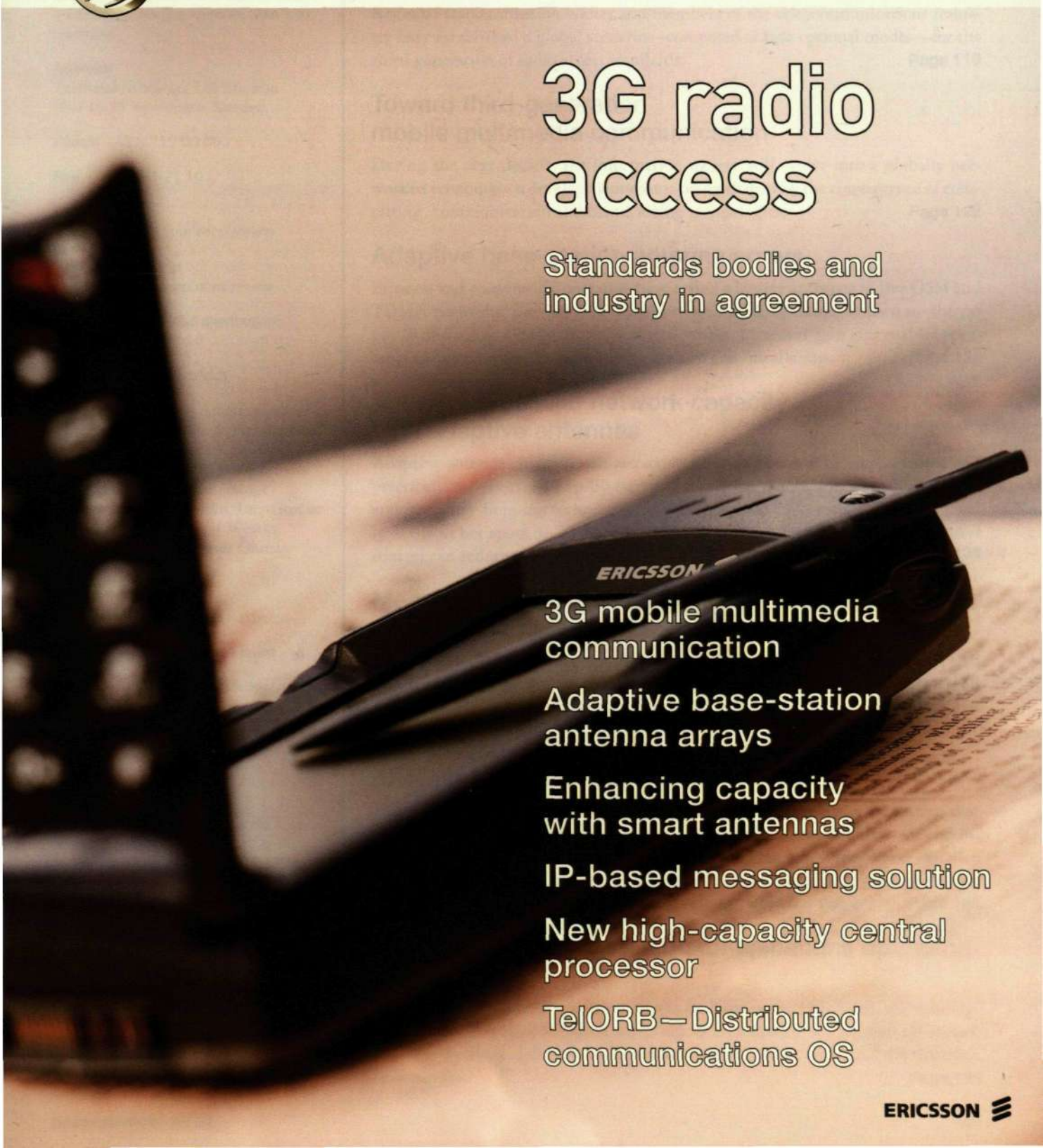
Ericsson
3/1999

REVIEW

THE TELECOMMUNICATIONS TECHNOLOGY JOURNAL

75 YEARS IN PUBLICATION



3G radio access

Standards bodies and industry in agreement

3G mobile multimedia communication

Adaptive base-station antenna arrays

Enhancing capacity with smart antennas

IP-based messaging solution

New high-capacity central processor

TelORB—Distributed communications OS

ERICSSON

Ericsson Review is also published on the Web. Now with a new look and improved search capabilities. Visit us at http://www.ericsson.com/Review.

Editorial

Bernt Ericson

Applied research

The role of research at Ericsson is to investigate new ideas, in order to understand their usefulness as applied to telecommunications. Promising ideas are prototyped to the extent that the development organization can integrate new findings into commercial products. Applied research of this kind relies heavily on input from the research community at universities and research institutes.

Academic research is often accused of being introverted and of little benefit to society. There are, however, many examples of how the needs of society and industry have inspired the basic research corps to expand theories into new fields—new-found knowledge gives rise to new products which inspire new research ideas, and so on.

I was recently invited to speak on the relevance of basic science at the "World Conference on Science" in Budapest. In my speech, entitled "Views from the electronics industry," I explained why and how cooperation between industry and academia can benefit each party. I began my speech with some historic highlights that have played a fundamental role in the evolution of telecommunications and the information society.

Historical highlights

Thomas Alva Edison was obsessed with the idea of turning electricity into light. After thousands of experiments with various materials, he succeeded, in 1880, when he perceived that the glowing carbon wire must be protected from oxygen in the air. Scientists who analyzed the light bulb found new phenomena that led to the discovery of the electron, in 1897. This addition to the pool of basic science led to the development of the diode, and later, in 1907, to the triode—the triode is needed to build an amplifier, which is an essential component in telecommunications.

Alexander Graham Bell filed the basic patent on telephony in 1876 (the same year in which Lars Magnus Ericsson founded Telefonaktiebolaget LM Ericsson). This new form of communication was as instant "hit" and quickly grew in use. Initially, the distances were bridged by open-air copper wires. Mother Nature, however, set a limit on how great a distance could be traversed (as explained by Ohm's law). Early trials to traverse the Atlantic Ocean failed because the required voltage was so great that the cables quickly broke. With the invention of the amplifier, however, it was possible to connect New York to San Francisco, in 1915, and at long last, in 1956, to inaugurate the first successful transatlantic cable.

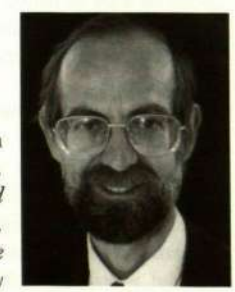
The triode also opened the era of electronic computers, although to build a practical machine, a veritable host of tubes was required. Not surprisingly, power consumption and heat dissipation were major issues, but the greatest limiting factor for operating and maintaining early computers was the high failure rate of individual tubes. The US military (which was quick to see the potential of automated computing) strongly urged and supported a better amplifying component. In 1949, the transistor was born when researchers at Bell labs found the answer they were searching for in semiconducting materials. Since then, engineers have steadily found new ways of integrating more and more transistors into the same silicon chip (a 400% increase every third year—an evolution that follows what has come to be known as Moore's law).

Moore's law

The integrated circuit plays a key role in most of Ericsson's products. Thanks to constant increases in performance, we have, at regular intervals, been able to release greater computing power in AXE. For instance, in this issue of Ericsson Review, you can read about the latest APZ 212 30 processor, which contains several highly integrated custom-designed circuits (the largest circuit is composed of 10 million transistors!).

In telecommunications, the need for greater capacity can to a large extent be traced to the rapidly growing use of wireless communication. As allocated frequency bands become saturated, new methods must be found that make optimum use of these sparse frequencies. The adaptive antenna (also discussed in this issue) is an example of one such method. Initially, the theory and first practical implementations of this technology were driven by the needs of the military. Today, the technology has matured and can be deployed in civilian systems.

According to Moore's law, we can expect increases in hardware performance that will allow most new functionality to be added through software. We must therefore have early access to, and become very adept at applying new technology, especially as relates to sotware. Consequently, we seek to establish good cooperation with academia and are constantly on the lookout for breakthrough research.



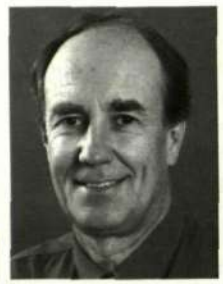
Bernt Ericson
Vice President, Research and Innovations, Ericsson Corporate Technology

Contributors

In this issue


Sören Anderson


Bengt Carlqvist


Anders Derneryd


Bo Hagerman


Lars Hennert


Per Holmberg


Björn Johannisson


Robert Lagerholm


Alexander Larruy


Janne Lundqvist


Mats Nilsson


Torbjörn Nilsson

Sören Anderson received an M.S.E.E. (1988) and Ph.D. (1992) in automatic control from the Linköping Institute of Technology, Sweden. In 1993, he was a postdoctoral research associate at Yale University. He joined the department for Radio Access and Antenna Systems Research at Ericsson Radio Systems, in 1994, where he conducted research on adaptive antennas in cellular systems and managed a research project for adaptive antennas for GSM. Since 1997, he has managed Ericsson's research of antennas and propagation.
soren.s.andersson@era.ericsson.se

Bengt Carlqvist works with Strategic Product Management and New Technologies at the GSM Radio Base Station Product Unit coordinating the introduction of adaptive antenna technology. He previously initiated the development of active antennas for single-carrier as well as multicarrier applications. He holds an M.S. in electrical engineering from the Royal Institute of Technology, Stockholm.
bengt.carlqvist@era.ericsson.se

Anders Derneryd joined Ericsson in 1978 to work on space and military projects in the Antenna Department at Ericsson Radar Electronics, Mölndal, Sweden. In 1992, he joined the antenna department at Saab Ericsson Space, and in 1995, he rejoined Ericsson Microwave Systems in Mölndal. In his current position, as Expert at Core Unit Antenna Technology, he works on novel array antenna concepts for mobile communication systems. He received an M.S. (1971) and a Ph.D. (1976) in electrical engineering from Chalmers University of Technology, Göteborg, and in June 1999, Dr Derneryd was appointed adjunct professor in antenna technology at Lund University, Lund, Sweden.
anders.derneryd@ericsson.com

Bo Hagerman received an M.S.E.E. (1987), Lic.Tech.E.E. (1993) and Ph.D. (1995) in radio communication systems from the Royal Institute of Technology, Stockholm. From 1987 to 1990, he was a member of the technical staff at the Ericsson Radio Systems R&D department, where he investigated signal processing as it applies to GSM receivers. In 1995, he joined the department of Radio Access and Antenna Systems Research at Ericsson Radio Systems, where he is currently employed in the research of adaptive antennas in cellular systems.
bo.hagerman@era.ericsson.se

Lars Hennert joined Ericsson in 1985 after graduating with a master of science degree in electrical engineering from the Royal Institute of Technology in Stockholm. Since 1990, he has worked with different aspects of TelORB and its predecessors.
lars.hennert@uab.ericsson.se

Per Holmberg is a specialist in computer architecture at Ericsson Utvecklings AB in Stockholm. Since joining Ericsson, in 1984, he has worked with the design and implementation of high-performance, high-availability and fault-tolerant computer systems for a number of applications, including aircraft, command and control systems, and data communication and telecommunication systems. During the APZ 212 30 project, he was responsible for the technical design of the instruction processor. He holds an M.S.E.E. from the Royal Institute of Technology in Stockholm.
per.holmberg@uab.ericsson.se

Nils Isaksson is currently employed as System Manager of AXE control system development at Ericsson Utvecklings AB, Stockholm. During the late 1970s, while working on early AXE design, he was a member of the research team that presented the original APZ 212 architecture and was responsible for the design of the control structure concept. He holds an M.S.E.E. from Chalmers University of Technology, Göteborg.
nils.isaksson@uab.ericsson.se

Nils Isaksson

Bo Svensson

Björn Johannisson worked on the development of the ERIEYE antenna at Ericsson Radar Electronics, Mölndal, Sweden, from 1987 until 1990, when he left Ericsson to work for Anaren Microwave, Syracuse, NY. In 1993, he returned to Ericsson Microwave Systems. He is currently a Senior Specialist in communication antennas at Core Unit Antenna Technology. In this role, he coordinates research on antennas at Ericsson Research. He holds an M.S. (1986) in engineering physics from Chalmers University of Technology, Göteborg.
bjorn.johannisson@ericsson.com

Robert Lagerholm, who joined Ericsson in 1994, is a product manager of third-generation radio base stations for TDMA systems. He holds a Lic.Tech. (1988) in microwave antennas from Chalmers University of Technology, Göteborg.
robert.lagerholm@era.ericsson.se

Alexander Larruy has worked with distributed systems since he joined Ericsson in 1982. Today, he manages the department for TelORB development at Ericsson Utvecklings AB. He holds an M.S. in physical engineering from the University of Uppsala.
alexander.larruy@uab.ericsson.se

Janne Lundqvist, who joined Ericsson in 1989, is currently in charge of the business development of IP-based messaging solutions at Ericsson Messaging Systems, Woodbury, NY. He has also worked as product manager in the same group. Before joining Ericsson Messaging Systems, he was the Director of Marketing, Mobile Networks, in Guangdong and Hainin, China. He holds an M.S. in industrial engineering and management from the Linköping Institute of Technology.
janne.lundqvist@ericsson.com

Mats Nilsson is responsible for Ericsson's strategies for future standards at Telefonaktiebolaget LM Ericsson, Corporate Technology. He joined Ericsson in 1987 to work as a systems engineer and project manager. From 1990 to 1992 he served as the technical manager of mobile systems at Nippon Ericsson KK. Between 1992 and 1998, he was responsible for Ericsson's technical strategies for future systems within the Mobile Systems business area. He holds a B.S. from the University of Uppsala and a Lic.Ph. in theoretical physics from Stockholm University.
mats.nilsson@lme.ericsson.se

Torbjörn Nilsson is presently Senior Vice President of Marketing and Strategic Business Planning at Telefonaktiebolaget LM Ericsson, a position he has held since 1998. From 1992 to 1998, he was Vice President of Strategic Business Development at Ericsson Mobile Systems. Before this, he was the Director of Product Development at Telefonaktiebolaget LM Ericsson. Mr. Nilsson began his career at Ericsson in 1978. He holds an M.S. in engineering from the Lund Institute of Technology and an MBA from Stockholm University.
torbjorn.nilsson@lme.ericsson.se

Bo Svensson is the chief architect of Ericsson's IP-based messaging solution. He joined Ericsson in 1985, and in 1989, began working with O&M-related aspects of AMPS/TDMA systems (CMS 88). In 1992, he transferred to Ericsson Research Canada, where he became an Expert in Operations Support Systems. In 1997, he moved to Ericsson Messaging Systems in Woodbury, NY. He holds an M.S. in electrical engineering from Chalmers University of Technology, Göteborg.
bo.svensson@ericsson.com

Third-generation radio access standards

Mats Nilsson

Second-generation radio access has been a major success story for the global telecommunications industry, delivering telephony and low bit-rate data services to mobile end-users. The growth rate of second-generation mobile telephony indicates that mobile communication is well on its way toward full mass-market penetration. Thanks to the tremendous growth of the Internet, multimedia is also penetrating the mass market at an explosive pace. Combined, the digital cellular footprint and the multimedia services of the Internet form the basis of tomorrow's integrated wireless communication and multimedia access system.

The transition to third-generation capabilities must be based on a feasible migration path that defines a way of integrating multimedia, packet switching and wideband radio access into the dominating second-generation systems of our day. Although the standards behind these systems were initially defined on a regional basis, standardization bodies and members of the telecommunications industry have finally agreed on a harmonized global scenario for third-generation radio access standards.

The author describes this newly harmonized global scenario, and how the developing third-generation radio access standards will cater for requirements for multimedia and flexibility, spectrum allocation, and successful migration from second-generation systems.

Second-generation standards

Various digital cellular standards were developed in several regional standards bodies during the late 1980s and early 1990s. The first-generation standards had been developed some ten years earlier.

GSM

The development of the new pan-European digital cellular standard got under way in 1985. GSM has since evolved into the leading global second-generation standard, in terms of number of subscribers and area of coverage (Figure 2).

GSM is an eight-slot, time-division multiple-access (TDMA) system with 200 kHz carrier spacing. In terms of service, GSM is mobile ISDN (integrated services digital network), with support for a wide variety of services. Intelligent network (IN) support in the mobile environment has also been defined for GSM—for example, the virtual home environment—as well as many advanced data services. And today, thanks to general packet radio services (GPRS), packet access can also be integrated into GSM.

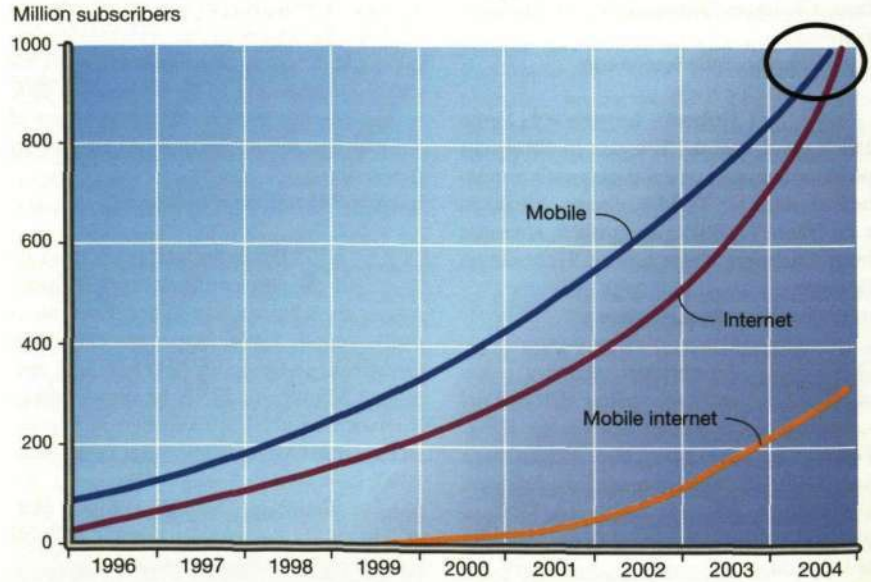
TDMA/136

The TDMA/136 specification, which was defined in the USA, in 1988, by the Telecommunications Industry Association (TIA), was developed with the aim of digitizing the analog advanced mobile phone service (AMPS). To maintain compatibility with AMPS, the TDMA specification stipulates 30 kHz carrier spacing in a three-slot TDMA solution.

PDC

The development of the personal digital cellular (PDC) specification was drafted by the

Figure 1
Growth of mobile telephony and the Internet.



RCR (1990), which later became the Association of Radio Industries and Broadcasting (ARIB). To ensure compatibility with the Japanese analog systems, a carrier spacing of 25 kHz was maintained in a three-slot TDMA solution.

CDMA/IS-95

The narrowband code-division multiple-access (CDMA) IS-95 specification stipulates 1.25 MHz carrier spacing for telephony services. TIA began defining this specification in 1991.

Each of the second-generation standards essentially defines a mobile telephony system—that is, a system that provides mobile end-users with circuit-switched telephony services. Apart from voice services, these systems support supplementary services and some low-bit-rate data services.

Drivers of third-generation development

Much as second-generation radio access brought mobile telephony capabilities to the mass market, third-generation radio access will introduce value that extends beyond basic telephony. The widespread growth of the Internet has created a mass-market base for multimedia and information services. The challenge is to merge mobile telephony coverage and the associated user base with the Internet and other multimedia applications. To successfully meet this challenge, third-generation radio access must provide

- flexible multimedia management;
- Internet access;
- flexible bearer services; and
- cost-effective packet access for best-effort services.

Most new multimedia services will be offered via the Internet. Therefore, a characteristic feature of third-generation radio access is that it provides mobile Internet. Multimedia requires considerable flexibility—that is, the cost-effective ability to support different bearer services with very different requirements, such as different bit rates (constant or variable), real-time or best-effort service, and packet- or circuit-switched service.

In addition, third-generation radio access must provide full-area coverage (same as second-generation voice service); high-peak bit-rate services (384 kbit/s full-area coverage, 2 Mbit/s local coverage); and any kind of service mix. Finally, third-generation

BOX A, ABBREVIATIONS

3GPP/3GPP2	Third-generation (3G) Partnership Project
8PSK	Eight-phase shift keying
AMPS	Advanced mobile phone service
ANSI-41	
ARIB	Association of Radio Industries and Broadcasting
BSS	Base station subsystem
CC	Call control
CDMA	Code-division multiple access
cdma2000	
CDMA/IS-95	Digital cellular standard IS-95
CN	Core network
CODIT	Code-division testbed (RACE II mobile project R2020)
DECT	Digital enhanced cordless telecommunications
DS	Direct sequence
ECSD	Enhanced circuit-switched data
EDGE	Enhanced data rates for GSM and TDMA/136 evolution
EGPRS	Enhanced GPRS
ETSI	European Telecommunications Standards Institute
FDD	Frequency-division duplex
FMA2	FRAMES Multiple Access 2
GMSK	Gaussian minimum-shift keying
GPRS	General packet radio services
GPS	Global positioning system
GSM	Global system for mobile communication
HCS	Hierarchical cell structure
IMT-2000	International mobile telecommunication
IN	Intelligent network
IP	*Internet protocol*
ISDN	Integrated services digital network
ITU	International Telecommunication Union
ITU-R	ITU Radio Communication Sector
ksps	Kilosymbols per second
MC	Multicarrier
Mcps	Megachips per second
MM	Mobility management
OHG	Operators Harmonization Group
PCS	Personal communication services
PDC	Personal digital communication
RACE	R&D in advanced communications technologies in Europe
RCR	Research & Development Center for Radio Systems (Japan)
RLC	Radio link control
RNC	Radio networkcontroller
SIR	Signal-to-interference ratio
TDD	Time-division duplex
TDMA	Time-division multiple access
TDMA/136	Digital cellular standard IS-136
TD-SCDMA	Time-division synchronous CDMA
TFCI	Transport format combination indicator
TIA	Telecommunications Industry Association
TPC	Transmit power control
UMTS	Universal mobile telecommunications system
UTRA	UMTS terrestrial radio access
UTRAN	UTRA network
UWCC	Universal Wireless Communications Consortium
VoIP	*Voice-over-IP*
WBTB	Wideband Test Bed
WCDMA	Wideband CDMA

BOX B, DIGITAL CELLULAR STANDARDS, END OF JUNE 1999

Standard	Subscribers	Countries/Network on air	Monthly growth (approximate)
GSM	183.3 million	120/284	7.6 million
PDC	42.3 million	1/30 (Japan)	0.6 million
TDMA	24.3 million	34/104	1.4 million
CDMA	31.5 million	12/31	1.5 million

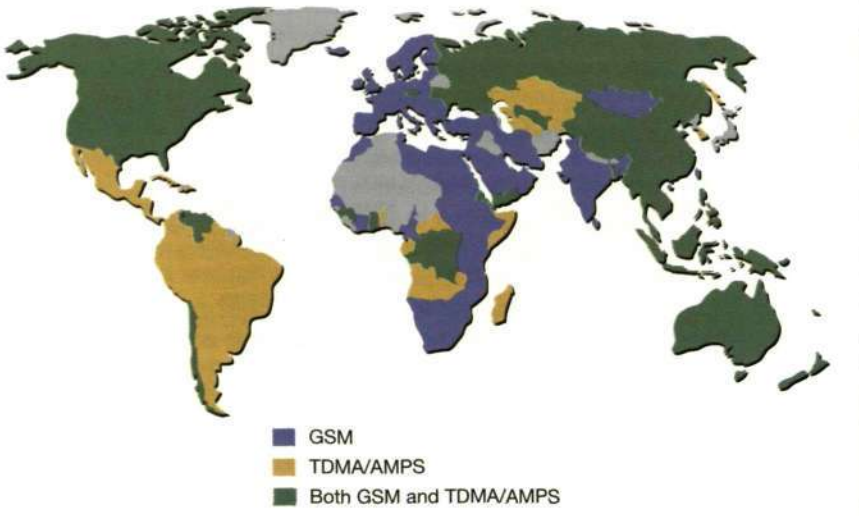


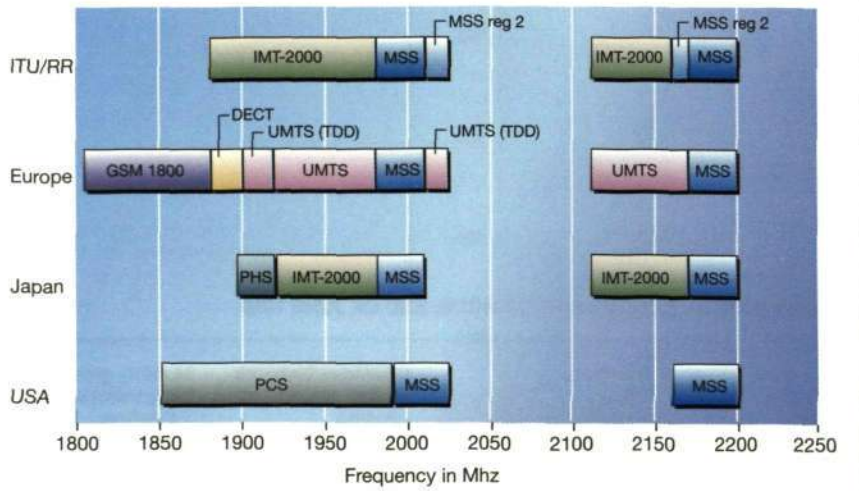
Figure 2
Examples of second-generation mobile radio coverage.

radio access must use the radio spectrum and network resources in a cost-effective fashion.

Migration

To succeed, the third-generation standards must facilitate efficient migration from second-generation radio access. The introduction of multimedia into mobile communication will proceed gradually over time. Thus, a step-by-step migration plan must be catered for that begins with the state of present-day second-generation systems.

Figure 3
Spectrum allocation.



Given that there are four separate second-generation standards (GSM, TDMA, PDC, and IS-95), different migration paths must be offered.

Spectrum allocation

Spectrum is allocated differently in different parts of the world. Moreover, the availability of spectrum varies greatly from operator to operator. In many regions of the world, new spectrum is to be allocated within parts of the 2 GHz IMT-2000 frequency band defined by the International Telecommunication Union (ITU). According to the ITU's recommendation, Europe will allocate 1920-1980 MHz and 2110-2170 MHz for the operation of frequency-division duplex (FDD), and 1900-1920 MHz and 2010-2025 MHz for the operation of time-division duplex (TDD). In Japan, an identical allocation has been made for the operation of FDD, but no allocation has been made for the operation of TDD. The allocation of spectrum in the USA differs from that of Europe and Japan (Figure 3), since parts of the 2 GHz frequency band have already been allocated for use by personal communication services (PCS) systems.

