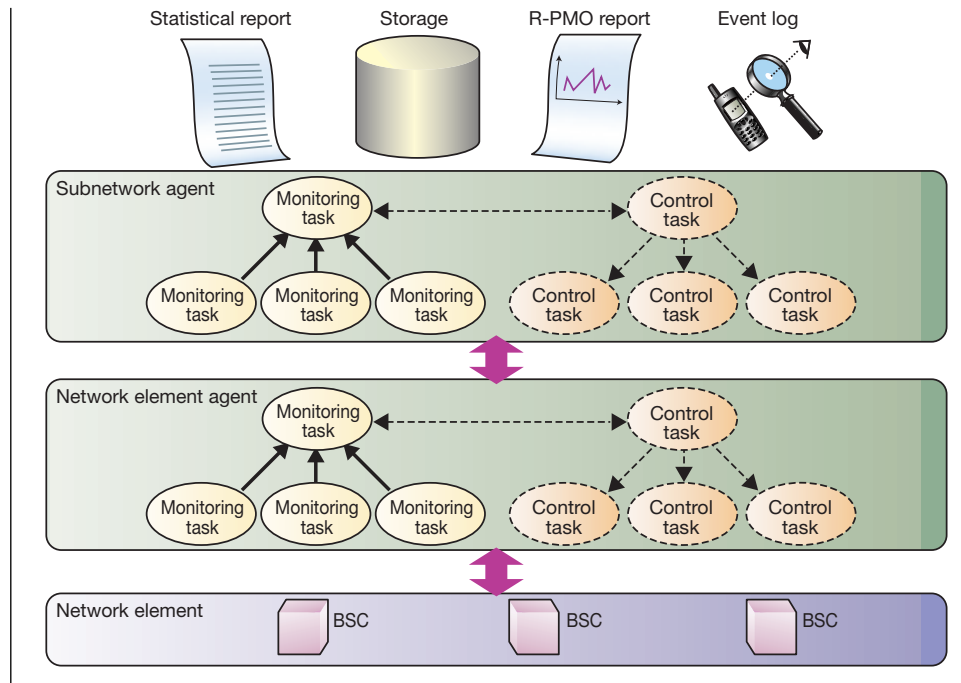


Figure 9  
An overview of the structure of the R-PMO.

The field trial has given valuable input that can be applied to future products with in this area. For instance we have learned that

- performance-critical parts should be implemented in C++ instead of in Java;
- the need for interprocess communication should be kept to a minimum; and
- when events are transmitted using CORBA calls they should be sent in batches (several events transmitted in the same message).

Latency measurements indicated that it took from 10 to 35 seconds for an event in

the network (for example, a dropped call) to be measured (percentage change in dropped calls) and displayed in the GUI. In most cases a latency of less than one minute is acceptable.

## From ideas to products

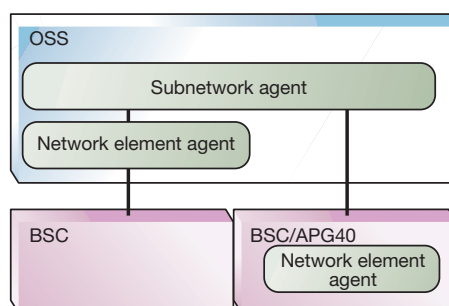
Based on input from previous product experience as well as on the results from the prototype and field test, Ericsson is now introducing event-based reporting mechanisms into GSM (Release 9.1) and WCDMA (Phase 2).

### Event mechanism in GSM

An event mechanism is being introduced into the BSC in Ericsson GSM BSS R9.1. The mechanism permits the operations and support system (OSS) to subscribe to real-time event data. The events are sent to the OSS over a TCP/IP connection.

In addition, a real-time performance monitoring (R-PMO) application is being introduced into the OSS to process real-time event data. The R-PMO provides real-time presentation of several basic monitors related to the performance of the radio network. Examples of data presented per cell are traf-

Figure 10.  
Options for distributing R-PMO functions.



fic level, TCH utilization, voice quality, and basic handover information. End-user reports provide an overview and detailed analysis of individual values.

The implementation of the R-PMO application has been greatly facilitated by the monitoring tasks (Figure 9), which perform simple and well-defined functions. The R-PMO implements a monitoring task framework that supports the administration of tasks. The framework also provides several basic services, such as giving access to configuration information, subscription mechanisms that allow monitoring tasks to subscribe to information from lower-level monitoring tasks or raw performance events. New monitoring tasks can be added to the OSS with little extra effort.

Control tasks are not currently supported, but the framework has been designed with them in mind.

The R-PMO server software has two main components—a subnetwork agent, which handles the overview of the network at the OSS level, and a network element agent, which is responsible for processing the data from a single network element.