Although spectrum has been reserved in certain parts of the world for IMT-2000 services, this does not mean that similar services cannot be provided in other bands. For instance, EDGE (which is a migration path for GSM and TDMA/136) and multicarrier cdma2000 (which is a migration path for IS-95) support the majority of IMT-2000 services. Accordingly, we can expect to see the following developments in the market:

- operators will be allocated new spectrum, either paired (FDD) or unpaired (TDD) bands; and
- operators will migrate existing second-generation spectrum, adding support for third-generation services.

In summary, the third-generation standards must effectively cater for requirements for multimedia and flexibility, second-generation to third-generation migration, and spectrum allocation.

Family of harmonized third-generation standards

Throughout the past decade, the ITU Radio Communications Sector (ITU-R) has elaborated on a framework for global third-generation standards. At the same time,

since the early 1990s, the industry has been actively researching third-generation radio access.

When the ITU-R issued a call for proposals, in 1998, ten terrestrial candidates were submitted. Although derived from different standards bodies, several of the candidates were quite similar.

The European Telecommunications Standards Institute (ETSI) responded with the UMTS terrestrial radio access (UTRA) interface (for IMT-2000).

In parallel to the wideband CDMA (WCDMA) activities in Europe, extensive work on third-generation WCDMA was being carried out in Japan, Korea, and the USA. The standardization bodies in these countries each submitted their own variants of WCDMA as candidates for IMT-2000. It should be noted, however, that considerable cooperation and coordination took place between the WCDMA proponents; therefore, in the end, the four variants of WCDMA were more or less identical. Today, there is only one WCDMA standard, since the regional standardization bodies—ETSI (Europe), T1P1 (USA), ARIB/TTC (Japan) and TTA(Korea) have joined forces in the 3G Partnership Project (3GPP).

The ITU also received other CDMA proposals:

- the cdma2000 specification proposed by standardization bodies in the USA and Korea (cdma2000 proposals contained two modes: a direct-spread mode across the entire 5 MHz band, and a multicarrier mode with three IS-95 carriers in a multicarrier format in the downlink); and
- the UWC-136 specification (EDGE and a wideband TDMA mode) proposed by TIA (USA).

In total, the ITU received three families of FDD proposals (WCDMA, cdma2000, and UWC-136) and three TDD proposals (UTRA/TDD, TD-SCDMA, and DECT—Europe, China and Europe respectively).

Since submitting these proposals to the ITU, the industry and standards bodies have coordinated their efforts to harmonize the IMT-2000 candidates and arrive at a smaller set of third-generation standards.

The most recent harmonization activity was initiated by the Operators Harmonization Group (OHG), a group of major operators from all parts of the world who operate different variants of second-generation systems (GSM, PDC, IS-136 and IS-95). The focus of this group's discussions has been on CDMA-based third-generation systems: UTRA FDD and TDD, cdma2000 (direct-spread and multicarrier) and TD-SCDMA. At its first meeting, the OHG set the framework for third-generation standards:

1. There must be three modes of CDMA-based radio interfaces:
   - direct-spread CDMA (DS);
   - multicarrier CDMA (MC); and
   - time-division duplex (TDD).
2. There must be two kinds of core network:
   - GSM-based core networks, including circuit-switched GSM and packet-switched GPRS; and
   - ANSI-41-based core networks, including circuit- and packet-switched networks.
3. All radio-access network modes must be able to connect to either type of core network (Figure 4).

In the spring of 1999, the OHG invited major manufactures in the telecommunications industry to join the debate, and in May 1999, the OHG and participating manufactures concluded the harmonization discussion at a meeting in Toronto, Canada, agreeing on the following CDMA-based third-generation systems:

- The direct-spread CDMA mode must be based on WCDMA (as specified by 3GPP), with some minor modifications

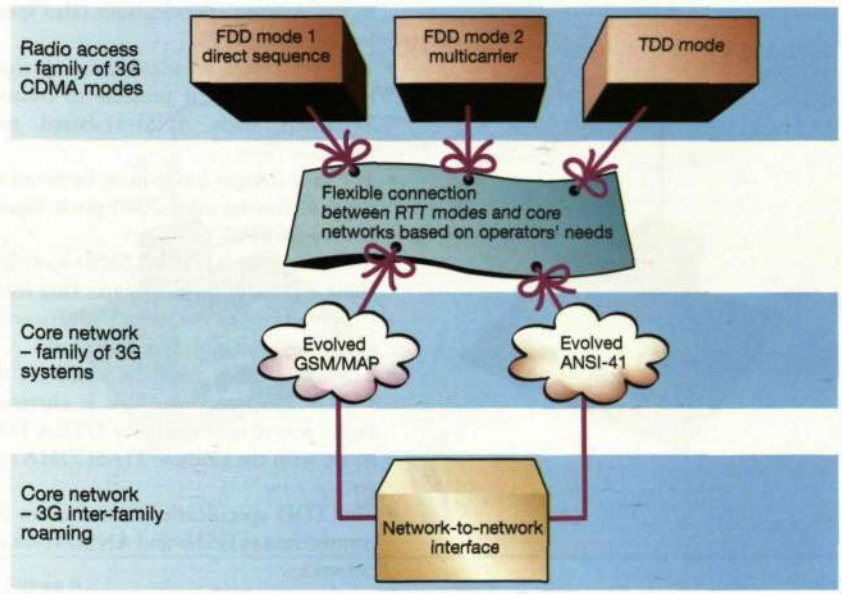


Figure 4
OHG framework of the third-generation harmonization activity for different CDMA modes.

to the existing specifications (also specified by the 3GPP).

- The WCDMA standard must include hooks that make it possible to connect WCDMA with ANSI-41-based networks.
- The multicarrier mode must be based on the multicarrier cdma2000 mode (specified by 3GPP2).
- The multicarrier cdma2000 specifications must also include hooks that make it possible to connect multicarrier cdma2000 to GSM-based networks.
- The TDD mode must be based on the UTRA TDD mode—work is currently under way to harmonize the UTRA TDD mode with the Chinese TD-SCDMA system.
- The TDD specifications must allow for connections to GSM- and ANSI-41-based networks.

In summary, the harmonization activities resulted in a single third-generation CDMA standard with three modes: a direct-sequence mode based on WCDMA, a multicarrier mode based on cdma2000, and a TDD mode based on UTRA TDD, plus one third-generation TDMA standard being developed for EDGE/UWC-136. At present, the 3GPP and 3GPP2 standardization projects are working out the final details of the different third-generation CDMA modes with the intention of delivering the first complete specifications by the end of 1999.

Harmonized family of four

For operators to provide third-generation services, the family of radio access standards must cater for every spectrum and migration scenario. To achieve this goal, the main task of harmonizing third-generation radio-access systems has been to minimize the number of different air-interface techniques.

In keeping with early proposals from Ericsson, the industry and global standardization processes have now established a third-generation radio-access scenario, which we call the harmonized family of four.

In brief, the spectrum and migration criteria give four main categories of incompatible requirements for radio-access standards:

1. New, or modified, spectrum in paired duplex bands for areas covering third-generation mobile services. This category requires a new, flexible, radio-access technique which has been optimized for multimedia and which can most efficiently utilize new or modified spectrum bands (at least 2 · 5 MHz, preferably 2 · 15-20 MHz per operator). The technique best suited for this is the harmonized WCDMA solution being developed by the 3GPP.
2. New, or modified, unpaired spectrum. For unpaired bands, TDD is the only feasible option.
3. TDMA spectrum migration solution. EDGE, which was defined to support the migration of TDMA bands (either GSM or TDMA/136) toward third-generation mobile communication, introduces high-level modulation and link-adaptation techniques that significantly boost the bit-rate capabilities of TDMA while using the same frequencies. In combination with time-slot aggregation and packet switching, EDGE becomes an especially effective tool for migrating third-generation radio access into TDMA systems (GSM or TDMA/136).

BOX C, GLOBAL THIRD-GENERATION STANDARD WITH FOUR OPTIONAL MODES

Wideband CDMA			**TDMA**
Direct sequence	*Multicarrier*	*TDD*	*EDGE/UWC-136*
•WCDMA as per 3GPP	•cdma2000 as per 3GPP2	•As per 3GPP, harmonized with Chinese TD-SCDMA	•As per ETSI/UWC
•New spectrum •FDD •Chip rate 3.84 Mcps •Asynchronous (synchronous operation supported)	•IS-95 spectrum overlay •FDD •Chip rate 3.6864 Mcps •Synchronous	•Unpaired spectrum operation (TDD) •Chip rate 3.84 Mcps	•Existing spectrum 200 kHz TDMA •high-level modulation •with link adaptation
•Network signaling – phase 1 GSM/MAP – phase 2 ANSI-41	•Network signaling – phase 1 ANSI-41 – phase 2 GSM/MAP	•Network signaling – phase 1 GSM/MAP – phase 2 ANSI-41	•Network signaling – GPRS-based for both GSM/MAP and ANSI-41

4. CDMA/IS-95 spectrum migration solution. To introduce wideband services into the existing IS-95 spectrum, a spectrum overlay solution utilizes either one 1.25 MHz carrier (1X) or three 1.25 MHz carriers (3X) for the downlink. This spectrum-overlay strategy enables second-generation and third-generation services to coexist on the same frequencies.

Direct sequence

WCDMA has been chosen as the basic radio-access technology for UMTS/IMT-2000 in all major areas of the world. Apart from high-bit-rate services, the WCDMA radio interface offers significant improvements over second-generation narrowband CDMA, including

- improved coverage and capacity, thanks to greater bandwidth and improved coherent uplink detection;
- support for inter-frequency handover, which is necessary for large-capacity hierarchical cell structures (HCS);
- support for capacity-enhancing technologies, such as adaptive antennas and multiuser detection; and
- a fast and efficient packet-access protocol.

UTRA includes FDD and TDD modes. The FDD mode is based on pure WCDMA, whereas the TDD mode includes an additional TDMA component according to the TD/CDMA proposal. This article focuses on the pure WCDMA-based FDD mode (UTRA/FDD).

Background to WCDMA

Extensive research on WCDMA has been carried out in Europe for almost a decade. A WCDMA concept that fulfilled the third-generation requirements was first developed in the RACE Code-division Testbed (CODIT) project (1992-1995). The CODIT concept also served as the basis for the hardware test beds that were used to evaluate and verify the performance of WCDMA technology. One example of hardware fashioned after the CODIT concept is the Ericsson Wideband Test Bed (WBTB).

The WCDMA technology was further refined into the FRAMES Multiple Access 2 (FMA2) concept developed within the FRAMES project. In March 1997, the FMA2 concept was submitted to ETSI as a candidate technology for UTRA. Within ETSI, the FMA2 proposal was then merged with other WCDMA proposals into the ETSI Alpha concept. Finally, in January 1998, ETSI selected the WCDMA-based Alpha concept as the main technology for UTRA FDD.

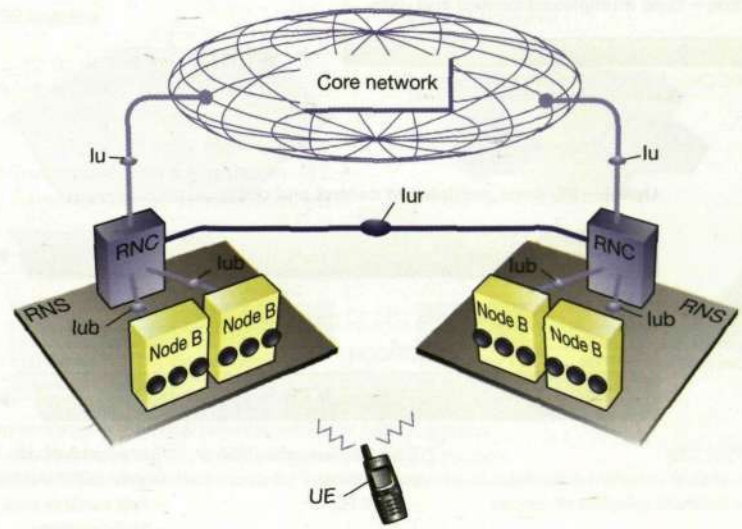


Figure 5
UMTS terrestrial radio access network (UTRAN).

Work on further enhancements and refinements of UTRA continued within ETSI. The aim of this work was to have a complete description of the radio interface drafted by the end of 1998. Since early 1999, WCDMA standardization has continued within the 3GPP, whose task is to specify a third-generation system with the UTRA radio-access network connected to a GSM-based core network.

WCDMA description

A close examination of the radio parts of WCDMA shows that several interfaces and functional splits must be standardized (Figure 5). A comparison of the WCDMA architecture with that of GSM shows that the UTRA network (UTRAN) corresponds to the base station subsystem (BSS). Likewise, the Iu interface between UTRAN and the core network corresponds to the A-interface in GSM (Gb interface in GPRS). The Iub interface corresponds to the A˝ interface. The Iur interface between the radio network controllers, which supports soft-handover functionality, is new. Finally, we have the air interface between the base station and the terminal (nodes B and UE in Figure 5).

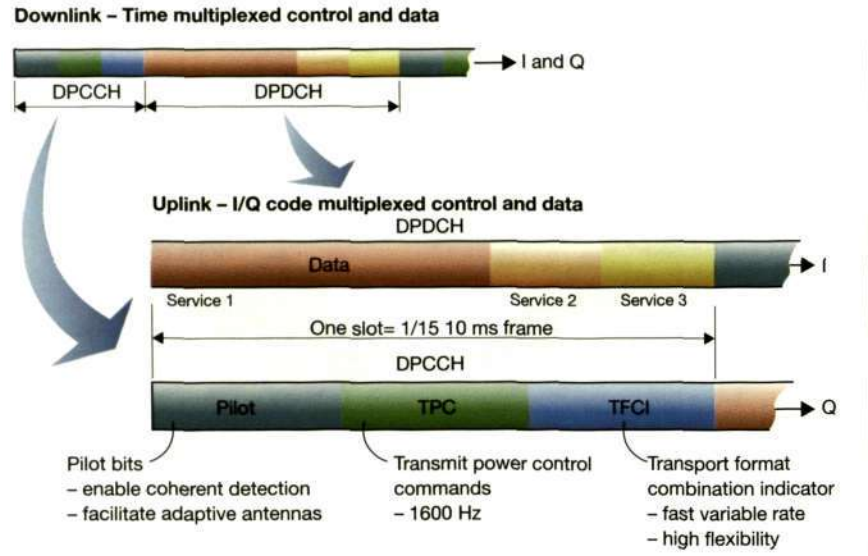


Figure 6
Structure of WCDMA physical layer.

Proceeding from the premise that a radio interface is to be based on direct-sequence WCDMA, designers can build a system whose properties fulfill the third-generation requirements. The key properties emphasized in WCDMA are improved performance, a high degree of service flexibility, and a high degree of operator flexibility. By improved performance we mean improved capacity, improved coverage, and third-generation services from second-generation deployment. By service flexibility we mean support of a wide range of services with maximum bit rates exceeding 2 Mbit/s, the ability to provide multiple services on one connection, and a fast and efficient packet-access scheme. And by operator flexibility we mean support for asynchronous inter-base-station operation (which makes for easier deployment in many environments), efficient support for various deployment scenarios, including hierarchical cell structures at hot-spots, and support for evolutionary technologies, such as adaptive antenna arrays and multi-user detection.

As was noted above, the WCDMA standard must contain necessary hooks for future performance enhancements, such as the introduction of smart antenna solutions. This is supported through user-specific pilot information, which is used for directing antenna beams to the end-user (Figure 6).

To facilitate channel estimation in the terminal, each connection uses a different antenna pattern, which is part of the channel. Other important details in the physical layer are the transmit power control (TPC) commands, and the transport format combination indicator (TFCI). Note: Although it is only necessary for the uplink, the TPC is included in both the uplink and the downlink, since it improves downlink performance. The TFCI facilitates support for fast variable rate, giving improved flexibility.

Another important feature is support for hierarchical cell structures (Figure 7). In an HCS scenario, the two cell layers operate on two separate frequencies. In CDMA, transmission and reception are usually continuous without time for inter-frequency measurements. This implies that measurement support is not provided for handover decisions. To overcome this problem, a compressed mode feature has been introduced. By means of this technique, data is compressed—basically by doubling the power and reducing the spreading factor by one half—which creates a free gap in time for making measurements.

Hooks and extensions

The initial WCDMA specifications from the 3GPP (prior to the OHG agreement) called for a CDMA system with a chip rate of 4.096 Mcps time-multiplex pilot information for all physical channels. However, the OHG agreement stipulates that the DS mode (WCDMA) should have a chip rate of 3.84 Mcps, to allow for easier implementation of DS and MC dual-mode terminals. The argument is that if the chip rate of the two modes differs by less than 5%, then it will be easier to use the same radio frequency chain in both modes.

The OHG also stipulates that the DS mode should have a common code-multiplex pilot. For cells that lack smart antennas or do not use sophisticated transmit diversity schemes, the common code-multiplex pilot is said to outperform the common time-multiplex pilot.

These modifications (chip rate and common pilot) have been incorporated into the most recent drafts of the 3GPP specifications.

As mentioned earlier, it must be possible to connect the WCDMA radio interface to an ANSI-41 network. Figure 8 shows how this is accomplished in accordance with the OHG agreement.

The connection of WCDMA to a GSM network or to an ANSI-41 network should use the radio protocols defined in 3GPP. Until now, however, 3GPP has only defined a WCDMA-to-GSM core-network connection. Consequently, some hooks (modifications of) and extensions (additions) to the WCDMA radio protocols may be required to guarantee that all services and functionality in an ANSI-41 network can be used together with the WCDMA radio interface. These hooks and extensions have been identified, and the work of specifying them is well under way in the 3GPP and 3GPP2 (completion is expected in March 2000). The hooks to WCDMA protocols must be incorporated into the 1999 draft of the WCDMA specifications; the extensions can be added to subsequent drafts. Call control (CC) and mobility management (MM) are taken from GSM when connecting to a GSM network, and from IS-634 when connecting to an ANSI-41 network.

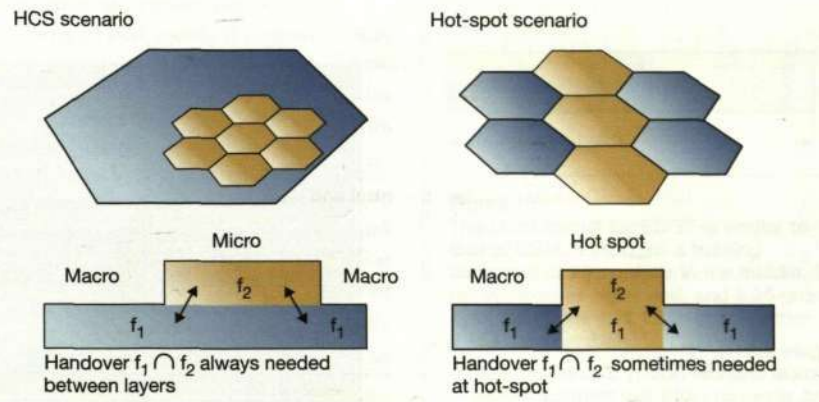


Figure 7
Support for HCS (inter-frequency handover).

EDGE

Two of the major second-generation standards, GSM and TDMA/136, have built the foundation from which to offer a common global radio access for data services. By exploiting a common physical layer (EDGE), each standard follows the same migration path toward providing third-generation services.

EDGE is currently the subject of standardization, the first phase of which will be finalized by the end of 1999, followed by a second phase in the year 2000. Compared to the data services currently available from GSM and TDMA/136, EDGE provides significantly higher user bit rates and spectral efficiency.

EDGE was first proposed to ETSI in the beginning of 1997, as a means of evolving GSM. EDGE reuses the GSM carrier bandwidth and time-slot structure and provides an efficient way of increasing bit rates, thereby facilitating the evolution of existing cellular systems toward third-generation capabilities.

In developing its third-generation wireless technology, the TDMA/136 community opted to base its proposal on the evolution of second-generation systems. In January 1998, the Universal Wireless Communications Consortium (UWCC) adopted EDGE as the outdoor component of the 136 High Speed radio interface, to provide 384 kbit/s data services. Arguments in favor of this approach are that the technology applies to both GSM and TDMA/136 systems and that it paves the way for global roaming. EDGE has since been developed concurrently by ETSI and the UWCC to guarantee a high degree of synergy for GSM and TDMA/136 alike.

The roadmap for EDGE standardization shows two phases. Initially, emphasis will be placed on enhanced GPRS (EGPRS) and

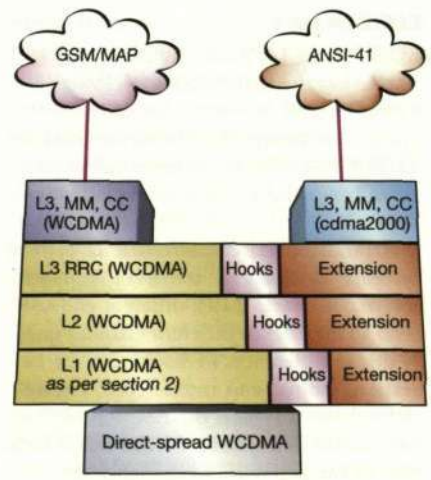


Figure 8
Connection of WCDMA to GSM and ANSI-41 networks. The words "Hooks" and "Extensions" indicate modifications and additions respectively.

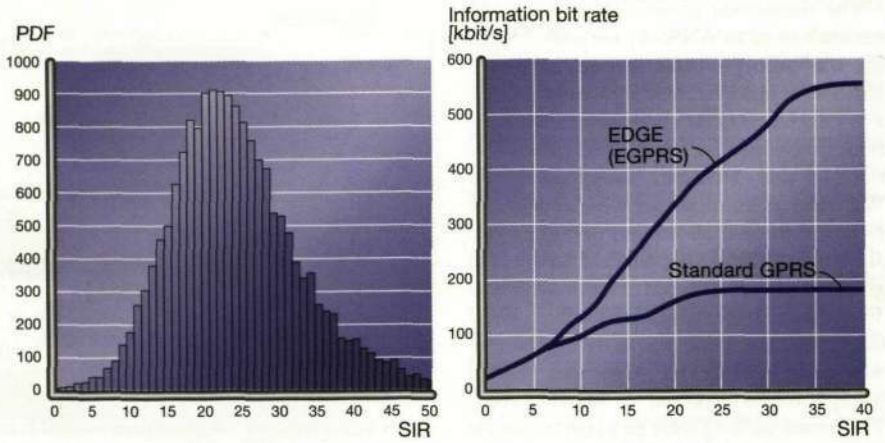


Figure 9
Typical channel quality distribution (left) and, given an eight-slot terminal (right), perceived user quality in terms of bit rate for EDGE and standard GPRS.

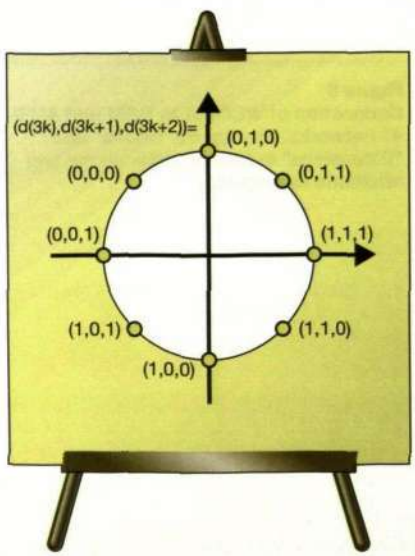


Figure 10
8PSK modulation: signal constellation in the phase plane.

enhanced circuit-switched data (ECSD). ETSI aims to release standards for these technologies in 1999. Products will appear shortly afterward. The second phase of standardization, which is targeted for release in 2000, will define improvements for multimedia and real-time services. Other objectives will include the alignment of services with the universal mobile telecommunications system (UMTS) and evolution toward a radio-access network which is independent of the core network (CN) and which will allow EDGE and UMTS to share a common core network.

EDGE phase I

EDGE phase I is primarily concerned with improving the radio interface. However, in a more general context, it can also be viewed as a system concept that allows the GSM and TDMA/136 radio access networks to offer a set of new radio access bearers to their core networks.

A fundamental characteristic of cellular systems is that, in terms of signal-to-interference ratio (SIR), different end-users tend to experience different channel quality due to differences in distance to the base station, fading, and interference. Figure 9 (left) shows a typical distribution of channel quality in a system. Notice that a large part of the population experience excessive SIR; that is, excellent channel quality from which they cannot benefit using traditional service, such as voice. Thus, we see that there is room for improving spectral efficiency.

Basic parameters of the radio interface

The EDGE air interface is intended to facilitate higher bit rates than those currently achievable in present-day cellular systems. To increase the gross bit rate, linear high-level modulation is introduced. The eight-phase shift keying (8PSK) modulation scheme was selected for its high data rates, high spectral efficiency and moderately complex implementation (Figure 10). The Gaussian minimum-shift keying (GMSK) modulation defined in GSM is also part of the EDGE system concept. The symbol rate of each modulation is 271 ksps, which yields a gross bit rate per time slot (including two stealing bits per burst) of 22.8 kbit/s for GMSK and 69.2 kbit/s for 8PSK. The 8PSK pulse shape is linearized GMSK, meaning that 8PSK fits into the GSM spectrum mask.

Many EDGE physical-layer parameters are identical to those of GSM. The carrier spacing is 200 kHz, and the TDMA frame structure of GSM is unchanged. Moreover, the 8PSK burst format is similar: a burst includes a training sequence of 26 symbols in

the middle, 3 tail symbols at either end and 8.25 guard symbols at one end. Each burst carries 2 · 58 data symbols, each of which is composed of three bits (Figure 11).

Radio protocol design

When possible, the EDGE radio protocol strategy reuses the protocols of GSM and GPRS. This minimizes the need for implementing new protocols. However, due to the higher bit rates and to new insights obtained through research, some protocols have been modified to optimize performance. The EDGE concept includes a packet-switched mode (EGPRS) and a circuit-switched mode (ECSD).

Packet-switched transmission—EGPRS

The current GSM/GPRS standard supports data rates from 11.2 to 22.8 kbit/s per time slot. By contrast, EGPRS will allow data rates from 11.2 to 59.2 kbit/s per time slot, which in a multislot configuration yields a data rate well over 384 kbit/s.

Due to the higher bit rate and the need for adapting the data protection to channel quality, the EDGE radio link control (RLC) protocol is somewhat changed from the corresponding GPRS protocol. The main changes involved improving the link quality control scheme. As mentioned, above, link quality control is a common term for techniques that adapt the robustness of the radio link to varying channel quality. Examples of link quality control techniques are link adaptation and incremental redundancy.

A link-adaptation scheme regularly estimates link quality and selects the most appropriate modulation and coding scheme for coming transmissions, in order to maximize the user bit rate. Incremental redundancy is another way of coping with variations in link quality. According to this scheme, information is initially sent with very little coding, which if decoding is immediately successful, yields a high bit rate. If decoding fails, additional coded bits (redundancy) are sent until the decoding succeeds. More coding means lower bit rate and greater delay.

EGPRS supports a combined link-adaptation and incremental-redundancy scheme, in which the initial code rate for incremental redundancy is based on measurements of link quality. The benefits of this approach are robustness and high throughput of the incremental redundancy operation in combination with shorter delays and reduced memory requirements.

Circuit-switched transmission—ECSD

The current GSM standard supports transparent and non-transparent radio access bearers. Eight transparent bearers are defined, offering constant bit rates in the range of 9.6 to 64 kbit/s.

A non-transparent bearer employs a radio link protocol to ensure virtually error-free data delivery. Eight bearers offer maximum user bit rates ranging from 4.8 to 57.6 kbit/s. The actual user bit rate may vary according to channel quality and the resulting rate of retransmission.

The introduction of EDGE does not affect the bearer definitions; that is, the bit rates remain unchanged. However, the way in which the bearers are realized, in terms of channel coding schemes, is new. For example, with EDGE, a 57.6 kbit/s non-transparent bearer can be realized with two time slots, whereas with standard GSM (using the TCH/F14.4 coding scheme) the same bearer requires four time slots. Thus, EDGE circuit-switched transmission yields high-bit-rate bearers using fewer time slots.

EDGE phase II

The main focus of EDGE phase II will be on real-time services delivered via the Internet protocol (IP); for example, voice over IP (VoIP). The introduction of these services will have an impact on radio access, the system architecture, and the core network. A single core network for UTRA and EDGE radio access will thus evolve.

Multicarrier

Originally, the cdma2000 specification contained the variants, 1X, 3X and direct-spread. After the harmonization work has been concluded, the global direct-spread multicarrier will be based on WCDMA, thereby putting an end to work on cdma2000 DS. In essence, 1X and 3X are specified in such a way that it is possible to add a spectrum-overlay solution.

To utilize spectrum overlay on current IS-95 services, the solution must reuse much of the IS-95 lower layers. The multicarrier (1X, 3X) signal must also maintain orthogonality to the underlying IS-95 carriers in the downlink. A 1.25 MHz carrier is used for 1X. Notwithstanding, several aspects of 1X have been enhanced to better suit third-generation services.

The use of 3X triples the bandwidth capabilities of reaching the highest bit rates envisaged for third-generation radio access.

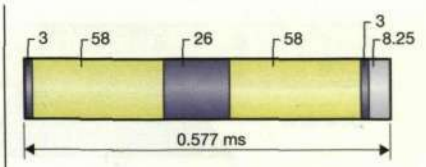


Figure 11
The burst format for EDGE is similar to that of GSM. It includes a training sequence of 26 symbols in the middle, 3 tail symbols at either end, and 8.25 guard symbols at one end. Each burst carries 2 · 58 data symbols.

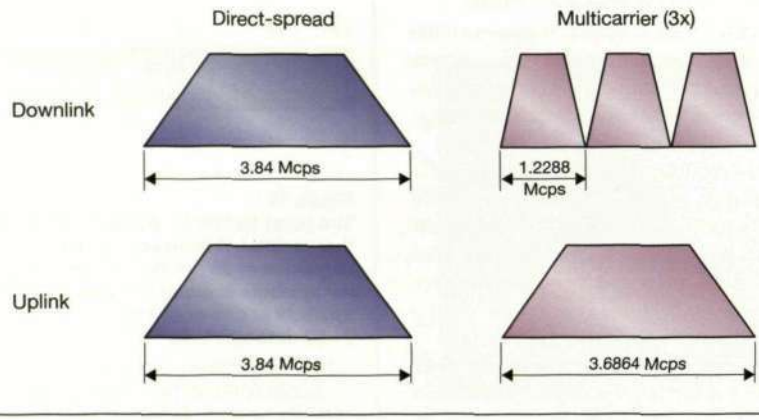


Figure 12
Overlaying in spectrum.

This is accomplished by multicasting the wideband signal in three 1.25 MHz carriers on the downlink. To avoid unnecessary terminal complexity associated with transmitting three carriers, direct spread is also used for the 3X uplink (Figure 12).

In contrast to the direct-spread mode, the multicarrier mode does not use a wideband CDMA carrier on the downlink. Instead, data is converted from serial to parallel and transmitted in several parallel narrowband CDMA carriers, each of which has the same characteristics as a second-generation IS-95 carrier.

Multicarrier downlink transmission facilitates the overlay of a third-generation system on top of a second-generation IS-95 system. Although slightly less efficient than the direct-spread mode, the multicarrier mode permits IS-95 operators with limited spectrum to offer third-generation services (operators who cannot exclusively allocate a part of their spectrum to third-generation services).

The uplink of the multicarrier mode is similar to the uplink of the direct-spread mode; that is, it uses a wideband carrier. However, compared to the direct-spread mode, the uplink of the multicarrier mode uses a somewhat lower chip rate in order to match the chip rate of the multicarrier downlink.

The 1X mode is evolved narrowband IS-95 with 1.25 MHz bandwidth. Regular operation of the multicarrier mode calls for three carriers in parallel (3X), which corresponds to a bandwidth of approximately 3.75 MHz. However, other modes may also be standardized (6X, 9X, 12X) with corresponding increases in bandwidth. Because each carrier of the multicarrier downlink should have the same characteristics as an IS-95 downlink carrier, tight intercell synchronization is required, typically by means of global positioning satellite (GPS) reception. This does not apply to the direct-spread mode, which is based on asynchronous WCDMA.

The direct-spread mode consists of wideband CDMA carriers on the uplink and downlink. The multicarrier mode consists of a wideband CDMA carrier on the uplink and numerous parallel narrowband CDMA carriers on the downlink. The chip rate of the carriers on the multicarrier downlink is identical to that of IS-95. The chip rate of the carriers on the multicarrier uplink is exactly three times that of the downlink carriers.

TDD

In WCDMA, multicarrier, and EDGE modes, the uplink and downlink are separated by means of frequency division. However, in the time-division duplex mode, the uplink and downlink are seperated by means of time division. Consequently, in the TDD

BOX D, APPLICATION SCENARIOS FOR THE TDD MODE

Scenario A—enhancing capacity at hot spots in FDD systems
The operator uses FDD in the initial deployment of the third-generation system. Over time, due to the popularity of mobile Internet, the operator might experience capacity problems at certain hot spots, such as shopping malls, airports, and city centers. A TDD system could then be used to provide extra capacity—much in the same way as GSM 1800 is currently used to enhance GSM 900.

Scenario B—uncoordinated business systems
A public FDD system operator licenses spectrum to private operators for use in indoor business systems in limited geographical areas. With a dual-mode FDD/TDD terminal, the public operator can cover the business market and extend the services of the public network to users in these environments. (Scenarios A and B are likely to occur after the FDD systems have been in operation for a few years.)

Scenario C—terminal-to-terminal communication
Since terminals must be able to transmit and receive in the same frequency band, according to an approved time schedule, the basic mechanisms for terminal-to-terminal communication are present, opening the door to a completely new range of possibilities:

- Because terminals can relay information between themselves, they can form their own *ad hoc* networks.
- Radio performance and operating times can be improved with multihop networks that transmit data—without path loss—from terminal to terminal until it reaches the base station.
- To prevent dropped calls, operators could use cheap relaying terminals to cover radio shadows.

mode only one frequency band is needed to carry both uplink and downlink traffic. This mode of operation is useful since unpaired spectrum is still available in some parts of the world, mainly Europe. Obviously, paired spectrum is preferred, since it supports FDD allocations. But because it may be difficult to find global allocations of paired spectrum, additional unpaired spectrum slots will probably be identified in the future.

Apart from the physical layer, the main radio parameters and spectral characteristics of TDD are the same as for WCDMA FDD. Because the uplink and downlink are separated in the time domain instead of in the frequency domain, synchronized base stations are needed to manage who may transmit, and when. Otherwise, severe interference results.

Each radio frame is divided into eight time slots, each of which carries a number of overlaid channels separated by means of code-division techniques (Figure 13). Thus, the TDD mode shares features from TDMA and CDMA. Each slot is allocated a direction for uplink traffic or downlink traffic. The allocation of slots is not predefined, however. Some flexibility is inherent in the TDD mode for accommodating asymmetric services. For example, if the greater part of the traffic is expected on the downlink, then the majority of the slots can be allocated to the downlink.

Midambles are used to estimate channel response. These and other important parameters are designed to facilitate multiuser detection techniques for better performance.

In terms of bit rates, switching method (packet or circuit), asymmetry, and so on, the basic services of the TDD mode are the same as for the WCDMA FDD mode. Although it has no immediate built-in limitations, the TDD mode is best suited for small cells (pico or micro) with little to moderate delay spread (Box D).

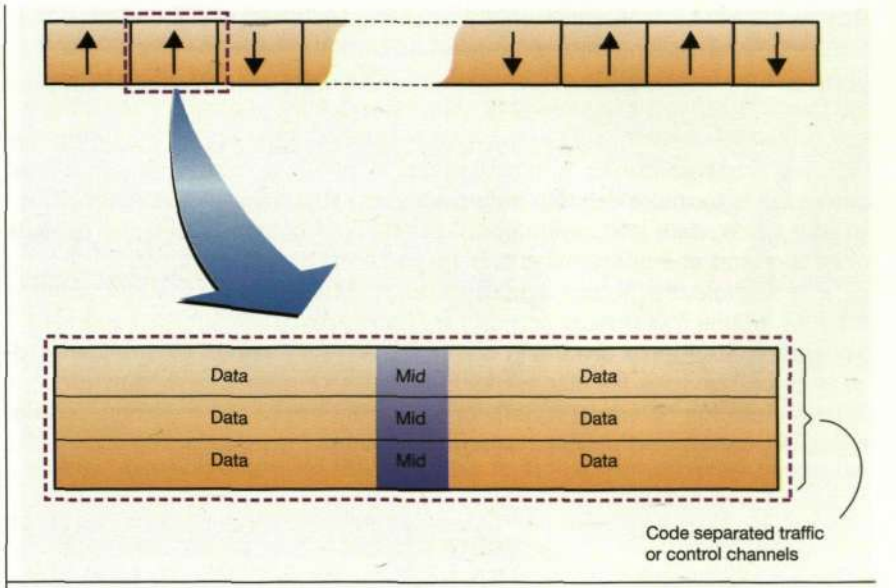


Figure 13
Aspects of time and code division. Slots separated by time. Each slot is assigned a direction—uplink or downlink.

Conclusion

Regional standardization bodies and members of the telecommunications industry have established a global scenario—composed of four optional modes (DS, EDGE, MC and TDD)—for the third generation of radio access standards. The newly harmonized family of third-generation standards caters for requirements for multimedia and flexibility, spectrum allocation, and successful migration from second-generation systems.

In terms of migrating major second-generation systems to a fullness of third-generation capabilities, the industry has agreed that there must be one mode of TDMA-based and three modes of wideband CDMA-based radio interfaces and two kinds of core network. In addition, each radio-access network mode must be able to connect to either type of core network.

The 3GPP aims to deliver the first complete specification of the WCDMA (direct-sequence) and TDD modes by the end of 1999. The 3GPP2 will also deliver a complete specification of the cdma2000 multicarrier mode by the end of 1999. Similarly, EDGE phase I will be finalized by the end of 1999. EDGE phase II will be complete in 2000.

Toward third-generation mobile multimedia communication

Torbjörn Nilsson

During the next few years, mobility-enabling technology will embrace the Internet. The resulting new dimension of communications will put the Internet into the pockets of hundreds of millions of people. Even today, we can witness the beginnings of these developments, which will further influence the development and evolution of future networks, including the architecture, backbones, applications, and mobile access.The immediate challenge is to make existing networks ready to deliver multiple services—voice, data and multimedia—in real time across public and private networks and at a guaranteed end-to-end level of quality.

In this article, the author describes the requirements for the third generation of mobile multimedia communication and how dissimilar second-generation standards are being converged and enhanced to provide third-generation services. He also walks the reader through an overview of third-generation networks, briefly outlining the backbone network, mobile network access, and mobile terminal networks.

Mobile communication is changing society's behavior. Mobile phones have become an everyday accessory for hundreds of millions of people . Today, more and more people use mobile phones as their sole means of personal voice communication. This trend is most apparent among young and single adults. In the Nordic Countries and the USA, as well as in many other mature and competitive telecommunication markets, this pattern is now well established.

During the next decade, the information society will evolve into a globally networked economy—a development that is being shaped by the convergence of computing, communication and broadcasting technologies. The emerging third generation of mobile communication ushers in a true paradigm shift. While mobile communication is presently voice centric, offering the benefits of person-to-person speech communication anywhere and at anytime, personal telephony is rapidly being transformed into a mass market of personal mobile multimedia services and terminals. Third-generation mobile communication will do much more than bring voice communication capabilities to our pockets. It will also make information services instantly available, including the Internet, intranets, and entertainment services—for instance, a third-generation terminal might function as a video camera from which end-users can send electronic postcards and video clips.

End-users will also be able to use their terminals as a tool for mobile electronic commerce (e-commerce). In essence, the end-user will have a retail outlet in his or her pocket, with the ability to reserve tickets,

Figure 1
Mobile telephony has become a mass-market service.

make banking transactions, pay parking fees, buy items from a vending machine, and so on. Third-generation mobile communication will also introduce a more powerful, flexible and efficient way of doing business. Mobile multimedia services and mobile or wireless office solutions will simplify the implementation of virtual enterprises. Similarly, appliance-to-appliance and appliance-to-people communication applications will grow in importance, vastly improving security and efficiency.

Market growth

Growing number of subscribers

The growth of traditional, fixed-voice subscriptions is beginning to slow and may level off in coming years. However, we foresee continued strong growth in mobile communication. In fact, we estimate that by 2003/2004 this number will approach one billion. The number of Internet subscribers is also on the rise. Here, too, we foresee close to one billion users by 2004. Of these, more

Figure 2
In all likelihood, mobile e-commerce will become a fundamental part of the emerging digital economy.

BOX A, ABBREVIATIONS

1XRTT	Technology for migrating CDMA/IS-95 toward 3G (The standard introduces 144 kbit/s packet data in mobile environments and greater rates in fixed environments)
ATM	Asynchronous transfer mode
CAMEL	Customized application for mobile enhanced logic
CDMA	Code-division multiple access
cdma2000	Telecommunications Industry Association (TIA) standard for third-generation technology that is an evolutionary outgrowth of cdmaOne
CDMA/IS-95 cdmaOne	IS-95 cellular digital standard cdmaOne describes a complete wireless system that incorporates the CDMA/IS-95 air interface, the ANSI-41 network standard for switch interconnection, and numerous other standards that make up a complete wireless system
CDPD	Cellular digital packet data
DWOS	Digital wireless office system
EDGE	Enhanced data rates for GSM and TDMA/136 evolution
FDD	Frequency-division duplex
GPRS	General packet radio services
GSM	Global system for mobile communication
HSCDS	High-speed circuit-switched data
IMT-2000	International mobile telecommunication 2000
IN	Intelligent network
IP	Internet protocol
MAP	Mobile application part
MPLS	Multiprotocol label switching
PDC	Personal digital cellular
PMR	Private mobile radio
P-PDC	Packet-mode PDC
PSTN	Public switched telephone network
QoS	Quality of service
SDH	Synchronous digital hierarchy
SMR	Specialized mobile radio
SONET	Synchronous optical network
TDD	Time-division duplex
TDMA	Time-division multiple access
TDMA/136	IS-136 cellular digital standard
UMTS	Universal mobile telecommunications system
VHE	Virtual home environment
WAP	Wireless application protocol
WCDMA	Wideband CDMA
WCDMA-DS	Direct-sequence WCDMA
WDM	Wavelength-division multiplexing
WIN	Wireless intelligent network
WWW	World Wide Web

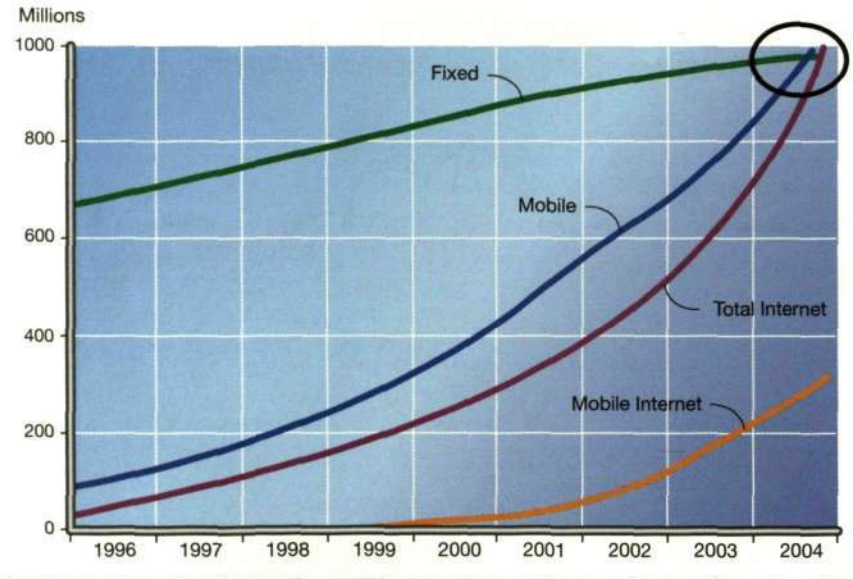


Figure 3
Forecast growth (subscribers) of fixed and mobile telephony and the Internet.

than 350 million will be mobile Internet subscribers (Figure 3).

The growth of mobile Internet will be stimulated by the growth of fixed Internet, by global mobile-data standards for the four major mobile cellular systems (GSM, TDMA/136, PDC and CDMA/IS-95), and by high volumes and short product life-cycles of mobile terminals (whose numbers will far exceed that of personal computers).

Growing volumes of traffic

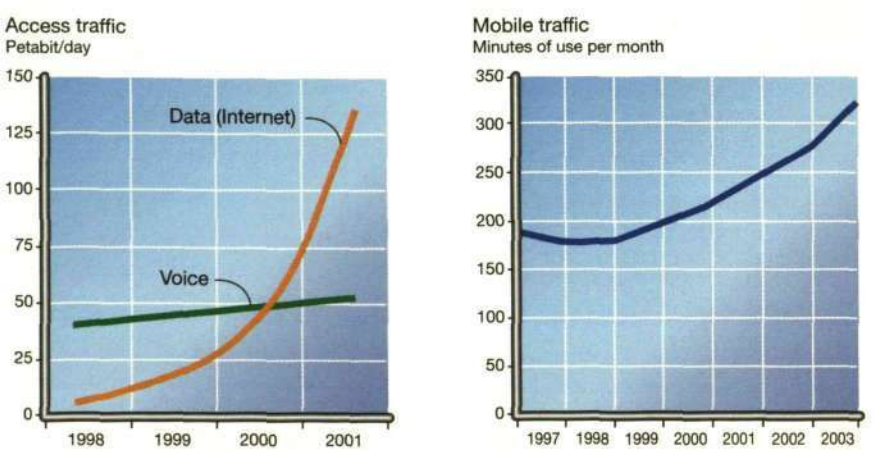
Traffic constitutes the second important area of growth. The accumulated volume of data traffic is currently on the verge of surpassing the accumulated volume of voice traffic in all public networks. In five years, data traffic will completely dominate (Figure 4). However, voice traffic is on the rise in the mobile network. We predict that the accumulated volume of traffic per subscriber will double or even triple by 2004, due primarily to lower tariffs.

Requirements

To succeed, third-generation mobile communication must provide the mass market with high-quality, efficient, and easy-to-use wireless mobile multimedia services. The third-generation systems must provide support for

- high data rates—up to at least 144 kbit/s (384 kbit/s) in all radio environments and up to 2 Mbit/s in low-mobility and indoor environments;
- symmetrical and asymmetrical data transmission;
- packet-switched and circuit-switched services, such as Internet (IP) traffic and real-time video;
- good voice quality (comparable to wireline quality);
- greater capacity and improved spectrum efficiency compared to present-day, second-generation wireless systems;
- several simultaneous services to end-users and terminals—that is, for multimedia services;

Figure 4
Forecast growth of voice and data traffic.



- the seamless incorporation of second-generation cellular systems and for the co-existence of, and interconnection with, mobile satellite services;
- roaming, including international roaming, between different IMT-2000 operators; and
- economies of scale and an open global standard that meets the needs of the mass market.

Applications and services

The evolution of mobile multimedia communication is driven by the demand for easy-to-use infrastructure- and terminal-related applications and solutions.

In coming years, data-centric and Internet- and intranet-based services will be adapted to, and become available over, the wireless network. The following services and applications embody the most highly valued capabilities of a third-generation wireless system (Figure 5):

- Full range of services—from narrowband voice to wideband, real-time multimedia services. Voice traffic is expected to remain an important application and source of revenue.
- Support for high-speed packet data, including
  - the browsing of information and the World Wide Web (WWW);
  - information delivery (news, weather, traffic, finance) via push techniques—the information might even be location-dependent; and
  - remote and wireless access to the Internet/intranets.
- Unified messaging services, such as multimedia e-mail.
- Real-time audio/video applications, such as videophone, interactive video-conferencing, audio and music, and specialized multimedia business applications, including telemedicine and remote security surveillance.
- Mobile e-commerce applications:
  - mobile banking; and
  - mobile shopping.
- Mobile office applications:
  - seamless multimedia for users who are on the move and at the office;
  - Specialized and private mobile-radio (SMR/PMR) services; and
  - intranet access.

New portable and pocket-sized wireless terminals will support these new multimedia applications.

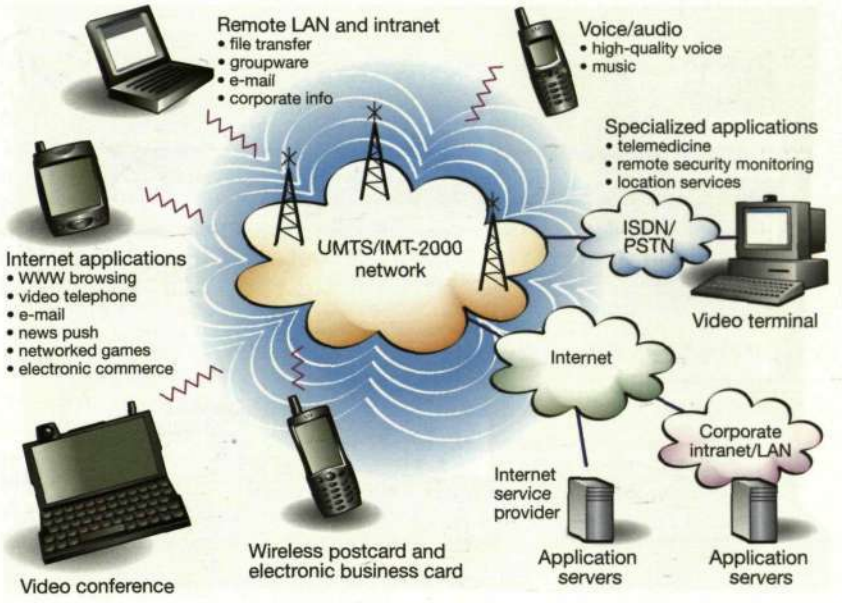


Figure 5
Examples of third-generation user applications.

Evolution of mobile radio standards

Present-day digital wireless standards continue to evolve—in particular, as relates to value-added services, capacity, coverage, quality, costs, bandwidth, and data or multimedia services. Each major cellular standard (GSM, TDMA/136, CDMA/IS95 and PDC)

Figure 6
Possible future mobile office tool for a traveling businessperson.

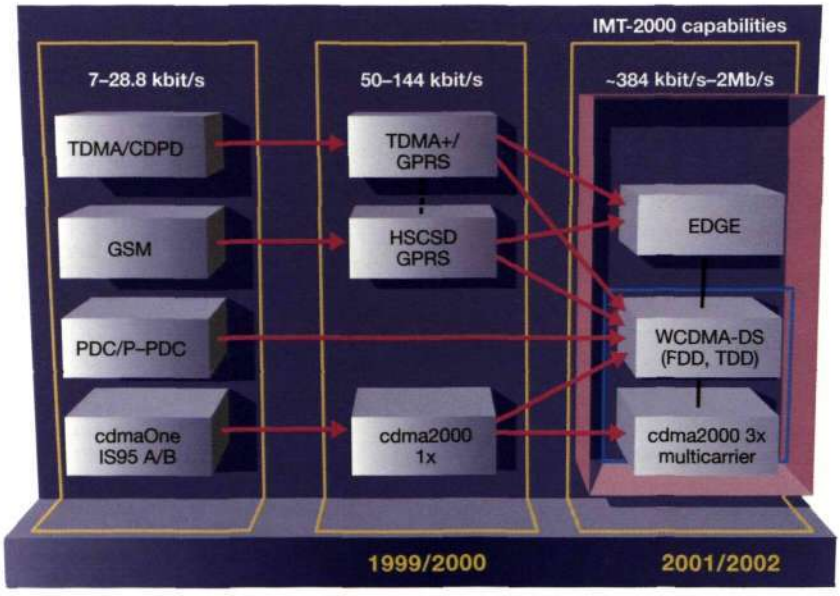
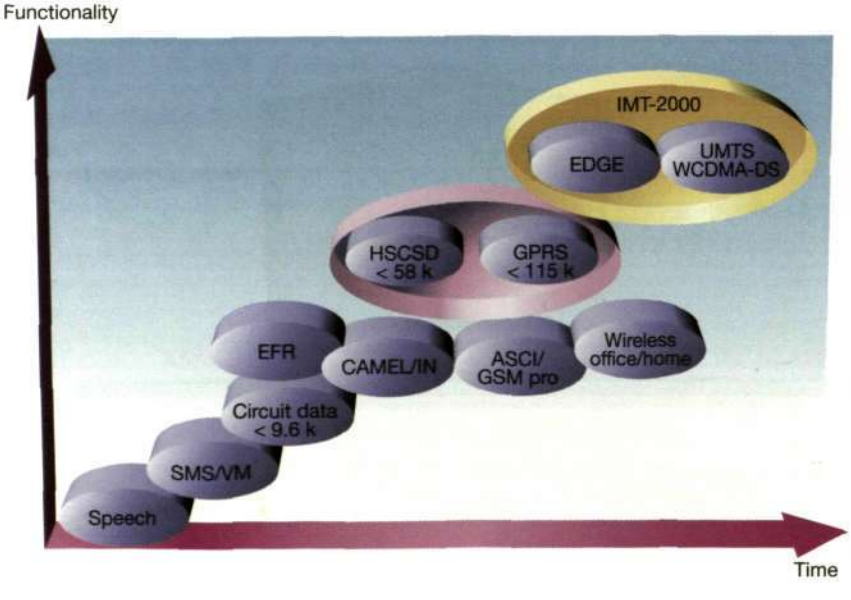


Figure 7
Main evolution toward global third-generation standards.