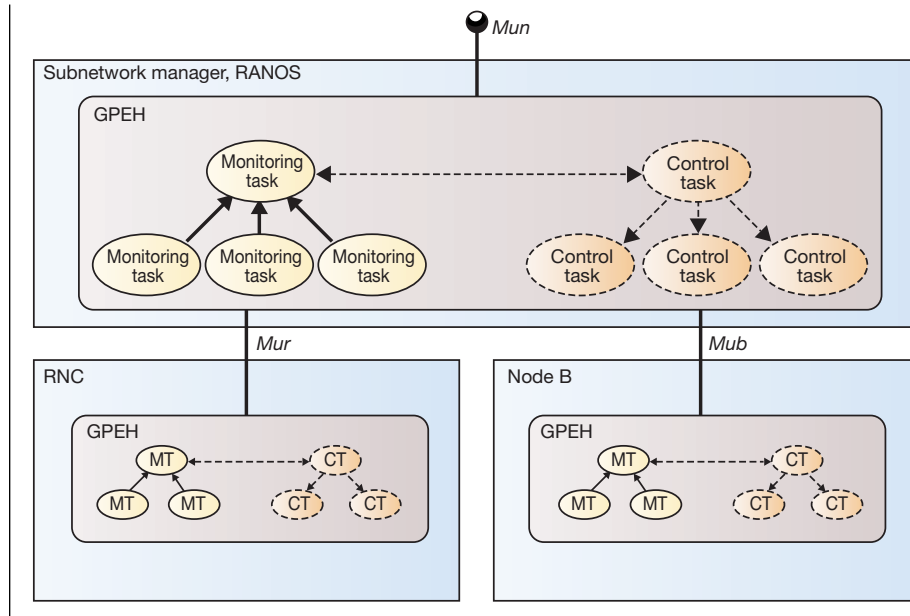
The amount of data that is sent from the BSC to the OSS is relative to the traffic volume. For large BSCs this can amount to several hundred kbit/s during peak hours. The data format used on the internal BSC-OSS link is proprietary to Ericsson. However, external applications can access the event data and also monitor values via an open PM IRP interface in the OSS.

The APG40, a new Windows NT-based input-output processor for the BSC, can be used to move part of the R-PMO functionality to the BSC itself (Figure 10). The distribution of functions to the BSC makes it possible to reduce the load on the OSS-BSC link significantly. The APG40 deployment is not part of the R9.1 version of R-PMO.

### Event mechanism in WCDMA

In Ericsson's WCDMA products (RNC, Node B and RANOS), general performance event handling (GPEH) is considered to be a cornerstone for providing real-time performance monitoring capabilities in future WCDMA releases (Figure 11).

The collection and recording of performance events is instrumented by monitoring tasks which are accessed through management interfaces (*Mun*, *Mur* and *Mub*) that support the CORBA solution set of the performance and notification integration reference points (IRP).



**Figure 11**  
Overview of the GPEH.

Thanks to the GPEH function, operators can create and modify monitoring tasks in a flexible way. Individual monitoring tasks can be defined to collect an arbitrary set of performance events for an arbitrary part of the radio network. Different monitoring tasks can be controlled as separate entities. In the current implementation, the performance events are available as periodic recordings written to a file. Subscribers are alerted to the availability of a new performance event file through the *Mu* interfaces.

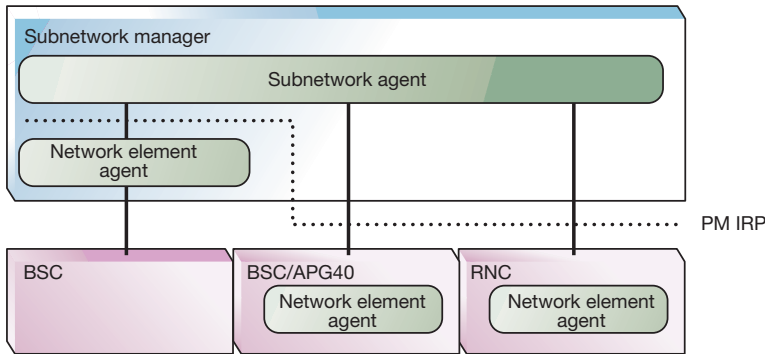
The architecture on which the GPEH function is based provides a solid base for future solutions, including the near-real-time delivery of performance events from network elements. Similarly, control tasks can be allocated to network elements; more complex control tasks can be allocated to subnetwork managers.

### Combining GSM and WCDMA

In the not-so-distant future, many mobile networks will contain GSM and WCDMA nodes. Interworking will be crucial, especially when the WCDMA part is new and has limited coverage. The management system will need to monitor both types of node. Special emphasis will be put on providing

### COPYRIGHT

Parts of this material are based on a previous article, *An architecture for Self-Configuring Systems*, © IEEE, published at Integrated Network Management 2001.



**Figure 12**  
Interface to common GSM and WCDMA functions.

tools for monitoring and tuning intersystem behavior, such as handover.

In terms of type of available event and architecture, GSM and WCDMA event mechanisms share many common principles. This means that the existing mechanisms provide a good base on which to build applications that support GSM and WCDMA. The event-reporting mechanisms in GSM and WCDMA can monitor

- control functions (connection handling);
- mobility functions (handover, cell update);
- capacity functions (admission, congestion);
- bearer functions (data throughput);
- measurement reports;
- configuration management functions; and
- inter-node communication.