Figure 8
Basic evolution of GSM toward third-generation capabilities.



is being developed to provide third-generation capabilities. The second- and third-generation systems will accommodate dual-mode terminals that work with different satellite systems, to offer complete coverage even in rural and remote regions.

When early volumes of third-generation systems become commercially available (around 2002), more than 700 million subscribers are expected to be using cellular services. Of these, some 50 million will also have access to data and Internet service. This number makes up a substantial customer base and represents considerable operator investments. Therefore, two of the most important requirements for third-generation systems are that they must provide a seamless path of migration from present-day digital wireless networks, and that they are capable of interworking with existing networks.

Global harmonization of standards

In general, the global telecommunications industry wants to reduce the number of third-generation mobile standards (from the current number of 2G standards), while at the same time respecting existing standards. With this in mind, the industry has recently made two major achievements: the convergence of TDMA/136 and GSM, and the convergence of CDMA modes.

The convergence of TDMA/136 and GSM begins with the general packet radio services (GPRS) standard, which creates a common core network architecture and shares network components, and continues with EDGE, which unifies the radio network and terminals (Figure 7).

The convergence of the CDMA modes creates a single radio-access family of third-generation CDMA modes:

- WCDMA direct-sequence, frequency-division duplex (FDD) mode;
- WCDMA direct-sequence, time-division duplex (TDD) mode; and
- multicarrier CDMA, FDD mode.

The WCDMA direct-sequence (WCDMA-DS) modes are the main modes being proposed for the universal mobile telecommunications system (UMTS). The multicarrier CDMA mode is mainly for the evolution of cdmaOne/cdma2000.

The PDC standard is being evolved directly into the WCDMA-DS mode.

In summary, we now have fewer radio standards and network protocols: MAP/GSM, IS41/TDMA and cdmaOne.

For the two third-generation standards or

modes and the four second-generation standards to work together, interworking functionality must be provided at the network level (with interworking units between protocols) and at the terminal level (dual-mode/multi-mode terminals).

GSM enhancements

The GSM standard is being enhanced to provide even better services, capacity, coverage, quality, and data rates. A series of developments has already been initiated to enhance the functionality of GSM networks (*Figure 8*):

- The customized application for mobile enhanced logic (CAMEL) will give subscribers continued support for intelligent network (IN) services when they roam into other networks; for example, by creating a virtual home environment (VHE) for visiting subscribers. The first phase of CAMEL has already been implemented.
- Several new, value-added applications are being implemented, including GSM on the net (GSM at the office), GSM Pro (SMR/PMR functions), and prepaid subscriptions.[1,2]
- The first commercially deployed enhancement for increasing data rates—called high-speed circuit-switched data (HSCSD)—will initially support data rates of up to 57.6 kbit/s, using from one to four 14.4 kbit/s time slots.
- GPRS is a packet-switched service that allows full mobility and wide-area coverage with data transmission rates of up to 115 kbit/s.[3]
- Enhanced data rates for GSM and TDMA/136 evolution (EDGE) uses enhanced modulation and related techniques, raising data rates to 384 kbit/s or higher.[4] The GSM carrier bandwidth (200 kHz) and the complete TDMA frame structure, logical channel structure, frequency plans and methods remain unchanged. Channels and transceivers with EDGE functionality will operate in either GSM/GPRS or EDGE modes. Coexistence of this kind in the same network enables operators to introduce EDGE technology incrementally into any of today's GSM frequency bands: 900, 1800 and 1900 MHz.

TDMA/136 enhancements

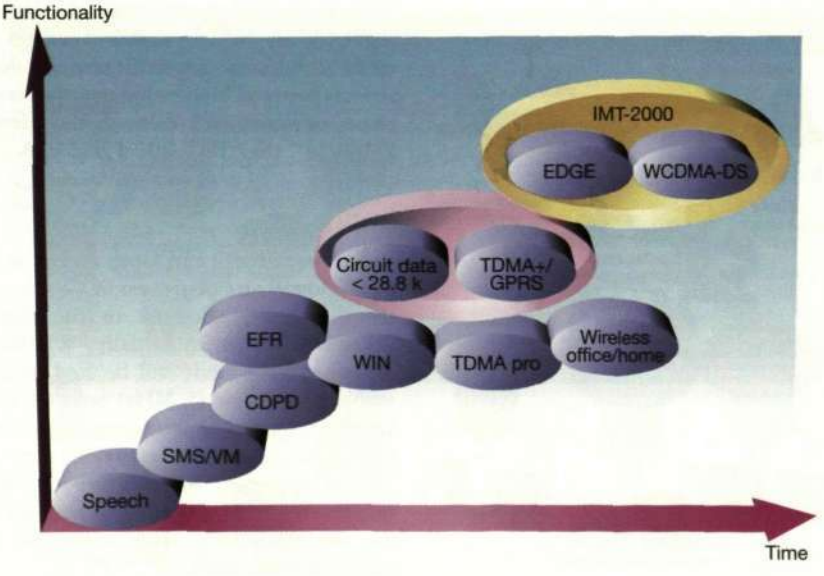
TDMA/136 (formerly IS-136 or D-AMPS) is also being enhanced to provide better services, capacity, coverage, quality, and data rates (Figure 9):

- Numerous service enhancements have recently been implemented, including wireless intelligent networks (WIN), digital wireless office systems (DWOS), and SMR/PMR services.[5,6]
- Two additional phases of data-rate enhancements are also envisaged. In the first phase, the bit rate of the 30 kHz radio carrier will be increased by means of high-level modulation to yield bit rates of up to approximately 64 kbit/s. Similarly, GPRS will be introduced into the core network. During the second phase (EDGE), a new air interface will be introduced (the same air interface as in GSM/EDGE). Initially, TDMA/136 will provide data rates up to and beyond 384 kbit/s. Operators will thus be able to offer third-generation wireless services in either of today's TDMA frequency bands, 850 and 1900 MHz. Later, operators might also be able to connect the WCDMA air interface (in new or reformed spectrum).

cdmaOne enhancements

As with GSM and TDMA/136, CdmaOne will also be enhanced to provide improved services, capacity, coverage, quality, and data rates. The current data rate is 14.4 kbit/s, but the introduction of

Figure 9
Basic evolution of TDMA/136 toward third-generation capabilities.



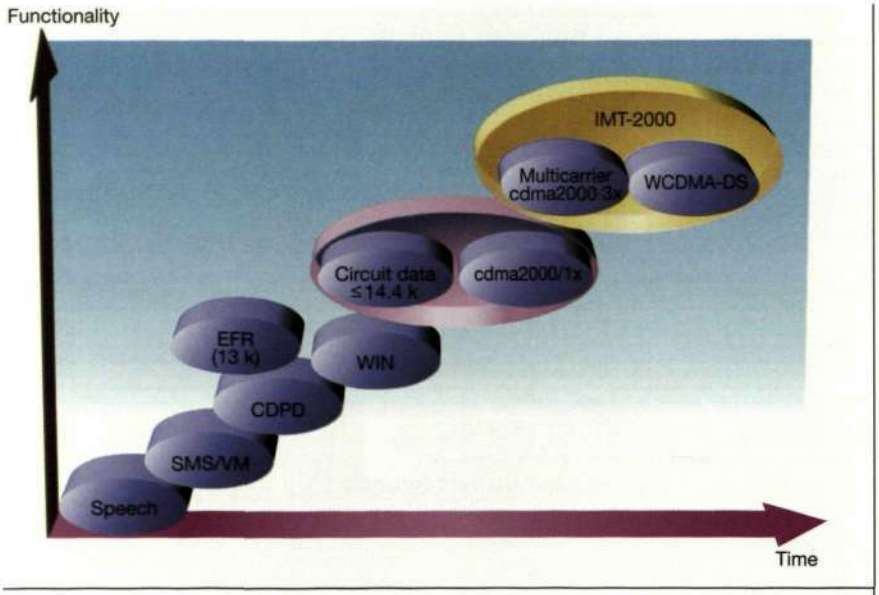


Figure 10
Basic evolution of cdmaOne/cdma2000 toward third-generation capabilities.

cdma2000/1XRTT will provide packet data rates of between 9.6 and 144 kbit/s. Later, the introduction of cdma2000 wideband (multicarrier CDMA and WCDMA-DS) will increase data rates to as much as 2 Mbit/s (Figure 10).

PDC enhancements

The PDC standard is also being improved to provide enhanced services, capacity, coverage, quality, and data rates. The standard has recently implemented the packet-mode (P-PDC), which gives packet data rates of up to 28.8 kbit/s. Several innovative data services have also been introduced. In a later phase, operators will directly move from PDC to the IMT-2000/WCDMA-DS mode.

WCDMA-DS

In Europe (UMTS) and Japan as well as in many other parts of the world, new spectrum is being allocated to the third-generation standards. Initially, WCDMA direct-sequence mode will be used, which employs wideband (5 MHz) carriers. This standard can also be introduced into other mobile frequency bands, provided there is enough spectrum to introduce 5 MHz carriers. WCDMA-DS will also function as the direct-sequence mode connected to cdmaOne/cdma2000.

Migrating second-generation mobile systems into third-generation mobile systems

The WCDMA-DS access will coexist with present-day as well as "evolved" GSM access (GPRS/EDGE), and the related dual-mode mobile terminals will support full roaming and handover from one system to another. In the introductory phases of WCDMA, the dual-mode terminals will ensure that subscribers can roam and interwork with the rest of the GSM community from the very outset (Figure 11). The evolved GSM core network will serve as the basis for a common GSM/UMTS core network that connects GSM/GPRS/EDGE and UMTS/WCDMA-DS access.

In Japan, the plan is to deploy IMT-2000 as a fully overlaid network on top of the PDC network, with interworking functions between the two networks. Dual-mode PDC/IMT-2000 terminals might also be introduced.

The cdmaOne community mainly aims to evolve the cdmaOne standard into multicarrier CDMA. However, it also plans to add and adapt harmonized WCDMA-DS to its core network.

The TDMA community is also looking into the possibility of adding global WCDMA-DS radio access, at least in Internet-based multimedia applications.

The GSM/UMTS migration strategy, which includes dual- and multi-mode terminals, is a generic migration strategy between second- and third-generation systems. Interworking units at the network level will enable different standards to interwork with one another.

Next-generation networks

The New Telecoms World has its roots in the convergence of fundamentally different communication-network domains. Today, several separate, vertically oriented, single-service networks have been optimized to deliver fixed telephony (PSTN), mobile telephony, data, and cable-TV services. For example, for more than 100 years, classic telephone networks have been optimized to carry real-time voice traffic between hundreds of millions of phones all around the globe.

By contrast, data communication comes to us from an environment in which service quality is grounded on a best-effort approach. Because data communication typically consists of non-real-time applications,

packet data is considered to be the most cost-effective communication solution.

Considerable support exists for packet-ready networks. Indeed, in some markets, packet data accounts for more than 50% of the traffic volume in access networks.

Next-generation network structure

The significant increase in data traffic, together with the convergence of services in a multiservice network, calls for a new generation of networks (Figure 12). The next-generation, packet-oriented network will be characterized by

- a layered network structure and that decouples applications, control and connectivity (that is, transport and switching) of bits;
- a connectivity layer composed of a common backbone network; separate wireless, copper, and fiber/coax access; and media gateways that connect different packet-oriented (ATM/IP) and circuit-switched networks;
- a client/server type of architecture between servers in the call/mobility control level and connectivity layer (media gateways). A client/server architecture will also exist between different communication applications and external content applications, as well as between end-user devices (clients). The end-user devices will also work with each other in client/server networks;
- the development of open interfaces and standards, which are essential to this type of architecture. Especially important are acceptance and support for third-party development, content and applications, and openness between end-user devices and the network; and
- end-to-end control through the layers, in order to provide a quality-of-service solution or application. Especially vital to end-to-end control and management are telecom management, network management (supervision, alarm handling, traffic control, QoS agreements), billing, security, and customer care.

The benefits of a multiservice network are a common network with shared management and services and servers (for multiple services) and an open, layered architecture with open interfaces for improving time to market and third-party applications. Put another way, multiservice networks cost less to operate and offer greater flexibility.

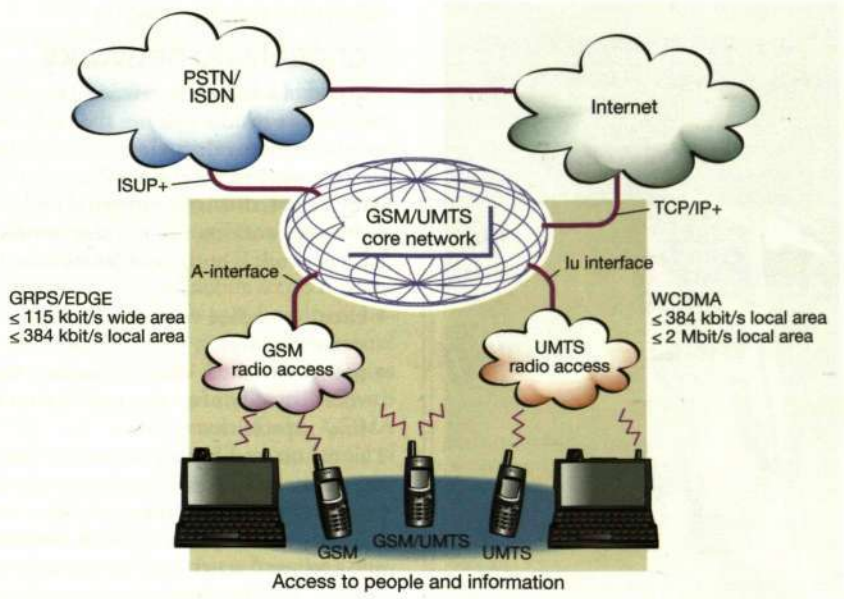
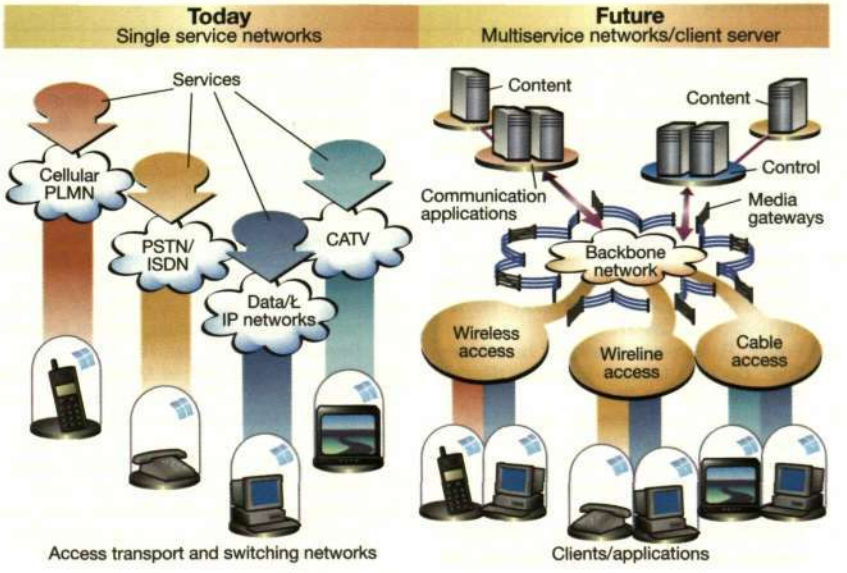


Figure 11
Network evolution and migration between GSM and third-generation mobile multimedia communication (UMTS).

Figure 12
Evolution toward next-generation multiservice networks.



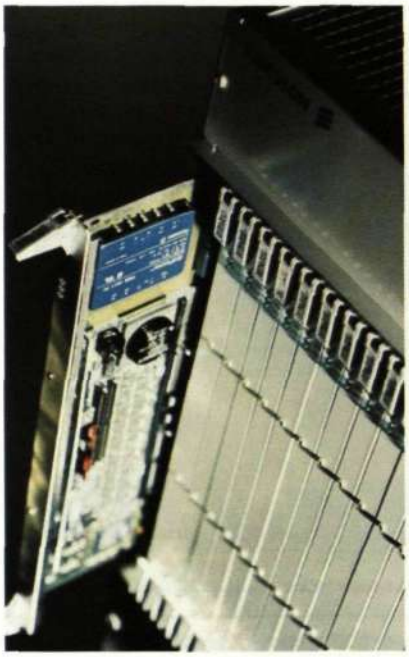


Figure 13
Ericsson's AXD301, a scalable, high-performance, packet-switching system for the carrier backbone.

Mobile multimedia connectivity networks

The base of subscribers connected to circuit-switched networks will not suddenly cease to exist. For some time, circuit-switched and packet-switched networks will coexist. Initially, data traffic is being introduced as an overlay network on top of voice networks. Each network is optimized for its basic services:

- circuit-switched voice provides real-time quality of service;
- packet-switched data provides "best-effort" quality of service and "always online" capabilities.

The circuit and packet networks will be interconnected by means of gateways that provide common mobility control.

To begin with, data must be developed into a robust "carrier-class" wireless, mobile service. Specific developments and standards are needed to make data work efficiently over the air (wireless) and in mobile environments. Examples include Mobitex, cellular digital packet data (CDPD) and GPRS.[7] These technologies will be followed by third-generation standards, such as EDGE and WCDMA.

The wireless application protocol (WAP), which gives mobile users access to various Internet-based services, is one example of how data can be adapted to wireless or mobile environments.[8] This new dimension of mobile communication is destined to put Internet services into the pockets of hundreds of millions of people. WAP also exemplifies an end-to-end wireless standard within the client/server architecture.

A second major challenge is to prepare connectivity networks to offer multiple services over the same network technology. In the process, many obstacles will have to be overcome in the public (carrier) networks. The objective is to provide efficient, real-time delivery with guaranteed quality of service in and between public networks, to private networks, and over the air, and to guarantee reliability and security as required by large public networks. Various standards bodies are currently grappling with this challenge.

Roadmap toward third-generation networks

The backbone network

The backbone and core network of next-generation mobile networks will evolve into a packet-centric architecture that makes use of various switching or routing and transport techniques, such as wavelength-division multiplexing (WDM), synchronous digital hierarchy and synchronous optical networks (SDH/SONET), asynchronous transfer mode (ATM) and the Internet protocol (IP).[9,10] Today, many multiservice networks use scalable, high-capacity ATM switching and transport and can provide ATM services and frame-relay services with IP services on top; for example, using the multiprotocol label switching (MPLS) standard.[11,12] To become more efficient in large, real-time, multiservice networks, the Internet protocol will have to be developed further. In the next release (IPv6), addressing and core network mobility will be much improved. Still further enhancements will be needed, however, to create a next-generation Internet protocol that efficiently handles public, real-time, multiservice and wireless networks. Ericsson and others in the industry are leading the work to define the optimum wireless Internet protocol.

Mobile network access

ATM and IP technology will be used to "packetize" nodes of the mobile access net-

REFERENCES

1 Granberg, O.: GSM on the Net. Ericsson Review 75 (1998):4, pp.184-191.
2 Gratorp, A., Nilsson, P. and Smedman, T. Ericsson Pro products—Adapting mass-market technology to fit specialized needs. Ericsson Review 76(1999):1, pp 8-13.
3 Granbohm, H. and Wiklund J. GPRS—General packet radio service. Ericsson Review 76(1999):2, pp. 82-88.
4 Furuskär, A., Näslund, J. and Olofsson, H. Edge—Enhanced data rates for GSM and TDMA/136 evolution. Ericsson Review 76(1999):1, pp. 28-37.
5 Foster, R. Wireless intelligent networks—The flexible future. Ericsson Review 75(1998):2, pp. 78-82.
6 Johansson R., Nilsson M. and Ward, T. Mobile Advantage Wireless Office—A digital wireless office sstem for TDMA/136 networks. Ericsson Review 76(1999):1, pp 20-27.
7 Wetterborg, L. CDPD—Adding wireless IP services to D-AMPS/AMPS wireless networks. Ericsson Review 73(1996):4, pp.151-156.
8 Erlandson C. and Ocklind, P. WAP—The wireless application protocol. Ericsson Review 75(1998):4, pp.150-161.
9 Grenfeldt, M. Erion—Ericsson optical networking using WDM technology. Ericsson Review 75(1998):3, pp. 132-137.
10 Hopfinger, J., Khodaverdian, O. and Saure, E. The SDH interface in AXE. Ericsson Review 75(1998):4, pp. 192-204.
11 Blau, S. and Rooth, J. AXD 301—A new generation ATM switching system. Ericsson Review 75(1998):1, pp. 10-17
12 Hågård, G. and Wolf, M. Multiprotocol label switching in ATM networks. Ericsson Review 75(1998):1, pp. 32-39
13 Waesterlid, A. Open communication devices using the EPOC operating system. Ericsson Review 75(1998):1, pp. 14-19
14 Haartsen, J. Bluetooth—The universal radio interface for *ad hoc*, wireless connectivity. Ericsson Review 75(1998):3, pp. 110-117.

work. Likewise, ATM and IP will each be used to provide efficient transport and routing capabilities. The packet network will also be adapted to meet the requirements for real-time service and wireless transmission, which encompasses wireless terminals (end-to-end perspective). The first leg of this evolutionary path, which will be characterized by continuous, step-by-step standardization and (end-to-end) development, begins with GPRS.

Mobile terminal networks

New multimedia applications will drive the market for third-generation services. The terminals used in the mobile multimedia era will nearly always be turned on, serving as the gateway to the Internet or to corporate intranets via packet-switched connectivity networks. This will eliminate delays associated with setup, and add convenience to the use of data and multimedia services.

Ericsson and other members of the industry are creating new, open platforms and standards which facilitate multimedia applications that can be accessed by and run on wireless terminals (Figure 14):

- The wireless application protocol easily adapts information from the Internet for access via mobile terminals.
- Ericsson is a founding member of Symbian, a joint venture that recognizes the need to standardize an operating system (EPOC) that supports mobile devices and usage.[13] The Symbian partners are working to create a toolbox that will enable third-party software developers to create innovative services for third-generation mobile devices.
- Bluetooth is a new, low-cost, short-distance radio technology that was designed to eliminate cables between portable and peripheral terminals and devices.[14] Bluetooth can, for example, connect mobile terminals, digital cameras, scanners, printers, and PCs.

Conclusion

During the next decade, the information society will evolve into a globally networked economy—a development that is being shaped by the convergence of computing, communication and broadcasting technologies. Accordingly, the third generation of mobile communication will enable end-users to enjoy the benefits of data and image or video communications while on the move—true mobile multimedia.

By 2003 or 2004, the number of subscribers of mobile communication will have climbed to nearly one billion. Similarly, by 2004, the number of Internet subscribers will approach one billion. Of these, more than 350 million will be mobile Internet subscribers. The explosive popularity of the Internet is testimony that people want multimedia communication and instant access to information.

Second-generation mobile access is being enhanced to offer high-speed multimedia-capable radio access—the convergence and migration of TDMA/136 and GSM begins with the GPRS standard and continues with EDGE; the convergence of the CDMA modes creates a single radio-access family of third-generation CDMA modes. At the same time, data, voice and image/video services are being converged into a single multiservice network that is based on a new, layered, network architecture.

The immediate challenge for suppliers and operators is to provide efficient, real-time quality of service and reliable, wide-area communication using packet-switching techniques, end to end. Over time, more and more connectivity networks (bit pipes) will migrate toward packet-switching technologies (ATM and IP).

Figure 14
Future pocket-sized mobile multimedia concept features voice and video—via a rotating camera. The wireless, hands-free earphone connects via a Bluetooth connection.

Adaptive base-station antenna arrays

Anders Derneryd and Björn Johannisson

Adaptive antenna systems represent an area in which considerable development efforts and field trials are being conducted to increase capacity in mobile communication networks. Ericsson has developed array antennas for use in the 900, 1800 and 1900 MHz frequency bands. Thanks to microstrip patch elements, Butler matrices, and several dual-polarized beams, these antennas yield high antenna gain and excellent spatial efficiency.

The authors describe adaptive antenna hardware and configurations as well as measured results from work on adaptive base-station antennas.

BOX A, ABBREVIATIONS

C/I	Carrier to interference ratio
ERP	Equivalent radiated power
GSM	Global system for mobile communication
RF	Radio frequency
TDMA	Time-division multiple access

Introduction

The rapid growth in the number of users of mobile communications means that many operators must find new ways of increasing the capacity of their networks. Their options include allocating more frequency, introducing frequency-hopping techniques, and adding microcells and adaptive antenna systems.

The introduction of new frequency bands at 1800 and 1900 MHz is an example of allocating frequency to increase capacity. However, compared to 800 and 900 MHz systems, mobile communication at 1800 and 1900 MHz requires more base stations or greater levels of radiated power.

Within the available frequency spectrum, capacity can be increased through the introduction of smaller cells, such as microcells, to create a dense network of base stations. Nonetheless, networks of this kind are often perceived as an aesthetic eyesore, due to the large quantities of associated antenna installations. Indeed, in many regions, the general public demand is for fewer antennas that are smaller in size and less conspicuous.

Another problem associated with adding numerous base stations is the cost involved in finding new locations for the antennas and base-station cabinets. Therefore, more and more operators are showing interest in adaptive antenna systems as a means of resolving their need for greater network capacity.

Ericsson has conducted extensive research and development of advanced base-station antennas for mobile communication. This work comprises both adaptive and active antenna systems. With the introduction of active antenna products, such as Ericsson's Maxite products, operators can now use small-sized base station units with high levels of equivalent radiated power (ERP) and low power consumption.

Ericsson has vast experience of array antenna products which, thanks to a superior design practice and the integration of antenna and electronic components, make attractive system solutions.[1-4] Product examples found in commercial and defense applications include

- Maxite active antennas (Figure 1);
- the MINI-LINK family (Figure 2;
- Erieye airborne early-warning radar (Figure 3); and
- Arthur artillery hunting radar (Figure 4).

Figure 1
Maxite active antenna and base station.

Figure 2
MINI-LINK with integral antenna.

Discrete base-station antennas

Traditional installations of mobile communication base-station antennas make use of space-diversity techniques, which require at least two antennas pointing in the same direction and separated from each other by a distance of 10 to 20 wavelengths.

An alternative to space diversity is polarization diversity, which also reduces antenna visibility.

Polarization diversity increases gain through the simultaneous reception of two orthogonal polarized signals from a single, dual-polarized antenna. In most instances, polarization diversity is just as efficient as space diversity. Thus, the introduction of polarization diversity means that the antenna installation may consist of a single antenna unit used for both transmission and reception. In addition to being more aesthetically pleasing, this solution is often easier and more cost-effective to install. Figure 5 shows an adaptive-antenna array installation at an existing site. In total, the site has nine sector antennas oriented in three different directions. In each direction, one transmission antenna and two receive antennas use space diversity. Each set of three antennas can be replaced with the much smaller installation of a single, adaptive-antenna array, which yields approximately the same antenna gain and coverage as the traditional arrangement.

Adaptive antenna configurations

Adaptive antenna systems offer a promising way of increasing network capacity. The antenna arrays in these systems have a horizontal extension that enables narrow antenna beams to be created in the azimuth plane. An antenna array also makes it possible to obtain angular resolution in the horizontal plane (azimuth angles) that can be accessed in order to identify the position of mobile terminals and to assess traffic distribution.

In principle, narrow beams directed toward a mobile terminal reduce interference levels in the network and thereby increase capacity. With the proper choice of receiver algorithms, it is possible to attain a good carrier-to-interference (C/I) ratio in signals received from a mobile terminal. Angular directions to mobile terminals and interferers may also be determined, which gives sufficient information for making an intelligent choice of transmit beam with

- high amplitude levels toward the target mobile terminal; and
- low interference levels in other directions.

Several methods can be used for directing radiated power from an array antenna into a narrow beam. The required phase front along the antenna elements that correspond to a scanned narrow beam can be generated either through digital beam forming in the transceivers or at radio frequency (RF), using passive networks or phase shifters.

Regardless of the method used, beam forming is required to ensure phase coherency all the way from the beam former to the antenna elements. When digital beam forming is performed in the transceivers at the base station cabinet, the feeder cables between the cabinet and the antenna must remain in phase for their entire lifetime.

One the other hand, a passive beamforming network at the antenna does not require phase coherence between the radio transmitters and the beam former, which allows the feeder cables to have arbitrary phase. With RF beam forming in the antenna, phase coherence is easily achieved, since the phase requirements only involve transmission lines within the antenna unit

Figure 3
Erieye airborne early-warning radar.

Figure 4
Arthur artillery hunting radar.

Figure 5
A TDMA 1900 adaptive antenna installed at a traditional three-sector site together with space-diversity antennas.

itself, as part of the standard design of an antenna array. One example of a passive network is the Butler matrix, which generates a set of simultaneous orthogonal beams from a single array antenna and minimizes beamforming loss. A crossover gain drop between the orthogonal beams must be considered in the system design. Ideally, gain at the crossover point using a Butler matrix is 3.9 dB less than beam peak gain.

Antenna arrays

Ericsson has developed two-dimensional antenna arrays for adaptive base-station systems. These arrays, which are developed for

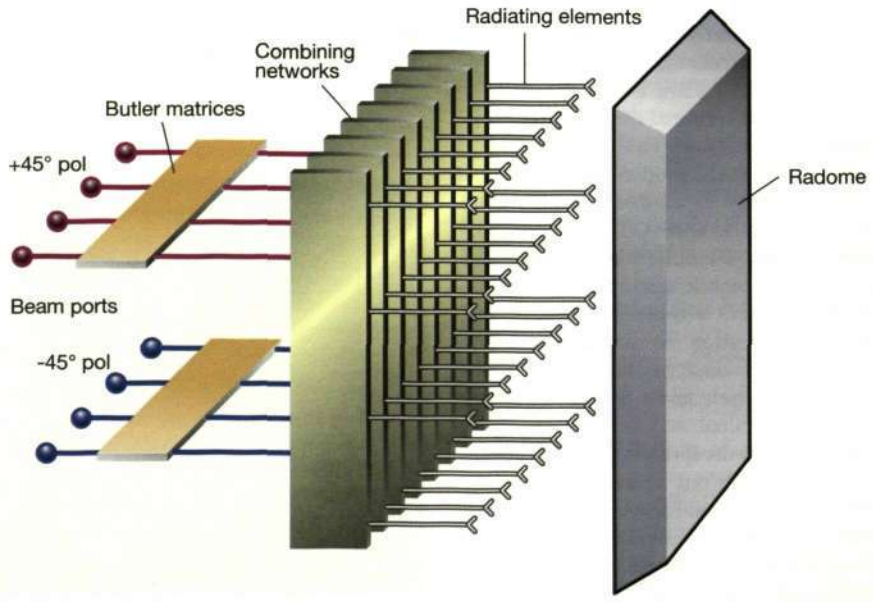


Figure 6
Principles of a multibeam array antenna.

systems based on GSM and TDMA (IS 136) standards, work in the 900, 1800 or 1900 MHz frequency bands. Together with Mannesmann Mobilfunk GmbH (GSM) and AT&T Wireless Services (TDMA), Ericsson has conducted field trials in live networks to evaluate the performance of the adaptive systems. The results show that adaptive antenna systems considerably increase capacity.

The adaptive array antenna transmits and receives radio-frequency signals in directed narrow beams. Figure 6 shows a principle view of a multibeam array composed of a dual-polarized multibeam antenna with four azimuth beams in each of two orthogonal polarizations. The orientation of the polarization is slant linear ±45°. The radiating elements are aperture-coupled microstrip patches, located in columns spaced half a wavelength apart. For each polarization, the radiating elements in each column are combined in a vertical network. One horizontal beam-forming network (that is, the Butler matrix of each polarization) combines the radiating element signals to beam ports: four beams with +45° polarization and four beams with -45° polarization. A radome is placed in front of the antenna to protect it from the environment. The same array antenna is used both to transmit and receive, and must work over the entire system frequency band. A broadband design of the aperture-coupled patches and feed networks makes this possible.

In GSM as well as TDMA, base stations must occasionally transmit a control channel simultaneously over the entire sector region. To satisfy this requirement, a separate sector antenna function has been introduced as part of the adaptive antenna system. An effective solution uses an additional column of radiating elements next to the array antenna columns. For best results, the deviation between the sector antenna radiation pattern and the array antenna multibeam pattern must be as small as possible.

Besides increased capacity, the increase in antenna gain may also be exploited to offer greater coverage. The gain of the GSM 900 array (Figure 7) is comparable to that of a

Figure 7
A GSM 900 multibeam array antenna.

REFERENCES

1 Dahlsjö, O., Ljungström, B. and Magnusson, H.: Fibre-Reinforced Plastic Composities in Sophisticated Antenna Designs, Ericsson Review 64 (1987):2, pp. 50-57.
2 Derneryd, A. and Wilhelmsson, H.: Dichroic Antenna Reflector for Space Applications, Ericsson Review 68 (1991):2, pp. 22-33.
3 Dahlsjö, O.: Antenna Research and Development at Ericsson, IEEE Antennas and Propagation Magazine 34, (April 1992): 2, pp. 7-17.
4 Ahlbom, S., Andersson, P. and Lagerlöf, R.: A Swedish Airborne Early Warning System Based on the Ericsson ERIEYE Radar, Ericsson Review 72 (1995):2, pp. 54-63.

BOX B, TERMS AND DEFINITIONS

C/I
The ratio between the received desired signal and interference signals, usually expressed in dB.

ERP
In a given direction, the relative gain of a transmitting antenna with respect to the maximum directivity of a half-wave dipole multiplied by the net power accepted by the antenna from the connected transmitter.

Grating lobe
A lobe, other than the main lobe, produced by an array antenna when the interelement spacing is sufficiently large to permit the in-phase addition of radiated fields in more than one direction.

Orthogonal polarization
In a common plane of polarization, the polarization for which the inner product of the corresponding polarization vector and that of the specified polarization is equal to zero.

Phase coherence
Each antenna element must be supplied with the correct phase of signal in order to preserve the overall desired beam shape. This means that phase matching must be exact from the point of beam forming to the individual radiating antenna elements.

Polarization diversity
Orthogonal polarization components are used to provide diversity reception. Two-branch diversity is provided with a single antenna unit with dual-polarized radiating elements.

Radome
A cover usually intended for protecting an antenna from the effects of its physical environment without degrading its electrical performance.

Space diversity
Multiple receiving antennas are used to provide diversity reception. Base-station antennas must be spaced far enough apart to achieve decorrelation.

Figure 8
Radiating element layout of an array antenna with integrated sector antenna.

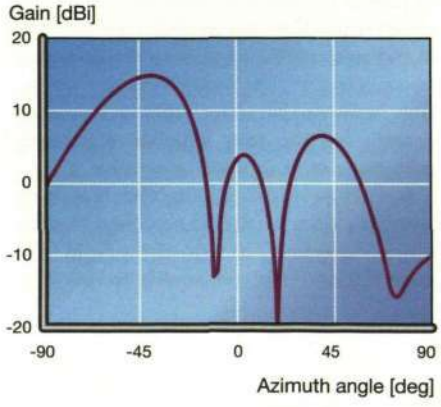


Figure 9
Beam radiation pattern: beam port no. 1.

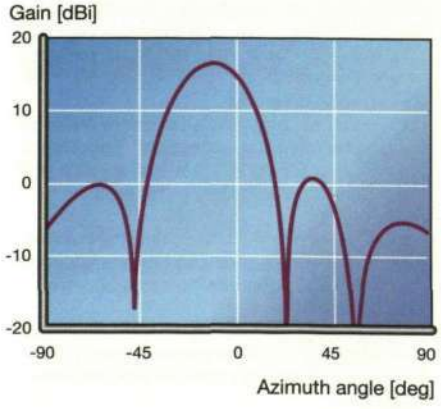


Figure 10
Beam radiation pattern: beam port no. 2.

traditional sector antenna more than twice as high.

A sparse grid of elements minimizes feed network losses and coupling effects between radiating elements. On the other hand, to maintain control over the beam pattern at all beam positions, grating lobes may not be generated. Figure 8 shows an effective element pattern layout in which the radiating element positions have been optimized to avoid grating lobes even at the outermost beams.

The antennas described in this article use Butler matrices to produce horizontal beamforming networks. A Butler matrix has an equal number of antenna ports and beam ports. For each polarization, a separate Butler matrix is connected to the columns of microstrip patches. By interleaving the beams of two polarizations, every other beam has the opposite polarization, which significantly reduces crossover depths between adjacent beams.

Figures 9-12 show the measured radiation patterns for the GSM 900 array. In particular, they show the four different beams of one polarization. Figure 13 shows each of the antenna's eight beams.

Improved power efficiency

Mobile communication base-station cabinets have traditionally been attached to pas-

sive antennas on a mast. To derive sufficient power radiation from these antennas, it has been necessary to use amplifiers with high output power and low-loss feeder cables.

Although high-power amplifiers are relatively efficient, the overall power efficiency of a traditional base station is low, since a lot of heat is generated at the base-station cabinets. Consequently, air conditioners must be installed, further reducing the total efficiency of the base station. Moreover, even when low-loss feeder cables are used, a considerable amount of power is lost in transit to the antenna as well as in the antenna power-combining/power-distribution network. A future introduction of adaptive antennas, which employ distributed power amplifiers along the antenna array close to the radiating elements, can greatly improve overall power efficiency.

Conclusion

Ericsson and cooperating operators have tested adaptive antenna systems in live GSM and TDMA networks, proving that these systems enable operators to increase the capacity of their mobile communication networks.

Moreover, the increase in gain derived from adaptive antenna systems enables operators to extend coverage from a compact antenna installation.

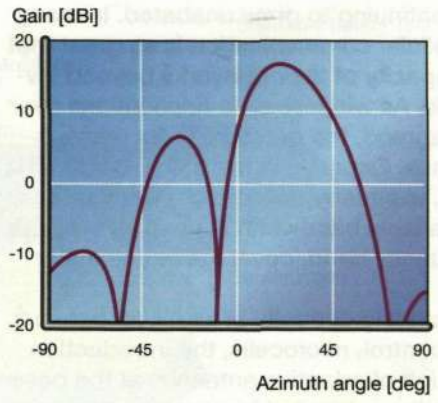


Figure 11
Beam radiation pattern: beam port no. 3.

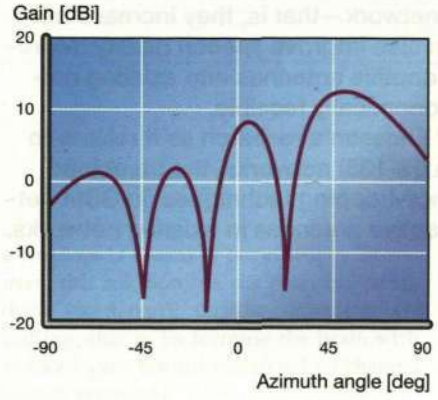


Figure 12
Beam radiation pattern: beam port no. 4.

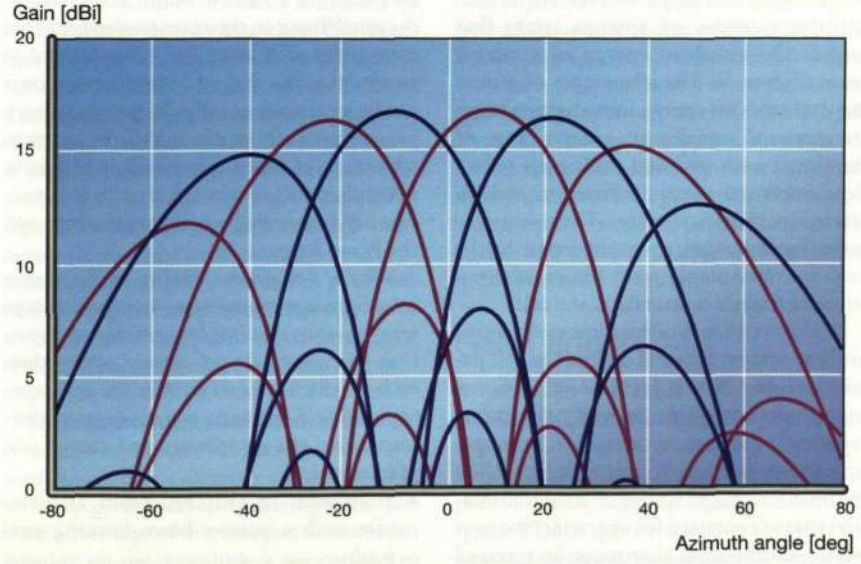


Figure 13
Measured radiation pattern of a GSM 900 array antenna showing all eight interleaved beams. The beams of one polarization are indicated by red lines; the beams of the other are indicated by blue lines.

Enhancing cellular network capacity with adaptive antennas

Sören Andersson, Bengt Carlqvist, Bo Hagerman and Robert Lagerholm

Wireless cellular communication is continuing to grow unabated. In many areas of the world, the demand for cellular communication is so great that operators are tempted to push the capacity of their networks beyond levels that current technology can handle. As wireless data applications over cellular networks become more widespread, the pressure to increase capacity will become even more intense. Capacity in the 800 and 900 MHz bands, where bandwidth is restricted, is already becoming a limiting factor. At 1800 and 1900 MHz, where available bandwidth is greater, the path loss is also greater. Thus, in this frequency band, coverage is one major aspect to consider.

There are a number of ways of enhancing capacity in a cellular network, including frequency hopping, power control, microcells, the introduction of half-rate codecs, and the introduction of adaptive antennas at the base station. Adaptive antennas have been the subject of increasing interest in recent years, and several manufacturers and operators have made them the focus of research and field trials. The main conclusions being drawn from Ericsson's studies indicate that adaptive antennas enable tighter reuse of frequencies within a cellular network—that is, they increase network capacity. Adaptive antennas can also improve speech quality. Moreover, a step-by-step introduction of adaptive antennas into existing networks appears to be practical and economically feasible.

The authors describe the results of Ericsson's research as it relates to adaptive antennas in GSM and TDMA (IS-136) networks, the combined use of adaptive antennas and frequency-hopping techniques (in GSM networks), and the implementation of adaptive antennas in existing networks.

Adaptive antennas, what are they?

Unlike conventional cellular antennas, which broadcast energy over the entire cell, adaptive antennas are antenna arrays that confine the broadcast energy to a narrow beam (Figure 1). The advantages of directing the broadcast energy into a narrow beam are increased signal gain, greater range of the signal path, reduced multipath reflection, improved spectral efficiency, and increased network capacity. There are also some disadvantages, the main one being the need to continuously track the angular position of mobile terminals in the cell.

In a conventional cellular network, a single base-station antenna defines the cell parameters and is the focus of all radiated communication. This includes the transmission and reception of revenue-generating data and voice traffic, as well as the broadcasting of system-related information that is necessary for operating the network—information that must be received continuously and simultaneously by every mobile terminal operating within the cell. System-related information includes cell identity, the frequencies in use within the cell, frequency-hopping sequences, maximum power levels, and so on.

An adaptive antenna design that increases system capacity requires that the conventional base-station antenna be replaced by one or more adaptive antenna arrays. Instead of flooding the cell with radiated information from a single source, adaptive antennas fill the cell with several narrow signal beams (typically four or eight). An immediate consequence of this new approach is that a different downlink strategy must be applied; that is, more complex information must be used for transmission from the base station to the mobile terminals in the cell. This is because the system needs to know

- which beam direction reaches which mobile terminals; and
- how it can get system information to every mobile terminal simultaneously.

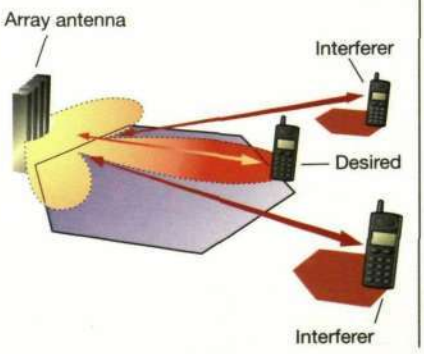
Two main downlink strategies meet these requirements. The first calls for a beam to be directed toward the mobile terminal; the second requires a beam to be selected from a set of fixed beams (Figure 2). In either case, the downlink beam relies on estimating the direction of arrival (DOA) of the uplink from the mobile terminal. The algorithms that determine the most suitable beam or beam path for downlink are thus vital elements of the adaptive antenna solution.

Several different antenna architectures (with different levels of complexity) can be used for directing radiated energy from an antenna into a narrow beam. For example, the phase front on the antenna elements that correspond to a beam can be generated at baseband using digital beam forming or it can be generated at radio frequency using a passive network or phase shifters. A main advantage of using a beam from a passive network is that it does not require phase coherency between the radio transmitter and the beam former.

While numerous system architectures exist for adaptive antennas—including separate antenna systems for uplink and downlink—Ericsson favors three approaches, each of which appears to offer the appropriate trade-off between system-level performance and the complexity and cost of implementation:

- multibeam or switched-beam architecture with a passive beam-forming network;

Figure 1
Array antenna arrangement showing the adaptive antenna concept.



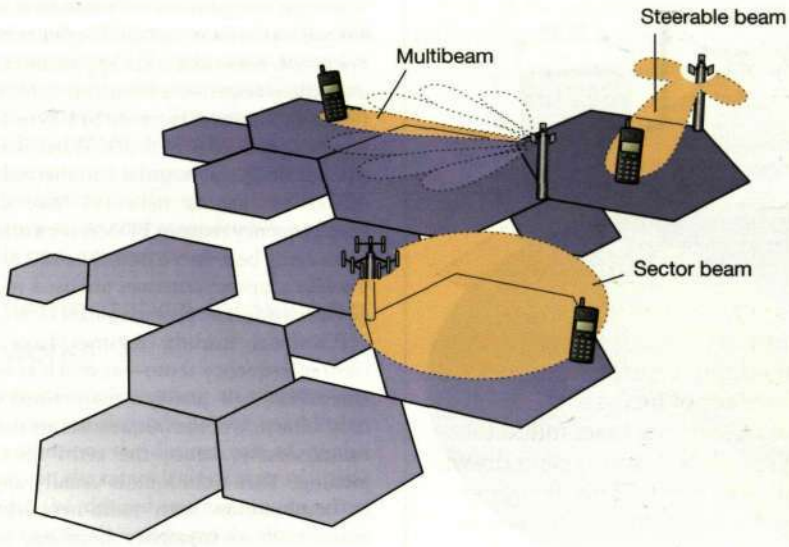


Figure 2
Steerable beam and switched multibeam antennas in network with cells.

- switched interleaved beams in the downlink; and
- fully steerable beams.

The passive beam-forming solution is the least complex one. Because the direction of arrival can identify the best uplink beam, phase coherency is not needed for the uplink or downlink.

The second solution, which requires additional downlink beams, forms beams differently in the uplink and downlink. In the uplink, the number of beams is limited by the number of receiver branches. The direction of arrival is calculated from the uplink information. This information (the DOA) is then used to select a beam from a larger set of downlink beams. In the downlink, several parallel beam-forming networks are present. After the beam has been formed, the signals to the antenna elements are combined. Compared to the steerable-beam approach, this method reduces the phase coherency requirements in the downlink. An accurate direction-of-arrival estimate might require coherent receiver branches and calibration in the uplink.

The fully steerable solution requires an individual transmitter for each antenna element plus phase coherency of the branches on the receiving and transmitting sides. The main advantage of this solution is that beam forming on the downlink is not limited to a fixed set of beams or beam shapes. Moreover, this solution has the potential to reduce interference on the downlink via nulling; that is, by forming the beam with reduced gain toward interfered co-channel mobile terminals.

Ericsson's adaptive antenna program has included extensive field trials for GSM and TDMA (IS-136) in cooperation with two major network operators, Mannesmann Mobilfunk and AT&T Wireless Services. One objective of the trials was to determine how adaptive antennas might be used in different propagation environments—urban, suburban, hilly, and rural areas. An important system-level verification from the trials is that the use of adaptive antennas enhances network quality by decreasing network interference. In particular, the narrow beams reduce received interference in the uplink and the distribution of interference in the downlink (Figures 3 and 4).

GSM and TDMA

As could be expected, differences in the GSM and TDMA standards carry over to the application of adaptive antenna technology. For example, TDMA does not currently support frequency hopping capability. Similarly, the TDMA specification requires that the base station output power on each carrier fre-

BOX A, ABBREVIATIONS

BCCH	Broadcast common control channel
C/I	Signal-to-interference ratio
DOA	Direction of arrival
GSM	Global system for mobile communication
TDMA	Time-division multiple access

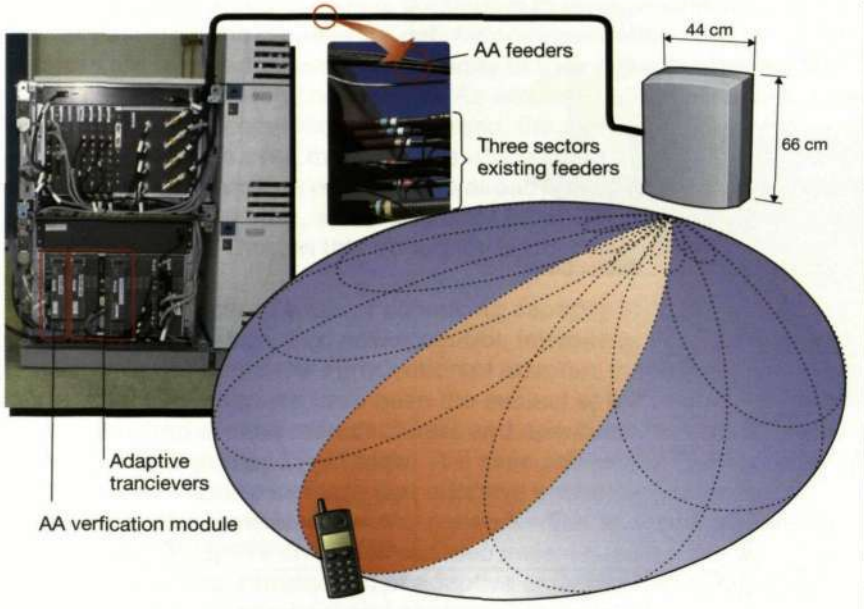


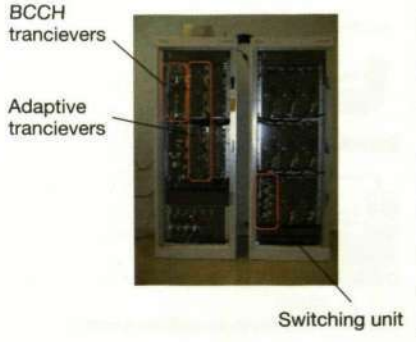
Figure3
Field trial equipment used in live-traffic TDMA network.

quency be held at a constant level for the duration of the frame once one of three available time slots is occupied. However, a proposal has been made to change the specification to allow beam forming and individual power control for each time slot. The current TDMA specification prevents the introduction of downlink beam forming and beam switching individually for each time slot. It is nonetheless possible to form beams on the downlink on a carrier basis. With a carrier-based beam-forming strategy, performance can be improved by introducing beam packing, whereby the system allocates slots on the same carrier frequency to mobile terminals that share a similar line of direction from the base station. Simulations indicate that this technique increases capacity in TDMA networks by approximately 75 to 130%, depending on system parameters.

With cellular networks, adaptive antennas offer two ways of increasing network capacity:

- using carrier signal-to-interference (C/I) gain to implement tighter frequency reuse; and
- using fractional loading.

Figure 4
Field trial equipment used in live-traffic GSM network.



The most straightforward solution is to use C/I gain to obtain tighter frequency reuse. For GSM networks, this approach can reduce the average reuse from nine to four, and typically improves capacity by100 to 120% for a C/I gain of 5 to 6 dB. What is more, field trials during regular commercial conditions in existing networks have shown that frequency reuse in TDMA networks can potentially be reduced from 21 to 12 or even 9 when adaptive antennas are used together with downlink power control.