GSM and WCDMA each use subscription-based principles for reporting events. The applications can thus specify which events are to be activated. Moreover, the mechanisms support the IRP concept.

### Applications

The event mechanisms in GSM and WCDMA provide a solid base that can be used for creating a wide range of applications.

### Performance monitoring

The R-PMO application provides immediate feedback on the performance of the network. Monitors, such as counter values and gauges, can be shown with low resolution and very short delay. Operators can thus react quickly to problems in the network and see the immediate effect of configuration updates. This is useful for troubleshooting and can also be used when new features are being introduced and tuned, and when parameters are being changed. The R-PMO application includes a range of predefined reports on traffic load, quality, and so on. Real-time performance events could also be considered as enablers of a wide range of performance monitoring services, such as

- alarms that supervise traffic conditions in the network;
- the identification of specified situations in the network that depend on the correlation between events. This gives almost unlimited possibilities of finding specific behaviors in the network. Examples are identifying cells with a large amount of fast moving mobile stations or ping-pong handover situations; and
- the monitoring of end-user performance to determine whether or not the customer gets what he pays for.

### Troubleshooting

Event data is a powerful tool for troubleshooting. The mobile traffic recording (MTR) application for GSM and the user equipment traffic recording (UETR) application for WCDMA allow operators to record the traffic behavior of select mobile stations, including measurement data relating to the air interface. This data can be used to check parameter settings in the network (for example, handover parameters).

The cell traffic recording (CTR) application for GSM and the corresponding application for WCDMA allow operators to record the traffic behavior of mobile stations in a specific recording area (for instance, a given set of cells).

### Optimizing applications

The frequency allocation support (FAS) and frequency optimization expert (FOX) applications in GSM help operators to choose

### TRADEMARKS

Sun, Solaris and Java are registered trademarks of Sun Microsystems, Inc. in the United States and other countries.

Windows NT is a registered trademark of Microsoft Corporation.

which frequencies should be allocated to the cells in their GSM networks. FAS presents the results in reports, so that the operator can decide which frequency changes to implement, whereas FOX automatically finds, suggests and implements frequency changes that will improve network quality.

The neighboring cell support (NCS) and neighboring optimization expert (NOX) applications give operators a way of optimizing the neighboring cell lists for all cells in their GSM network. NCS and NOX help operators to add and remove cells from the neighboring cell lists that every GSM cell needs to have in order for handovers to work between cells. NCS presents the recorded measurements and statistics in several reports, which the operator can study to see which neighboring cell relations should be removed and added. NOX has the added functionality of suggesting which neighboring cells should be removed or added, and of implementing these changes automatically. Real-time performance events could also be an enabler for a wide range of optimization services, such as

- neighbor cell list optimization for combined GSM and WCDMA networks;
- optimization of power settings; and
- loadsharing between cells, frequencies and systems.

#### *Event-based statistics*

Ideally, to preserve flexibility, the software that manages the radio network should not also have to calculate statistics. As the radio network is enhanced (has more features added to it) it becomes more complex. The addition of a counter-based interface to the switch requires a massive addition of new counters—to keep network performance visible. The same results can be achieved with a small set of performance events. Events contain detailed information about individual mobile stations and can be correlated with network elements, cells, parts of cells, and even individual calls or phone types.

Extracting data from the traffic machine also gives operators the opportunity to gather and correlate data from more than one network element. This gives performance data on relations between cells served by different network elements and different systems,

such as GSM or WCDMA. New statistics can quickly be introduced and tailored to individual customers' needs.

The statistics can then be processed by the same storage, monitoring and presentation systems that are currently used to fulfill compatibility requirements. Event-based statistics allow for a flexible approach—operators can easily add new counters without having to update the software in the network elements.

#### *Inter-system performance management*

The event mechanisms in GSM and WCDMA are at the same level, which means that future products can support both systems simultaneously. Operators can thus generate statistics for a combined GSM/WCDMA network, including specific inter-system statistics, such as inter-system handover.

Moreover, operators will be able to adaptively tune inter-system aspects, such as inter-system neighbor cell lists.

#### *External applications*

The use of event data is not limited to the applications described above, but can also be used to create new functions that help operators to increase their revenue. The event data contains detailed information about the type and distribution of traffic in the network. When combined with other external data this can serve as the basis for services provided by the operator or a third party. One such example is road traffic information.

## Conclusion

This article describes an architecture and prototype based on real-time performance events for monitoring and optimizing cellular networks. The events-based architecture admits fast and flexible development of performance monitoring and optimization applications. It also provides open interfaces that enable customized performance monitoring.

Existing mechanisms in GSM and WCDMA products adhere to many of the architectural principles described, which in turn allow efficient monitoring of combined systems.

## REFERENCES

- 1 HYPERLINK3GPP TS 32.102 V4.2.0 (2001-09) 3G Telecom Management Architecture [ftp://ftp.3gpp.org/specs/2001-09/Rel-4/32\\_series/32102-420.zip](ftp://ftp.3gpp.org/specs/2001-09/Rel-4/32_series/32102-420.zip)
- 2 ETSI GSM Technical Specification 05.08