Fractional loading regimes have even tighter frequency reuse—as much as one to three. Network quality is maintained when only a fraction of the frequencies are used simultaneously, hence the term fractional loading. This technique is usually applied in combination with radio-network features, such as frequency hopping, power control, and discontinuous transmission. The increase in capacity that results from fractional loading depends on a wide range of variables, such as frequency reuse, C/I gain, discontinuous transmission, and power control. However, field trials and system-level simulations have shown that fractional loading has the potential to increase capacity in GSM networks by as much as 280% under regular conditions.

The combination of adaptive antennas and frequency hopping in GSM networks offers the greatest potential for increasing capacity. Moreover, existing cell sites can be used to provide capacity increases over a large area.

C/I gain and fractional loading do not have the same effect on the characteristics of system interference. C/I gain brings interferers closer together, whereas when fractional loading is applied they remain in the same position.

A benefit of frequency hopping is that frequency diversity balances the quality between slow- and fast-moving users. Frequency hopping also introduces interference diversity, which improves performance. Strong interferers are shared by different users; time-varying interference increases interleaving and coding efficiency, which improves receiver performance.

Although complex to implement, fractional loading networks that use frequency-hopping techniques are efficient. This premise is supported by system-level simulations in which each adaptive antenna base station was equipped with eight fixed beams. The results showed that interference diversity is always obtained regardless of

traffic load and interference-reducing techniques, such as discontinuous transmission and power control. This is not the case for networks with conventional omnidirectional or sector antennas. The simulations also showed that GSM networks that combine adaptive antenna arrays and frequency hopping are spectrum-efficient, cope with tight frequency reuse, and considerably improve mobile tracking performance.

Implementation in existing networks

The cost of implementing an adaptive antenna solution depends on the complexity of the solution, the desired ease of implementing it, the target level of network quality, and the desired increase in capacity. Simulation trials using actual cell and data traffic supplied by Mannesmann Mobilfunk suggest that a cost-effective, step-by-step migration from a conventional antenna solution to one based on adaptive antennas is feasible. The simulation trials, which were based on existing radio networks, also showed that by installing only a few adaptive antenna base stations, operators could improve the overall quality of the network Figure 5).

Most operators are expected to approach the migration in a step-by-step fashion, since doing so is more manageable and cost-effective. The majority of today's cellular networks are composed of a mixture of large macrocells and smaller microcells. Ericsson's field trials in commercial networks were based on implementing adaptive antennas in a macrocell. Three alternatives have been identified for

- boosting capacity—interference reduction means tighter frequency reuse and an increase in transceivers;
- saving frequency spectrum—instead of increasing the number of transceivers, the current traffic can be served by fewer frequencies; and
- reducing interference—a reduction in macrocell-to-microcell disturbance makes it possible to increase capacity by reusing frequencies in the microcell layer.

The study considered cell relationships, the impact that introducing an adaptive antenna array would have on those relationships, and the addition of frequencies in the target cell and surrounding cells. Uplink performance was also investigated, as were the effects of introducing several adaptive antennas into a network. In each case, the simulations showed that capacity and quality could be enhanced, and that the introduction of only a few adaptive antenna base stations can significantly reduce interference.

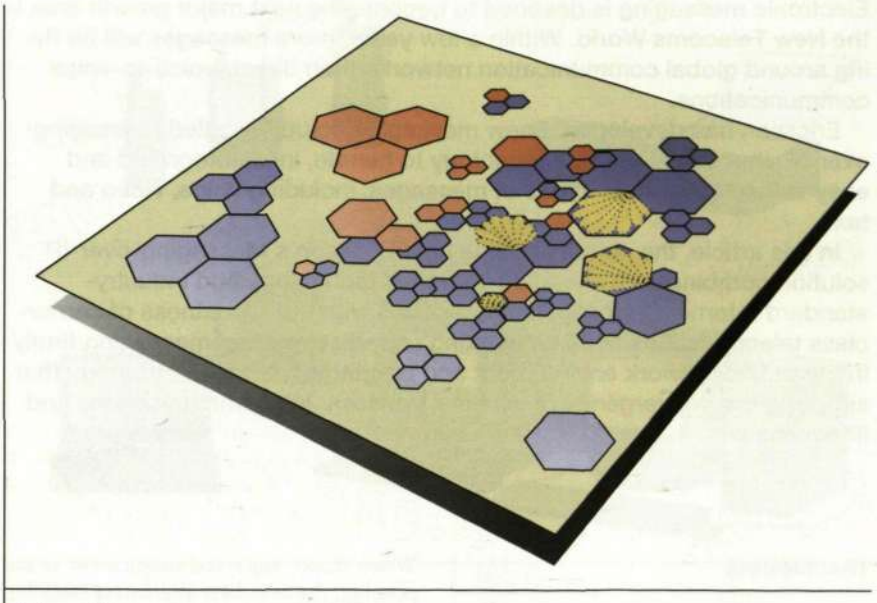


Figure 5
Migration strategy: Adaptive antennas are deployed in a few target cells. This reduces interference in several neighboring cells and substantially increases capacity.

Conclusion

Ericsson has developed a system-level concept that uses adaptive antenna arrays to meet demands for greater capacity in cellular communication networks.

In GSM networks, the combination of adaptive antennas and frequency-hopping techniques is an especially attractive solution that has the potential to increase capacity by nearly 300% at hot spots.

In TDMA networks, the combination of adaptive antennas and downlink power control has the potential to reduce the frequency reuse pattern from 7/21 to 3/9.

The introduction of only a few adaptive antenna base stations can significantly reduce interference.

Adaptive antennas make up an attractive solution that can be implemented in a practical, cost-effective, step-by-step process.

Messaging-over-IP—A network for messaging and information services

Janne Lundqvist and Bo Svensson

Electronic messaging is destined to become the next major growth area in the New Telecoms World. Within a few years, more messages will be flying around global communication networks than direct, voice-to-voice communications.

Ericsson has developed a new messaging solution, called Messaging-over-IP, that uses Internet technology to handle, in an integrated and easy-to-use manner, all types of messages, including voice, video and text.

In this article, the authors outline how Ericsson's Messaging-over-IP solution combines advanced off-the-shelf technology and industry-standard Internet protocols and interfaces with the robustness of carrier-class telecommunications networking expertise, putting messaging firmly in the public network environment and offering a messaging strategy that supports the convergence of wireline, wireless, telecommunications and IP services.

When direct voice communication is not possible, the next-best alternative is the indirect message—usually written, although more recently indirect messages are also recorded as audio messages. In the modern world of telecommunications, the electronic message is becoming a primary communication option. Driven by the widespread adoption of electronic mail (e-mail), voice mail and short message services (SMS) in the workplace and at home, electronic messaging is now growing at unprecedented rates.

Some forecasters estimate that the volume of messages flowing around global communication networks will soon exceed that of direct voice communication, which continues to grow in its own right.

Today, electronic messages typically take the form of facsimile (fax) transmissions, voice mail, e-mail, radio paging, and SMS. Each associated messaging system requires a message source, a message store, a message-retrieval device and a control system. Often, some kind of message-waiting notification system is also provided.

The source of a message can be a wireline or wireless communication terminal on PSTN, ISDN, GSM, TDMA, GPRS, WAP, and UMTS networks and systems or on a variety of computer-based networks, such as the Internet. Today, the majority of message stores are in the hands of private network operators or Internet service providers (ISP). Separate stores are used for each type of medium; for example, the home answering machine, company voice-mail devices and ISP e-mail message banks. Retrieval devices are similar to the terminals used to create the message source; that is, a variety of computer and telephony equipment (Figure 1).

The explosive growth of electronic messaging caused by the new communication culture brings new challenges and opportunities to the public network operator. The primary concerns are network capacity and the implications of managing large volumes of data traffic.

To the end-user, the variety of message sources, stores, and retrieval devices (each with its own solution) implies that he or she must learn to use several different devices to keep fully in touch. For instance, a user could have a mobile phone message service, a fax message service, a mobile data message service, a home answering machine, and a business voice-mail service. Before the potential of messaging can be exploited in full, it will be necessary to bring some order to this fragmented scene. A comprehensive solution will provide tremendous, new, value-added business opportunities for public network operators and service providers. Likewise, it will reduce the costs of ownership and of providing new and more advanced services.

BOX A, ABBREVIATIONS

ATM	Asynchronous transfer mode
DTMF	Dual-tone multifrequency
FTP	File transfer protocol
GPRS	General packet radio services
GSM	Global system for mobile communication
IMAP	Internet messaging access protocol
IP	Internet protocol
ISDN	Integrated services digital network
ISP	Internet service provider
IT	Information technology
LDAP	Lightweight directory access protocol
MEMA	Messaging enterprise management agent
MER	Messaging event repository
MIB	Management information base
MIME	Multipurpose Internet mail extension
MUR	Messaging user register
MWS	Messaging Web server
OA&M	Operation, administration and maintenance
PSTN	Public switched telephone network
RMON	Remote monitoring
SMS	Short message service
SMTP	Simple mail transfer protocol
SNMP	Simple network management protocol
SQL	Structured query language
TCP	Transmission control protocol
TDMA	Time-division multiple access
UMTS	Universal mobile telecommunications system
WAP	Wireless application protocol

Messaging-over-IP

Ericsson has extensive experience of developing messaging systems. For example, in the United Kingdom, Ericsson supplied four million easy-to-use mailboxes for Vodafone, the region's leading wireless service provider. Ericsson's Messaging-over-IP solution is a natural fit in the New Telecoms World, which is characterized by a "network of networks" made up of several access networks—wired, wireless, cable, satellite, and

so on—for all multimedia, voice, data, and Internet services. The Messaging-over-IP solution puts messaging firmly in the public network environment, relieves the pressure for capacity in traditional telephony circuits (by using an IP backbone) and is easy to use by service providers and their customers. End-users manage their own services, such as their messaging profile, which reduces the need for complex management operations, particularly in the service provider's customer-care organization.

Messaging-over-IP can incorporate the automatic provisioning of messaging subscriptions and can easily be integrated into existing management and billing systems by means of standard protocols. It can also play a key role in the convergence of wireline and wireless communications and of telephony and Internet services. With it, the many service providers around the world that offer mobile and fixed services (and perhaps IP communication) have a powerful unified messaging solution that can be used across all networks (Figure 2).

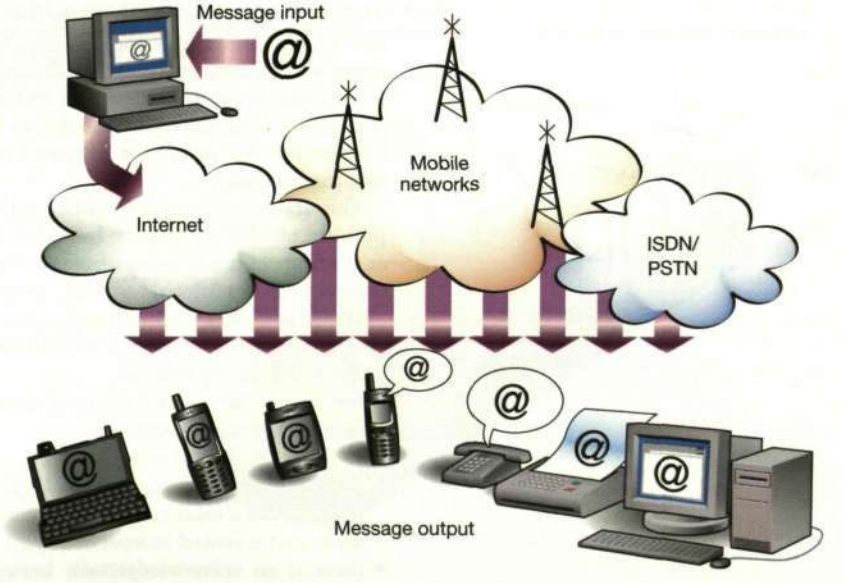
As the name implies, Messaging-over-IP uses Internet protocol-related (IP) technology to handle messages. This technology provides the characteristics of Internet services—for example, truly open standards with proven interoperability, plug-and-play hardware and software compatibility, scalability, uniform functionality anywhere in the world, ease of use, and disregard for geographic distance.

Messaging-over-IP makes use of existing IP networks that can have been built in any number of different ways. For instance, it can make use of broadband packet-switched ATM transport or narrowband circuit-switched transport. The design of the IP network determines the quality of service that can be offered. If users can tolerate delays, the Internet may be used for messaging services.

Messaging-over-IP primarily means that messaging services will integrate very efficiently into Internet services. E-mail is already fully compatible with IP. Voice, SMS, and fax messages can also be treated as e-mail with MIME-encoded (multipurpose Internet mail extension) attachments. Attachments can be opened with plug-in applications. Depending on the kind of terminal used, some media conversion might be required. Ericsson's Messaging-over-IP solution already includes text-to-voice conversion for services that can read incoming e-mail to users over fixed or mobile phones.

Figure 1
Typical retrieval devices/terminals.

Figure 2
Messaging-over-IP can play a key role in the convergence of wireline and wireless communications and of telephony and Internet services. With it, service providers can offer messaging solutions that give access across all networks to any messaging medium (for instance, to e-mail as depicted in this figure).



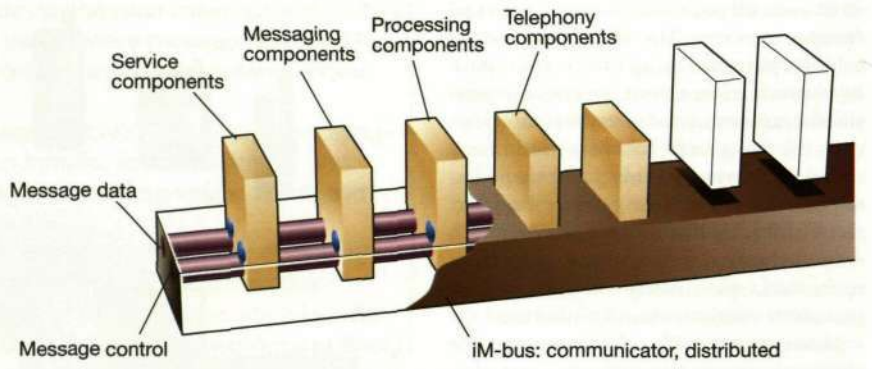


Figure 3
The iM-Bus provides a very efficient way of integrating software components.

IP is different

The Messaging-over-IP architecture is based on that of the Internet. Although important pieces of the Internet design are under continuous evolution, it can be argued that the fundamental architecture resides in the protocols. Since its inception, certain basic principles have consistently guided the evolution of the Internet. These principles can be summed up by a few key phrases, such as the end-to-end principle, IP-over-anything, and connectivity is its own reward.

End-to-end

The end-to-end principle advocates that the user should always have the final say, that trying to supplement the user with intelligence in the network is redundant, and thus that networking functions should, to the greatest possible extent, be delegated outside the network.

This argument extends well beyond the problem of relaying packets. Indeed, it lies at the heart of the Internet ideology. Compared with traditional, virtual, circuit-switched networks, the Internet (and the Internet protocol) adheres to a very different architecture. In the Internet,

- the concept of a circuit does not exist—each packet is independent;
- all control is end-to-end;
- there is no route setup—every IP datagram carries a source and destination address, and is routed independently;
- there is no acknowledgement between the client terminal and the network switch, nor are there any guarantees that the network will not lose packets. Packets are acknowledged end-to-end by means of the remote transmission control protocol (TCP). If the sending terminal does not receive a TCP acknowledgement, it simply resends the packet across the network.

IP-over-anything

The second principle of the Internet architecture is that the "inter-network" is built by layering a unique internetworking protocol on top of various network technologies, in order to interconnect the various networks. IP is layered on top of connecting networks, such as Ethernet, Token Ring, and telephone networks. Each of these provides its own set of services, and IP basically only sends packets from one point to a neighbor, who de-encapsulates the IP packet from its local network envelope, examines the destination address and, if necessary, relays it over another network.

Connectivity is its own reward

The TCP/IP suite is loaded with interesting applications. It offers good support for client/server applications on a local network. Moreover, a rich set of routing and management protocols makes for easy organization of corporate backbones. From the very start, the success and real power of the Internet can be attributed to its internetworking design.

Messaging-over-IP architecture

Messaging-over-IP is a modern, distributed, message-processing system that provides carrier-class performance. It defines an open network architecture, a system of components, their structure and relationships, and a set of design rules. Because it is based on an open architecture with industry-standard interfaces and protocols, the Messaging-over-IP solution can be built almost entirely from off-the-shelf software with little additional proprietary development. This reduces development and implementation costs, ensures short time-to-market, and enables systems to be built from the latest and best-available technology. Typical off-the-shelf components include an operating system (UNIX or Windows), databases, middleware, directory servers, messaging servers, text-to-speech conversion software, and Web servers. Getting the system to offer carrier-class performance, however, is not so simple. This is where Ericsson's expertise in communication software development, system design, implementation, integration, operation, administration, and maintenance of robust carrier-class systems comes into play.

A significant aspect of the system architecture is its scalability. As messaging becomes more popular, service providers will need to expand their messaging systems to meet demand. The Messaging-over-IP system concept easily facilitates expansion.

The heart of the concept

At the heart of the Messaging-over-IP concept—and the key to the success of a range of messaging applications, such as advanced voice mail, wireless e-mail, and unified messaging—is the iM-Bus, which defines a number of interfaces between the components that form the system and specifies the protocols for these interfaces. It thus enables software components to interact with one another. The iM-Bus is distributed over the computer nodes of the system and supports multiple operating systems.

The iM-Bus—which is a combined message-data bus and message-control bus—provides a very efficient way of integrating third-party components: each bus uses industry-standard Internet protocols. For example, the message-data bus might use SMTP and IMAP for transporting messages, whereas the message-control bus might use SNMP to manage a component. Which protocols are chosen depends on the source and type of message, system design, and the applications being used (Figure 3). Typical protocols (Box B) include

- SNMP—for operation and maintenance;
- LDAP—for directory services;
- IMAP—for message store; and
- SMTP—for notification and message transit.

System components, which are dedicated to processing, telephony, messages, services, and other functions, are plugged into the bus as required. Because the components communicate via the iM-Bus, designers and operators can build applications in manageable pieces. Components plugged into the iM-Bus are listed in a component register, which enables other components to locate and use the services they offer (Figure 4).

From local to global

The iM-Bus-centered system and the use of standard protocols and interfaces mean that compatible components can be plugged in as required by an application and network capacity. Thus, the iM-Bus is fully scalable; that is, components can be duplicated or replicated to achieve high availability. Ser-

BOX B, SYSTEM PROTOCOLS

In addition to standard, off-the-shelf hardware, Messaging-over-IP uses standard Internet protocols, the main ones being LDAP, IMAP, SNMP, SMTP, and FTP.

The lightweight directory access protocol (LDAP) reads and writes to a directory server. It is primarily used for accessing user information and resolving network addresses. It can be used to access information in directory servers outside the system as well as to allow external directory queries in the system. The protocol has been optimized for reading information and includes mechanisms for scaling, distributing and replicating information across multiple nodes.

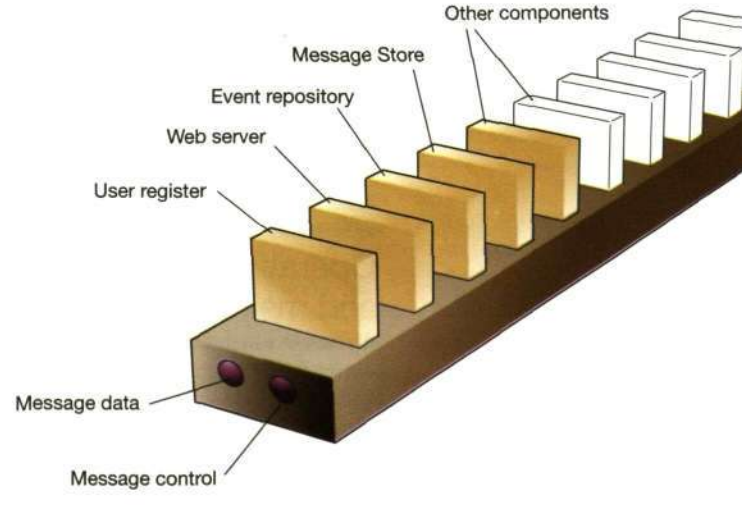
The Internet messaging access protocol (IMAP) accesses electronic mail kept on a mail server. It is used for accessing mailboxes and retrieving messages in an internal or external message store.

The simple network management protocol (SNMP), which is the *de facto* standard for management in the Internet, consists of a simple set of network-communication specifications that covers the basics of network management without putting undue stress on the network. This protocol is used for interfacing external operation and maintenance systems.

The simple mail transfer protocol (SMTP) transports messages between components, internally as well as to external systems.

The file transfer protocol (FTP) provides the reliable transfer of text and binary files between two systems.

Figure 4
Thanks to the iM-Bus, service providers can begin offering a single-medium service for a part of their network, adding new service components and extending geographic assess as the demand for messaging grows.



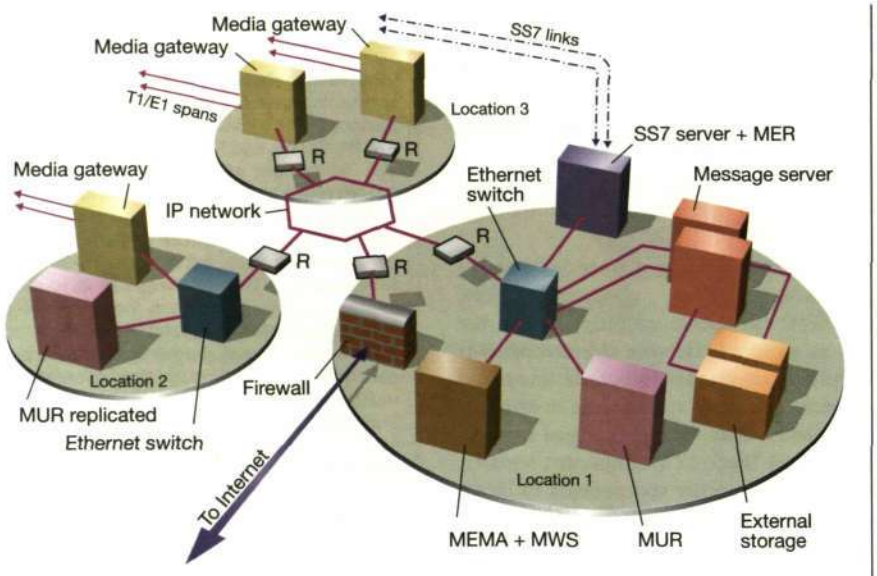


Figure 5
Messaging-over-IP can be deployed in a geographically dispersed environment using existing and new IP infrastructure. Thus, it gives public network operators the opportunity to create new value-added messaging services with immediate global access points anywhere in the world.

vice providers can thus begin offering a single-medium service for a part of their network. As demand for messaging grows, new service components can be plugged in and geographic access extended. The ultimate messaging service could be offered on a global network basis with access points anywhere in the world (Figure 5). The open, multiprotocol architecture and system design of Messaging-over-IP can be adapted to handle virtually every situation.

User access to the global messaging network is similar to that of Internet user access to global databanks: users simply dial into a local Messaging-over-IP node or local ISP; the network handles the rest. The global messaging network may also be accessed from the Internet via local ISP gateways and firewalls.

Operation, administration and maintenance

An area in which Ericsson's special expertise comes to the fore is operation, administration and maintenance (OA&M), which is essential to carrier-class operation, in-service performance, and accurate network planning and dimensioning. The use of SNMP for OA&M facilitates integration into existing OA&M systems. Thanks to a messaging enterprise management agent (MEMA)—which collects and correlates network information before presenting it to external systems—the outside world does not see individual nodes and components but instead perceives them as a single, integrated system (Figure 6).

Messaging-over-IP OA&M functionality covers the network-wide requirements for statistics, traffic data, configuration, charging and surveillance. The industry-standard protocols used for these applications typically include:

- SNMP in association with several managed information bases. For example, MIB II and remote monitoring (RMON) for device monitoring, internetwork monitoring, inventory, and management;
- FTP—for bulk data transfer; for example, for the output of traffic data; and
- SQL-2—for transaction-orientated database management.

Value-added services

Besides immediately alleviating the requirements for capacity in public voice networks, Messaging-over-IP gives public network operators the opportunity to create new value-added messaging services. For example, they can offer wireless e-mail, voice mail, or unified messaging service.

Unified messaging is perhaps the most exciting opportunity with the greatest long-term potential for generating revenue. The research organization, Ovum, predicts that the market for unified messaging will be worth $48 billion by 2002. By 2003, unified messaging services are expected to be in use by 57% of the North American population and 47% of the European population. Its first users are expected to be mobile business people and home workers, although IT-literate residential users will not lag far behind.

Unified messaging enables subscribers to use any terminal—without regard to geographic location or time of day—to access messages of any origin and format. For instance, a user who accesses his message box from a laptop computer is presented with a list of waiting voice messages, e-mail and faxes that can be downloaded. If, on the other hand, the user accesses his message box from a mobile handset, he can have the list of voice, e-mail and fax messages read to him. He can then choose to listen to any of these

messages: voice messages are played back and text messages are automatically converted to voice.

In terms of flexibility, the ultimate unified messaging system will be virtually unlimited. The technology for building such a system is already available, but until Ericsson offered its Messaging-over-IP solution, a well-designed architecture and user interfaces in a system that offers good management, control and maintenance had been lacking.

Compatibility and the future

Messaging-over-IP is already compatible with most mobile, fixed, and Internet access networks, making it a valuable tool as relates to the convergence of wireline and wireless communication networks and of telephony and IP services. Its open architecture means that new access technologies and service applications can be added with ease as they are developed. New technologies, such as the wireless application protocol (WAP), general packet radio services (GPRS) and the universal mobile telecommunications system (UMTS), will make the offering of messaging services even more user-friendly; for instance, visual interfaces will replace dual-tone multifrequency (DTMF).

Conclusion

Messaging-over-IP is deployed in a distributed messaging network that makes use of the existing transport infrastructure. This approach provides the necessary scalability to meet current and future needs. The choice of IP as the foundation for the networking of components gives Messaging-over-IP the same scalability and plug-and-play characteristics as the Internet.

Messaging-over-IP is fully ready to exploit the third-generation mobile system, and all the bandwidth enhancements that are introduced along the way (for example, GPRS and EDGE). With e-mail as the foundation of message handling, Messaging-over-IP can already handle multimedia message types, such as video and images with attachments. Furthermore, Messaging-over-IP benefits instantly as higher bandwidth is made available.

Messaging-over-IP in no way restricts access to messages on the basis of terminal or type of access network. On the contrary, Messaging-over-IP stands firmly in the center of convergence. Its component-based architecture ensures that as new access networks are introduced, new gateways to the system can be added without affecting the core of the service.

There is little doubt that electronic messaging is an exciting growth opportunity. Messaging and information services are destined to be big in the New Telecoms World, where users increasingly see their communications terminals as the access point for generalized and personalized information services, rather than just for call setup.

Ericsson's Messaging-over-IP concept is an innovative, easy-to-use solution that puts messaging firmly in the public network environment and, for the first time, creates a common platform for all forms of electronic messaging in all forms of communication network.

Messaging-over-IP enables public network operators to create a significant new business activity, by developing a portfolio of messaging and information services that generate extra revenues. It also allows them to offer market-specific service packages, thereby reducing customer churn and improving competitive positioning.

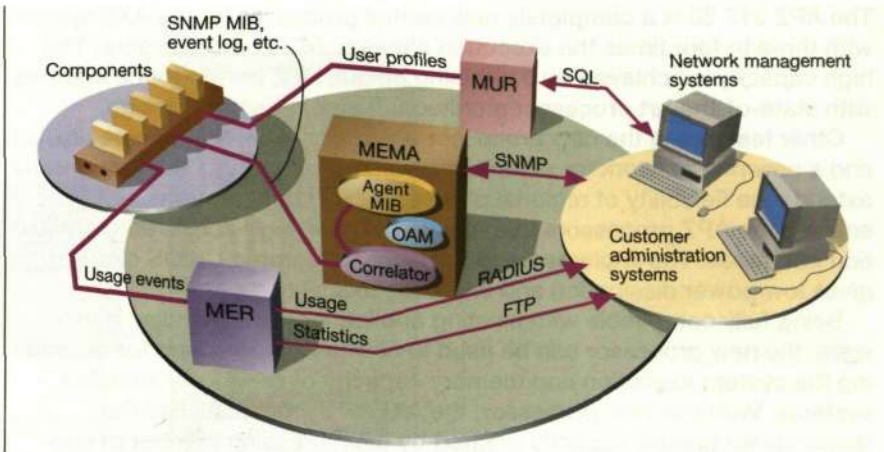


Figure 6
The messaging enterprise management agent (MEMA) collects and correlates network information before presenting it to external systems. Thus, the outside world perceives individual nodes and components as a single, integrated system.

APZ 212 30—Ericsson's new high-capacity AXE central processor

Per Holmberg and Nils Isaksson

The APZ 212 30 is a completely new central processor for the AXE system with three to four times the execution capacity of its predecessors. The high capacity is achieved by combining unique APZ performance features with state-of-the-art processor architecture and innovative design.

Other features in the new processor are improved memory capacity and a new ring network for external communication interfaces that greatly extends the flexibility of regional processor bus configurations and enables the APZ processors to make use of new, high-speed communication interfaces. The implementation of highly integrated CMOS circuits gives low power dissipation and improves reliability.

Being fully compatible with existing applications and installed hardware, the new processor can be used in new installations and for upgrading the system execution and memory capacity of previously installed systems. With this new processor, the AXE 10 system satisfies the demands for greater capacity created by the increasing number of subscribers in mobile networks and by new, revenue-generating service offerings.

The authors describe the architecture and implementation of the new APZ 212 30 processor, paying special regard to its advanced execution and communication mechanisms.

APZ 212 30 architecture

The APZ 212 30 is a completely new design (Figure 1). It retains and further enhances the unique high-performance architecture of the APZ 212 series of processors, implements a state-of-the-art execution pipeline, and introduces several new performance features. Its entire architecture has been optimized for the characteristics of telecommunications—efficient context switching, memory access and communication enable the processor to execute thousands of tasks in parallel. Instead of relying on a single processor unit to do all the work, the task of execution has been divided between two dedicated processors:

- the instruction processor unit (IPU), which executes application code; and
- the signal processor unit (SPU), which terminates protocols and schedules jobs (in conventional computers, these functions are usually associated with the operating system).

Another feature retained from previous APZ 212 processors is the pure Harvard architecture, in which the IPU has separate instruction and data caches and separate memory for instructions (program store, PS) and data (data store, DS). This design permits parallel access to instructions and data even at cache misses.

Program execution in the IPU is very advanced: instructions are decoded and executed in parallel (superscalar execution); the instructions are also dynamically reordered ("out-of-order" execution) for optimum performance. Instructions from application programs are decoded into internal RISC-style (reduced instruction-set computer) instructions. To handle jumps in the code, the processor employs dynamic branch prediction, executing on the predicted path (speculative execution). Innovative features include the pre-decoding of instructions as they are loaded into the program store, and a high-performance data store architecture.

The APZ 212 30 communicates with the AXE system via the regional processor bus handler (RPH), which implements a new ring network that allocates communication bandwidth to serial and parallel regional processor (RP) bus interfaces and high-capacity networks.

Job signal flow

The regional processor bus handler connects directly to up to 32 regional processor bus branches. External job signals (messages) that arrive over the RP bus are forwarded to the signal processor unit, which analyzes the signal, assigns a priority to it, and queues it in the job buffer where it awaits execution in the instruction processor unit. The SPU loads one job signal at a time into the IPU. When a job signal arrives, the IPU identifies it, looks up the start address of related program code in the program and reference store (PRS) table, and then begins executing the program. Programs that execute in

BOX A, ABBREVIATIONS

ALU	Arithmetic logic unit
ASIC	Application-specific integrated circuit
BGA	Ball grid array
CMOS	Complementary metal-oxide semiconductors
CP	Central processor
CPS	Central processor operating system
DMA	Direct memory access
DRAM	Dynamic random-access memory
DS	Data store
ECC	Error-correcting code
I/O	Input-output
IPU	Instruction processor unit
ISP	In-service performance
MAS	Maintenance system
MAU	Maintenance unit
MTBSF	Mean time between system failures
POWC	Power controller
PRS	Program and reference store
PS	Program store
RISC	Reduced instruction set computer
RP	Regional processor
RPH	Regional processor handler
RS	Reference store
SDRAM	Synchronous DRAM
SPU	Signal processor unit
SRAM	Static RAM
SSRAM	Synchronous, static RAM
UMB	Update and match bus

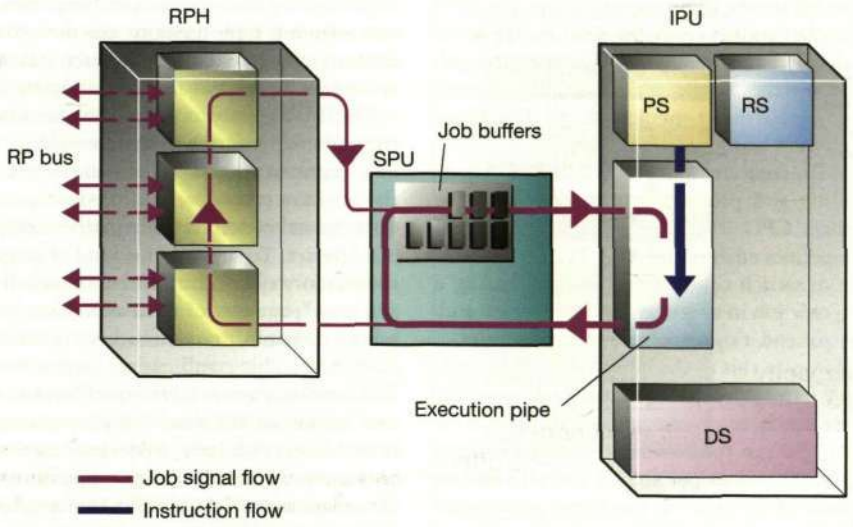


Figure 1
The APZ 212 30 architecture.

the IPU can send new job signals to the SPU. Signals that are designated to other program blocks are queued in job buffers; signals designated to other processors in the system are routed for transmission over the RP buses.

Job signal pipeline

The IPU-SPU interface has been optimized to support high throughput of job signals, which are transported through a pipeline between the SPU and IPU. While the IPU executes one job, the SPU preloads the next job signal directly into an extra bank of processor registers in the IPU. Thus, when the IPU finishes the first job, it swaps register banks and immediately begins executing the preloaded job without first having to copy registers (Figure 2).

Job signals to the SPU are also transported through a pipeline. The processor registers included in the signals are copied to a send buffer using the full internal band-

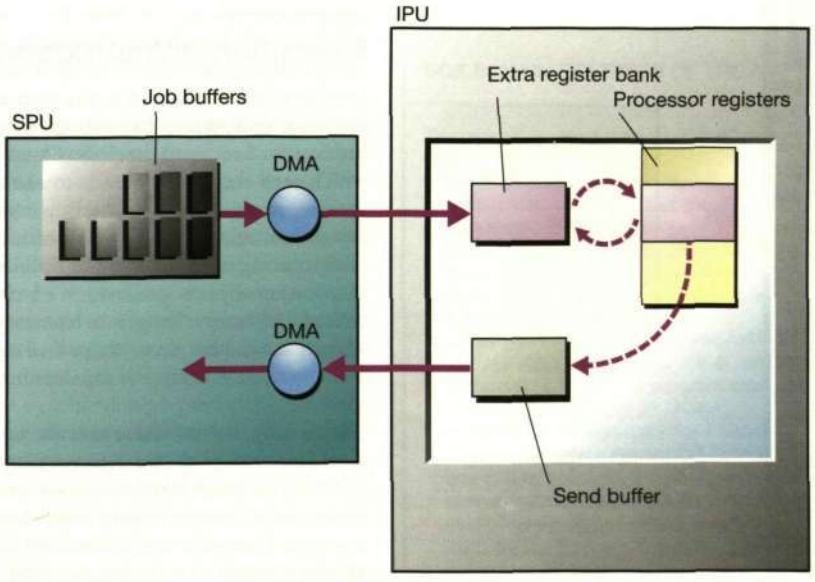


Figure 2
The IPU-SPU interface: the jobs signal pipeline.

width on the IPU processor chip. The SPU fetches signals from the send buffer while the IPU continues executing the same job or switches to the next. The SPU uses autonomous direct-memory-access (DMA) engines to transport job signals.

The combination of off-loading job scheduling and protocol termination from the main CPU to the SPU and of job signal pipelines enables the APZ 212 30

- to switch contexts and start executing a new job in as few as 30 clock cycles; and
- to send a signal in just 15 clock cycles.

Compare this to the hundreds or thousands of clock cycles that a standard microprocessor needs to do the same task. The APZ 212 30 can thus efficiently switch context 300,000 times per second and still devote most of its time to executing application code.

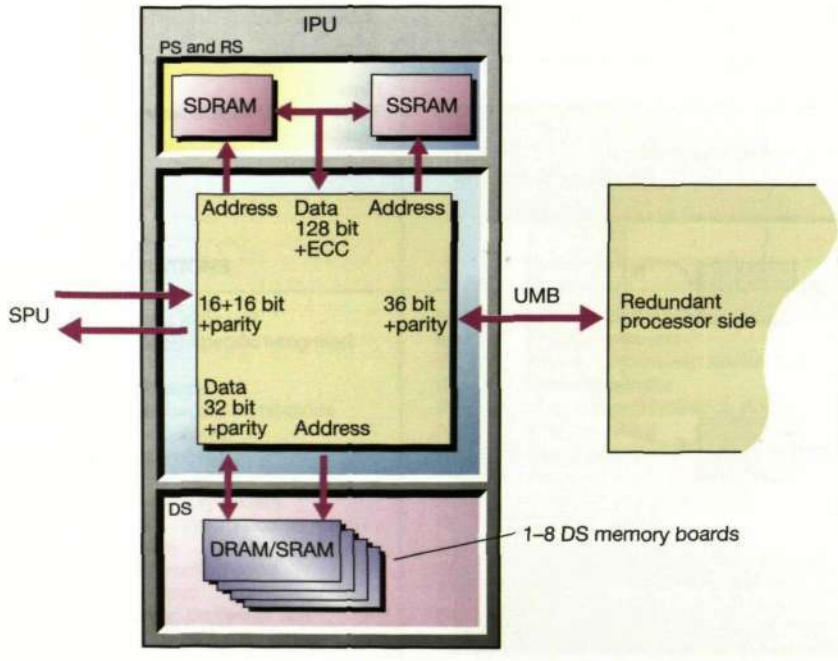
IPU structure

Compared to ordinary microprocessors, the APZ 212 30's instruction processor circuit is not limited to a single processor bus for communicating with main memory and other processors and networks. Instead, separate high-capacity buses connect to the program and reference store memory, the data store memory boards, and the SPU. Each bus operates at full processor frequency (Figure 3).

The IPU memory interfaces have been optimized to suit the characteristics of telecommunications applications. Access to the program and reference store is often sequential and concentrated to a narrow range of addresses. To support this kind of access, the memory system implements a wide bus and uses "page mode" in modern, synchronous, dynamic, random-access memory (SDRAM). This combination gives almost instantaneous access (three clock cycles) to data within an 8 Kword (16 Kbyte) range of addresses. Similarly, frequently used tables and program blocks are copied to synchronous, static RAM (SSRAM) for access in just two clock cycles.

By contrast, access to the data store is usually non-sequential and distributed between several different memory addresses. The IPU supports this kind of access by dividing the data store into banks and allowing up to eight parallel access attempts—provided no two attempts address the same bank simultaneously. The memory area on each data store memory board is divided into 16 banks, which means that one board can give full memory bandwidth in the system. The data store, which is highly configurable, can hold any combination of from one to eight memory boards. Two types of memory board are currently available: a 512 MWord (1 GByte) DRAM board, and a 32 MWord (64 MByte) high-speed SRAM board.

Figure 3
The IPU block diagram.



Execution of instructions

Instructions are executed in the instruction processor unit, which has a short, six-stage instruction-execution pipeline (Figure 4) in which each stage corresponds to one clock cycle. Instructions are shown flowing from top to bottom, with a new pair of instructions starting each clock cycle. In telecommunication applications, which are characterized by many changes in control flow (short jobs and frequent jumps and calls to other software blocks), the pipeline must be short.

Internally, the processor uses the equivalent of internal RISC microinstructions. Application program instructions are decoded into these microinstructions before they are executed. Complex instructions are decoded into a stream of microinstructions.

The first stage, called *Instruction fetch*, fetches instructions in 128 bit memory words (eight 16 bit words) from the on-chip program cache, the external second-level cache, or from the program and reference store. Given an average instruction length of 1.5 words, there are approximately five instructions in each memory word. Thus, five new instructions are loaded every clock cycle.

In the second stage, called *Partition*, up to two instructions are extracted from the memory word. These are decoded in the third stage, *Decode*. Instructions that perform simple operations—such as an ADD instruction, which adds the context of two processor registers—are directly decoded into a single microinstruction. Instructions that perform complex operations—for instance, an end program (EP) instruction, which ends the execution of the current job and switches context—are decoded into a stream of microinstructions.

The forth stage is called opread. Depending on the type of *Operand*, the microinstructions are written to one of five queues, called reservation stations. Here, the instructions wait for their operands to be fetched from the register file or from memory, or they wait for the results of earlier instructions. Up to eight instructions can be active in this stage simultaneously.

When an instruction has received all its operands, it is passed to the fifth stage, *Execute*. In this stage, up to two instructions can be executed in parallel, in separate arithmetic logic units (ALU).

The final stage, *Commit*, writes the results of the instructions to a register or memory.

The ultramodern design of the execution pipeline features the following characteristics:

- superscalar execution—two instructions can be decoded, executed and committed in the same clock cycle.
- branch prediction—when the processor performs a conditional jump, it does not wait until the branch condition is known; instead, it predicts the most probable branch and continues execution on that branch. Branch prediction is based on a very large 64 K entry prediction table used to achieve high prediction accuracy in the telecom application;
- speculative execution—execution on a conditional branch is speculative until the branch condition is known. The results from executed instructions are stored in temporary registers. If the processor failed to predict the correct branch, then the registers are cleared and execution is restarted. Otherwise, the results are committed in the last stage of the pipeline (Commit).
- Out-of-order (non-sequential) execution—if one instruction is delayed (for example, from waiting for data to arrive from memory), then subsequent instruc-

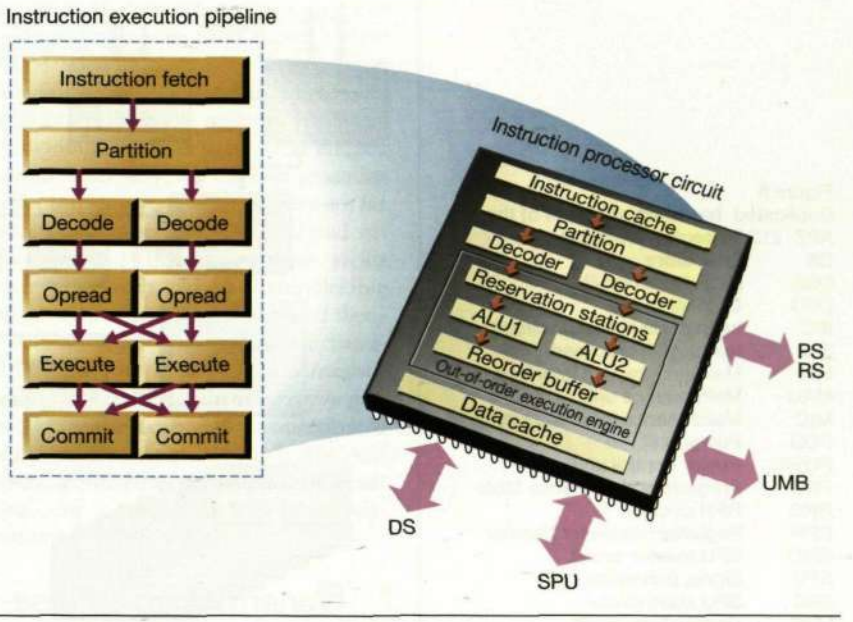


Figure 4
Instruction pipeline.

BOX B, MAIN FEATURES OF THE APZ 212 30 PROCESSOR

The main features of the APZ 212-30 processor are:

- Large processing capacity—three to four times more capacity than the APZ 212 20 (depending on application characteristics). To expand capacity, designers combined a new, advanced processor architecture, a higher clock frequency and high-speed static RAM data-storage boards;
- Large memory capacity—4 GWord (8 GByte) of data storage (up from 1.5 G Word), 96 MWord (192 MByte) of program storage (up from 64 MWord), and 32 MWord (64 MByte) of reference storage (up from 4 MWord);
- New functions—a generic communication bus interface based on a ring network that allows for adaptation to high-speed networks. A new communication-buffer mechanism allows program blocks to share
  - data in a buffer; and
  - access, using new, rapid-communication buffer read-and-write instructions.
- Optional configurations with high-speed SRAM boards and standard dynamic RAM boards for data storage. By using additional SRAM boards, the system can be configured from optimal price/capacity to best capacity.
- In-service performance (ISP)—integration into custom complementary metal-oxide semiconductors (CMOS) improves system reliability. In most configurations, mean time between system failures (MTBSF) is now more than 10,000 years. Well-defined interfaces and fewer boards improve diagnostics when hardware faults occur.
- Improved hardware maintenance—board number, revision.
- Reduced size—the processor fits into a single 600 mm, double-sided cabinet (half the size of its predecessor).

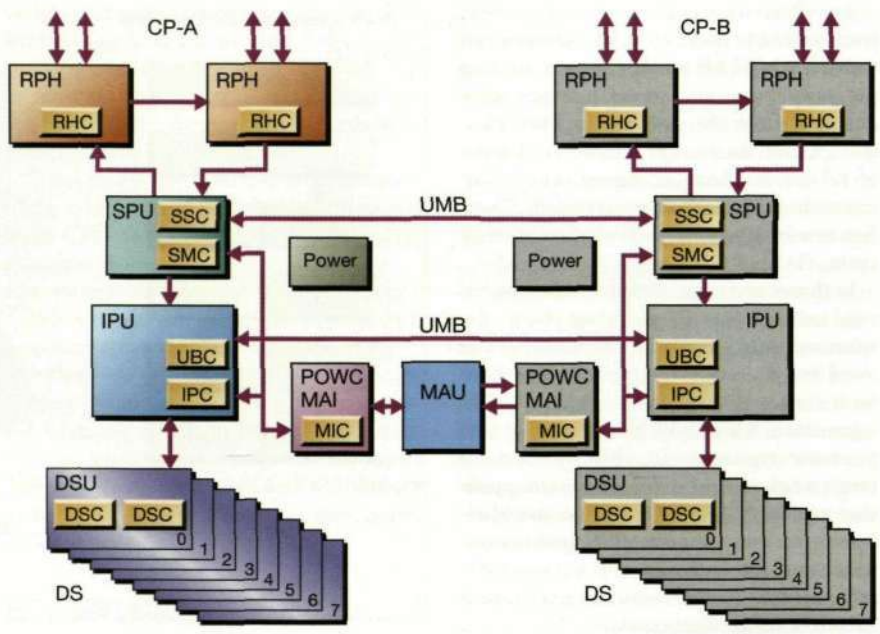


Figure 5
Duplicated hardware structure of the APZ 212 30 central processor.

DS	Data store
DSC	Data store circuit
DSU	Data store unit
IPC	Instruction procesor circuit
IPU	Instruction processor unit
MAI	Maintenance unit interface
MAU	Maintenance unit
MIC	Maintenance interface circuit
POU	Power unit
POWC	Power control unit
PRS	Program and reference store
RHC	RPH circuit
RPH	Regional processor handler
SMC	SPU master circuit
SPU	Signal processor unit
SSC	SPU slave circuit
UBC	Update bus circuit

tions are allowed to bypass the waiting instruction. The hardware first confirms inter-instruction dependency to ensure that an instruction does not bypass any instructions on which it depends for data.

- register renaming—the processor has access to more physical registers than the programmer can see. Dependencies are avoided by assigning a new, temporary register to the results of each instruction. This further enhances the capacity of out-of-order-execution;
- multilevel instruction cache system—support for fetching instructions is provided by a small, single-clock-cycle access, on-chip, level-one cache and a larger SRAM-based level-two cache; and
- data cache—support for fetching data is provided by a small, on-chip data cache together with optional, low-latency SRAM boards.

In addition to these features, the APZ instruction processor implements the following unique features to further improve capacity:

- Harvard architecture—the design of separate instruction and data memory allows simultaneous access to instructions and data.
- Load-time pre-decode—instructions are mapped to a new optimized format when loaded into the program memory. This action, which is performed by the loader in the operating system, is not visible to the user. Thanks to the new format, the IPU is able to extract two instructions from a memory word in just one clock cycle.
- Early jump extraction—jump instructions and their target addresses are identified in the partition stage before the instructions have been decoded. This enables the IPU to fetch instructions from the new path earlier and minimizes the penalty for taking the jump. This feature is especially important in processors for telecom applications, since the associated code has a high frequency of jump instructions.
- Early load-extraction—the one factor that most affects capacity in modern processors that run applications, such as a telecommunications control system (which uses a large amount of data storage) is access time for reading data from

DRAM. The new, optimized instruction format enables the IPU to identify and extract the variable address early on in the pipeline (during the partition stage), which decreases the access time for reading data.

- Loop unroller—instead of running loops in the microprogram, the loop unroller generates sequential instructions on the fly. This completely eliminates jumps in loops and improves capacity when data is copied to and from registers; when data is copied from memory to memory; and during linear search operations in memory.

Signal processor unit

The signal processor unit (SPU) is equipped with two specialized processors: the SPU master processor and the SPU slave processor, each of which is a micro-programmed RISC processor whose instruction set has been optimized for its specific duties.

The SPU master processor schedules jobs and preloads them for execution in the IPU. It also schedules periodic jobs in the system by scanning the job table and creating and scheduling the job signals to start them.

The SPU slave processor administers the RPH ring network and terminates the RP bus protocol (for example, retransmission). Outgoing messages are routed to the correct RPH. Incoming messages are forwarded to the SPU master processor.

Direct memory access engines serve as automatic data transports of signals between SPU buffer memory and the IPU as well as to the RPH.

Regional processor handler

The regional processor handler connects the central processor to the regional processors by providing interfaces to up to 32 RP bus branches. Internally, the regional processor handler implements a new ring network for communication between the SPU and interface boards. The ring network yields greater bandwidth and facilitates flexible configuration of interface boards and flexible allocation of communication bandwidth. There are currently two interface board types: one for connecting to two parallel RP buses; and one for connecting to four of the new, serial RP buses. The ring network in the RPH also supports the addition of new, high-speed data-communication interfaces (Box C).

In-service performance

With its fault-tolerant configuration, using two central-processor sides (CP-A and CP-B) that execute in parallel, the APZ 212 30 furthers the tradition of APZ robustness (Figure 5). A maintenance unit (MAU) supervises operation, selecting one side to execute and the other side to operate on standby. The standby side performs the same operations as the executing side, trailing it by 12 clock cycles. Having two sides guarantees tolerance against hardware faults and enables operators to conduct maintenance activities without loss of service. For example, one side can be extended with new hardware or software

BOX C, DESIGN AND POTENTIAL OF THE RPH/RING NETWORK

Logical design

Several point-to-point connections from the SPU to the RPH with two kinds of communication channel:

- a signaling channel
- a broadcast channel (from the SPU to every RPH)

Physical design

- Ring topology with a time-slot protocol that guarantees bandwidth per RPH.
- Synchronous clock operation supports fault tolerance in the central processor.
- 160 Mbit/s bandwidth
- Highly configurable
  - 1 to 16 interface boards of available types
  - free configuration of parallel and serial RP bus interfaces
  - (prepared for) dynamically allocable bandwidth per board
  - support for new high-speed interfaces
  - automatic configuration identifies new boards

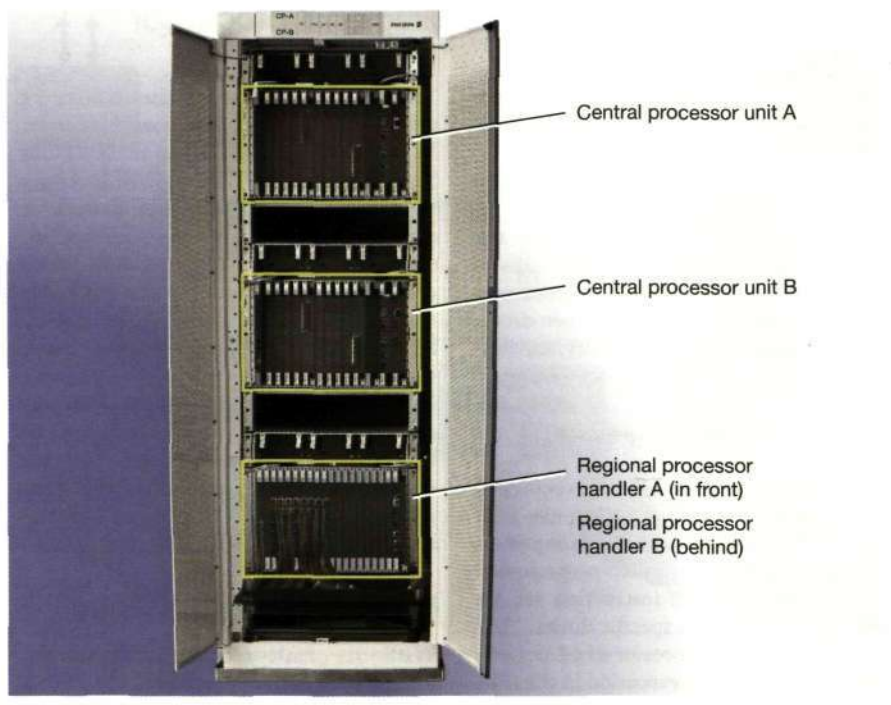


Figure 6
APZ 212 30 cabinet.

while the other side continues executing system operations.

In addition to the fault-tolerant configuration, each CP side has been designed to provide high availability. Memory (SRAM and SSRAM) is protected by error-correcting code (ECC), which corrects single bit errors. The EEC also corrects faults in whole circuits. The data store, for example, contains a large number of memory circuits, providing up to 8 GByte of memory. Consequently, if faults are detected in these circuits, the ECC corrects them.

The extensive use of application-specific integrated circuits (ASIC) makes for a very clean design (circuit boards solely contain custom circuits and memory) and reduces power dissipation. These features contribute toward exceptional mean time between system failures (MTBSF) for hardware faults—10,000 years in most configurations.

Technology

The APZ 212 30 is composed of eight separate circuit designs: one for the data store unit, two for the IPU boards, two for the SPU board, one for the RPH, one for the power controller (POWC) board, and one for the test unit (MIT trace equipment). All circuits have been implemented in high integration 0.35 micron CMOS. This circuit technology facilitated the advanced architecture of the APZ 212 30 that was needed to attain high capacity. To a large extent, the processor capacity of telecommunications applications is limited by the access time to large external data stores. At 80 MHz system frequency, the APZ 212 30 with its advanced, superscalar IPU, Harvard architecture, SRAM boards, and use of load-time pre-decode procedures, can easily keep up with memory access attempts.

The processor circuit of the IPU is housed in a 735-pin ball grid array (BGA) package. It is the largest circuit, with 2.8 million transistors in logic and 7.4 million transistors in memory.

The APZ 212 30 processor is housed in a single cabinet (600 mm wide) that holds four subracks, two for each CP side (Figure 6). The CPU subrack of each side holds the processor and memory boards; the RPH subrack holds the interface cards to the RP buses.

Software support and adaptation

Accompanying the new hardware are new releases of the operating system (CPS) and the maintenance system (MAS). These include support for the new processor hardware and functionality:

- Communication buffers—operating system support for allocating and deallocating communication buffers and for managing buffer pools.
- Data store SRAM boards—the operating system measures and allocates frequently used data to SRAM.
- Measurement functions—support for built-in performance counters and for measuring software behavior; capacity-enhancing mechanisms in the processor.
- RPH ring network—configuration of interface boards.

Software design support

While hardware designers developed the new APZ processor, other developers worked on an upgrade for the software design support:

- PLEX compiler—the new release offers better capacity and supports the new communication buffers.
- EMU CP emulator—a new version, based on new emulation technology, speeds up emulation and more exactly emulates the APZ processor.
- MIT trace equipment—this single-board trace device, which fits into a spare slot in the CPU subrack, can record every operation in the CP, including the execution of all application instructions, the execution of microinstructions in the IPU and the two SPU processors, and signals between units. This device is invaluable when debugging the new system. In real systems, collected traces give input for detailed analyses of application behavior.

Future directions

The APZ 212 30 is fully upgradable. Its advanced architecture will deliver substantially greater capacity when adapted to future silicon processes. Moreover, new features can be included for further decreasing system downtime and simplifying system handling.

Subsequent designs will make greater use of standard computer interfaces and components and support standard Internet protocols.

Conclusion

The APZ 212 30 is a completely new processor design that is now in operation in several markets around the world. It uses an advanced architecture for achieving high capacity in telecommunications applications. The processor yields three to four times greater execution capacity and extends data store capacity to a full 4 Gword (8 Gbyte).

A new, generic communication bus interface gives greater flexibility and opens the system to new types of communication buses.

The advanced processor architecture—implemented through standard CMOS technology—fulfills objectives for performance, integration, and power dissipation, and offers exceptionally high mean time between hardware-related system failures.

TelORB—The distributed communications operating system

Lars Hennert and Alexander Larruy

Future telecommunications platforms must fulfill both traditional requirements for availability and performance and increasingly stringent requirements for open-endedness and scalability.

TelORB is a distributed operating system for large-scale, embedded, real-time applications that require non-stop operation. It is composed of a modern OS kernel, a real-time database, software-configuration control, and an associated development environment for writing task-specific application code. A CORBA-compliant object request broker and a Java virtual machine run on top of TelORB.

The authors describe the TelORB operating system platform, its unique characteristics, and processing entities. These include device processors that directly control hardware with stringent real-time requirements; the TelORB operating system, which controls traffic availability and soft, real-time performance; and UNIX or Windows NT, which provide standard programming environments for less critical, real-time platform functionality and applications.

Introduction

Future platforms for use in telecommunications must offer extremely high availability, their systems must operate non-stop, regardless of hardware or software errors, and they must allow operators to upgrade hardware and software during full operation without disturbing the applications that run on them. These rigorous requirements for robustness must not affect system performance.

In the field of telecommunications, performance-related requirements are often specified in terms of statistics. For example, it must be possible, 90% of the time, to perform a certain operation within a specified period. During the remaining 10% of the time the stipulated threshold may be exceeded. Real-time performance of this kind, sometimes referred to as soft, real-time performance, is generally sufficient for telecommunications operating systems.

The platforms must also be scalable in terms of capacity: operators should be able to increase system capacity simply by plugging in new processing equipment, as opposed to having to replace the processors in a plant with more powerful ones.

Finally, there is a strong trend nowadays toward open systems. Openness, however, comes in many varieties:

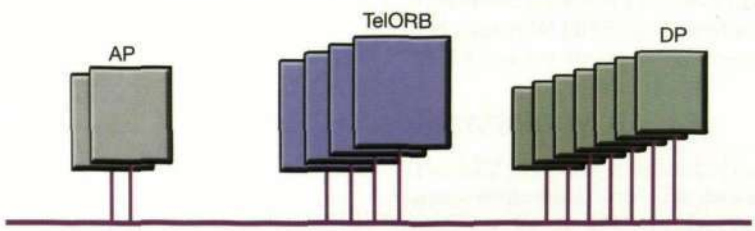
- Open hardware platform. Operators must be able to use commercially available off-the-shelf hardware. This requirement guarantees that the systems can keep up with, and take advantage of, the latest advances in hardware design.
- Programming languages. Operators should be able to hire skilled developers who can become productive quickly without first having to attend lengthy courses.
- Interoperability. Open systems must support standard protocols so as to communicate with external systems from a variety of vendors.
- Compatibility with third-party software. The application program interfaces (API) in open systems should run on standard operating systems.

TelORB-based systems support each of these requirements. However, the processors that run the TelORB operating system do not meet the last requirement; it is fulfilled by the system's adjunct processors, which use UNIX or Windows NT.

General architecture

In short, TelORB provides an environment for applications that control traffic and require soft, real-time responsiveness, high throughput, high availability (minute-per-year downtime), and scalability (in the sense that capacity can be increased by adding processors). Applications that run on TelORB are expected to serve numerous system end-users. These applications should also be permitted to evolve or to be developed continuously, thereby offering users new services. TelORB does not provide an environment for controlling small, low-cost hardware devices that require stringent, real-time responsiveness (for example, bounded, worst-case behavior). Nor does it provide a standard programming environ-

Figure 1
TelORB-based systems may consist of processors running the TelORB operating system, adjunct processors running UNIX or Windows NT, and device processors that run commercial, embedded, real-time operating systems.



ment into which third-party software can be integrated and custom adaptations quickly introduced—these environments are accommodated in the system architecture by device processors (DP) and adjunct processors (AP).

DPs are typically low-cost processors that control a handful of hardware devices. While the memory footprint of the DP operating system may be an issue, DPs require few (if any) middleware components, and the application software is fairly stable and well defined. To simplify DP software further, the DPs are owned and managed by applications running on TelORB.

The APs run standard operating systems, such as UNIX or Windows NT. They host operation and maintenance-related (O&M) platform components, such as off-line databases for complex queries, logging facilities, and management models, but can also host application software, such as purchased protocol stacks or specialized end-user services. A system may be composed of up to several hundred DPs, two or more TelORB processors, and one or more APs (Figure 1).

Within the system, one or more networks interconnect the various processors. Present-day TelORB systems use dual Ethernet, but ATM or practically any other network solution can be used as long as adequate bandwidth is provided. Different networks with different media may even coexist in the same system: the TelORB processors and the adjunct processors could, for example, be interconnected with Ethernet, whereas the device processors could be accessed through an ATM network.

The TelORB inter-process communication (IPC) protocol is used for transporting data between TelORB processors. A variation of that protocol is used between TelORB and the device processors. The standard user datagram protocol/Internet protocol (UDP/IP) and the transmission control protocol/Internet protocol (TCP/IP) are used for transporting data between TelORB and the adjunct processors. Applications can use CORBA or dialogs (when within TelORB) on top of these transport protocols. TelORB communicates directly with the device processors via the transportation layer.

Characteristics

Some key features of TelORB are its real-time characteristics and its support of continuous operation.

Real time

TelORB is intended for use with soft, real-time applications; that is, applications with load-dependent, statistically deterministic behavior. Support for stringent, real-time applications (applications with bounded, worst-case behavior) affects performance and application flexibility and is therefore left to the device processors.

Priorities and scheduling

Processes execute on one of four priority levels: high, normal, low, and background. The normal level is intended for ordinary telecom traffic applications, whereas the low

BOX A, ABBREVIATIONS AND TERMS

AP
Adjunct processor
API
Application program interface
ATM
Asynchronous transfer mode
Callback function
Function that is called as a result of an external event (for example, an incoming message on a communication link)
CORBA
Common object request broker architecture
DBMS
Database management system
Delos
One of the interface specification languages in TelORB
delux
The Delos compiler
DOA
Database object agent
DP
Device processor
DU
Distribution unit
Forlopp
Chain of interconnected processes resulting from an external event; several resources may be allocated during this chain of processes in order to handle the originating event
IDL
Interface definition language
IDP
Internal delivery package
IIOP
Internet inter-ORB protocol
IPC
Inter-process communication
LM
Load module
LPC
Linked procedure call
Managed object
Object that can be accessed from the O&M system
MI
Managed item
MIB
Managed information base
NTP
Network time protocol
O&M
Operation and maintenance
OMG
Object management group
ORB
Object request broker
OS
Operating system
OU
Object unit
Persistent object
Database object
SCC
Source code component
Scheduling queue
Queue in the kernel that holds processes that are ready to be executed
Supervisory mode
Execution mode in a microprocessor that contains the complete instruction set; application processes execute in the user mode, which has a slightly limited set of instructions
SWI
Software interface
TCP/IP
Transmission control protocol/Internet protocol
TMN
Telecommunication management network
Trigger (database)
Optional user-defined function that is called when certain events take place (for instance, when a database object is created or deleted)
UDP
User datagram protocol
Zone
Cluster of interconnected TelORB processors

level is intended for maintenance. The high-priority level should be used frugally for processes that involve the servicing of hardware. The background level could be used for audits and hardware diagnostic tests. TelORB uses a simple scheduling policy: the highest priority process that is ready to execute is allowed to execute for at most one time slice (about two milliseconds) until it becomes blocked, idle, or its time slice expires. If its time slice expires, the process is queued last at its priority level. This procedure is then repeated with the subsequent highest priority process that is ready to execute.

Obviously, average scheduling delays depend on the load of the processor. TelORB has a load-regulation mechanism that permits applications to reject parts of the traffic operations, in order to sustain real-time performance.

Interruptible kernel

By allowing operations within the OS kernel to be interrupted, the interrupt response time can be kept within reasonable bounds. However, to keep the kernel from becoming overly complex and error-prone, an interrupt-service routine restricts the number of operations that are allowed to run in the kernel. One such operation is the scheduling of a delayed interrupt-service routine, which (because it is synchronized with respect to other kernel operations) can use additional operations. Applications are advised to do the least amount of work in the real interrupt-service routine, postponing the rest of the work for the delayed interrupt-service routine.

Continuous operation

Today, more and more systems must operate non-stop—they are expected to provide year-round service, 24 hours a day, seven days a week. TelORB was designed specifically for these kinds of system. Its memory-protection hardware protects systems from ordinary application faults, and its fault-tolerant software permits in-service upgrades.

Memory protection

TelORB processes have their own memory space, which cannot be manipulated from other processes. Data in the OS kernel is also protected from manipulation by processes. Thus, errors in any given process cannot affect other processes. This minimizes the impact that a software error might have on the overall system.

Fault-tolerance implemented in software

To handle, and recover from, hardware errors and errors in the OS kernel, TelORB reconfigures the processes of a failing processor to other processors in the system. Consequently, TelORB systems do not require expensive, fault-tolerant hardware (Box B).

Network redundancy

To handle catastrophic situations (earthquakes and fire, for example), TelORB supports the option of having a redundant, geographically separate system that works in a standby fashion. The redundant system is continuously updated during normal operation and can take over immediately without any loss of data.

In-service upgrade

New software can be loaded and taken into operation while the system is running and providing service. Obviously, this requires cooperation between TelORB and the application programs that is facilitated by the framework for implementing processes. Since the same mechanisms are used for in-service upgrades and for reconfiguring the system, operators can easily verify the aspects of an application program to be upgraded.

BOX B: RECOVERY LEVELS IN TELORB

TelORB supports the following recovery levels:

- Database transaction rollback. The application is informed when a transaction is unable to commit successfully.
- Process abort/restart. If a static process encounters an internal error from which it cannot recover (for example, division by zero) it is restarted. Similarly, if a dynamic process encounters an internal error from which it cannot recover, then it is aborted.
- Forlopp recover. TelORB informs processes that are interconnected with communication links when the remote side of the link disappears. This enables the application to clean up any resources allocated to a certain chain of events.
- Processor reload. TelORB attempts to reload the processor after an error has occurred in the hardware or in software in the kernel code.
- Zone reload. If there is a risk of inconsistency in the system state—for example, when two or more processors fail simultaneously—TelORB reloads the entire cluster with the most recent backup.

Overload protection

In rare situations, events in the outside world create disproportionate load on the systems, exhausting system resources and simultaneously rendering several processors unusable. TelORB protects the system from situations of this kind by measuring the length of the scheduling queue and rejecting dialog setup attempts when the queue length exceeds set limits. In this case, system response time slows and fewer operations are carried out successfully.

Rapid last resort

Although many precautions have been taken to protect the system, there is always a remote chance that it will fail completely. For instance, several processors might conceivably fail at the same time, making it impossible for TelORB to maintain a consistent system state without reverting to a previously saved state. Thus, as a last resort, TelORB reloads all processors with a backup of the database. To minimize downtime, TelORB uses a multicast loading protocol that enables the parallel loading of all processors in the system without a bottleneck in the communications medium, and without requiring disks to be attached to each processor. If the network redundancy feature is being used, then even "catastrophic" situations can be handled without loss of data.

Program environment

Programs written for TelORB execute as one or more cooperating processes using the database to store configuration parameters and other data that needs to persist. The programs can use several operating system services through the TelORB API. The programs are written in standard C/C++ or Java, with interoperability interfaces specified in the CORBA interface definition language (IDL). Delos, a proprietary specification language, provides the constructs that standard programming languages lack for specifying processes and database objects (Figure 2).

Processes

Specification

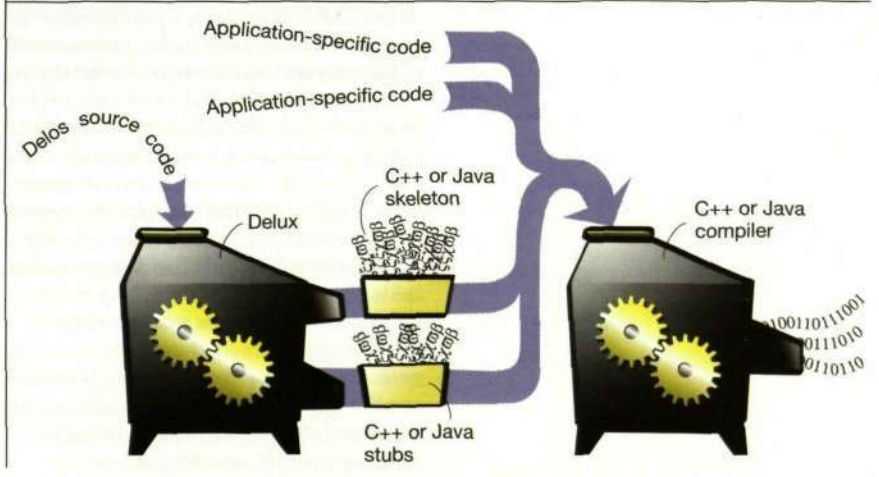
The processes that run on TelORB are specified in Delos, which assigns a name and characteristics to each process type. From the process type, TelORB creates process instances as defined in the specification.

Processes are declared as being static or dynamic. Static process instances, which are created when the system is started or when the process is installed, are recreated after a failure. Dynamic process instances, which are created when addressed by another process, are not recreated after a failure.

The process type specification indicates whether or not a process instance is to be replicated on several processors. When a process instance is replicated, TelORB directs any other process that wants to cooperate with it to its replica (if one exists) on the same processor. If a replica does not exist, the process is directed to an arbitrarily selected replica on some other processor. Note: replicas do not know of one another's existence unless required to do so by the application.

Finally, the process type specification indicates how multiple instances of the same type are to be differentiated and installed. For example, as an alternative to simply being instantiated anywhere whenever addressed, the instances of a dynamic process type can be differentiated by a primary key (that is consistent with corresponding database objects). TelORB can thus direct several calls with the same key to the same instance until it is terminated. For static process types, instances can be created automatically when the system starts, or the application software can install them through the TelORB API. This option is particularly well suited to process instances

Figure 2
The Delos source code is fed into *delux* (the Delos compiler), which generates C++/Java stub and skeleton files. The application-specific code is then manually added to the skeleton files before final compilation (C++ or Java compiler) together with other application code. The IDL source code is handled in a similar way.



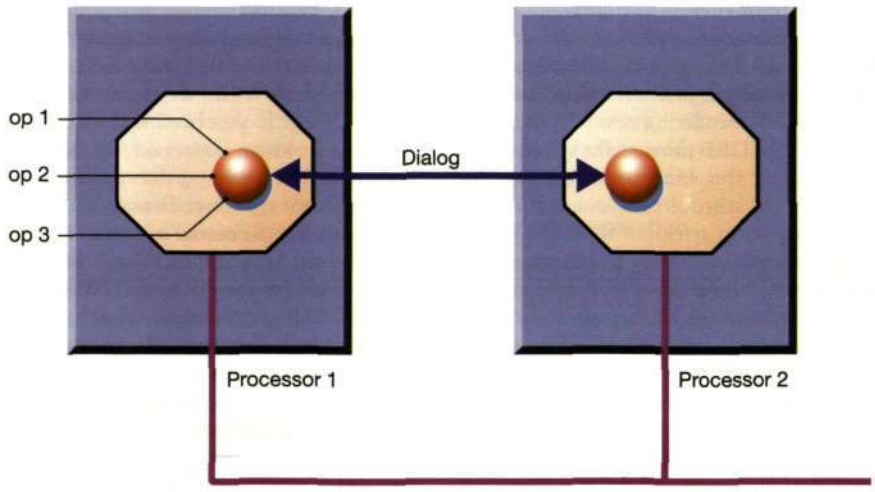


Figure 3
Processes cooperate by means of dialogs, which are specified as a pair of object types that has operations on it (in this example, op1, op2 and op3 can be called from the process in processor 1).

that are directly associated with the hardware configuration, which might vary from site to site and as plants are extended.

Implementation framework

The specified processes are implemented within a framework given by C++ or Java code that was generated from the Delos specification. The framework gives entry points for handling

- startup (initially, and after system crashes);
- the reconfiguration of instances between processors;
- in-service software upgrades; and
- termination.

When TelORB initiates a static process instance, it also indicates (in the instance) one of three reasons for starting it. Either the instance was created for the first time, or it has been recreated after a system crash. Alternatively, the instance was recreated after system reload. In this case, a serious system error occurred making it impossible to preserve database consistency. Consequently, the system was reloaded with a consistent backup of the database, which by necessity did not contain the most recent updates. The hardware state could thus be inconsistent with the state in the database. Notwithstanding, applications that can differentiate between the two usually accept the hardware state.

To facilitate non-disturbing reconfigurations and in-service upgrades, TelORB allows dynamic processes running on the original processor (with the original software) to terminate naturally over a period of time. New processes (running the new software) are created on the new processor. For static processes, TelORB creates another instance on the new processor (with the new software) that is allowed to run in parallel with the old instance. The new instance is also given a reference to the old one, so that the two can communicate—by means of an application-specific state-transfer protocol. Other processes in the system perceive the pair as a single instance. Nonetheless, if operation is not to be disturbed, certain special provisions may be needed for application-specific cooperation between the processes.

The execution paradigm

All execution within a TelORB process takes place as the execution of callback functions (most often as member functions of C++ or Java class instances) associated with events that are external to the process. Ordinarily, the program cannot choose to disable the handling of events—they are executed in the order in which they take place. Programmers must thus adapt to an inherently asynchronous outside world. The handling of all events, however, is serialized.

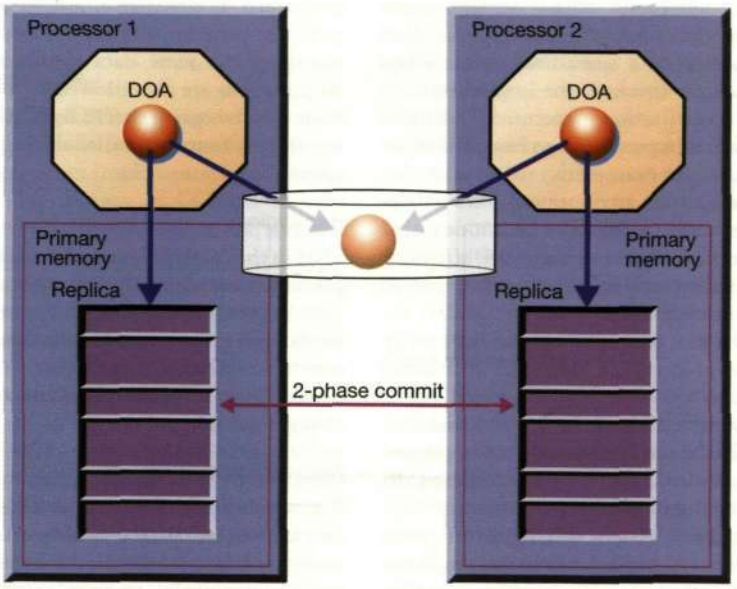


Figure 4
The database object agent (DOA) causes database objects to appear as ordinary objects belonging to a global database. In reality, however, each object is stored as a local replica in the primary memory of the corresponding processor. The database management system (DBMS) manages data access and keeps the replicated data in different processors consistent. Data replicas are updated automatically using a two-phase commit protocol.

Thus, programmers need not bother with the problems of concurrent programming (one callback function is finished before the next is called, which means there is effectively only one thread of execution in a TelORB process). On top of the execution paradigm, programmers are free to make use of lightweight threads. In Java, however, multithreading is supported as described by the language specification.

As an exception to this scheme, TelORB actually provides mechanisms for blocking the process-only execution thread. In this state, it cannot be unblocked except by a single corresponding event or time out.

Process cooperation

Processes cooperate by invoking remote operations that are collected in the Delos notion of dialogs. A dialog is specified as a pair of object types that has operations, and sometimes results, on it. C++ classes are generated from the specification. Some classes marshal and unmarshal the operations with their arguments; others provide the framework for implementing the actions of invocations within the processes (Figure 3).

To set up a dialog, one process addresses the other process by its type and, where needed, data that differentiates a particular instance. After the dialog has been set up, a connection is established between each member, to enable members to signal in the event that one of them dies. When the processes terminate, a dialog-shutdown procedure safely closes the connection.

The database

While data inside a process is volatile (it is lost if the process or the processor it runs on crashes), data stored in the TelORB database persists even after a process or processor crashes. Besides storing data, the database shares data between processes. The TelORB database is an object-oriented, real-time database that stores data in primary memory on the processors (Figure 4).

Specification

Data in the database consists of instances of persistent object types (specified in Delos) that have attributes of the persistently stored data and an associated set of methods that the application can use to manipulate the data.

The attributes include ordinary data types—integer, enumeration, record, and so on—as well as a data type which is specific to the TelORB database, and which holds references to other database objects, much like pointers in programming languages. These reference attributes may have either a single value or a multiple value. Object types may be derived from other types, in which case the behavior of the methods be-

comes polymorphic; that is, an application can open an object of a basic type and find an instance of a specialized type. When methods are invoked, the implementation of the specialization is executed. Delos also allows certain properties to be specified for attributes; for example:

- an array type attribute can have an element access property, which means that instead of retrieving the entire attribute, individual array elements can be retrieved from the database;
- a reference type attribute can have an inverse property (in this case, a reference type attribute in the referenced object must refer back to the referrer); and
- attributes can be optional, making it possible to lower the expense of storing default values for large attributes.

C++ and Java classes, which provide necessary interfaces to the database and the framework for implementing methods and triggers, are generated from the Delos specifications.

Operations

With the object types thus specified, applications can

- create new objects;
- open existing objects;
- fetch the values of, and assign new values to, attributes;
- invoke methods; and
- close and delete objects.

Database consistency is maintained by grouping operations into atomic transactions: either all operations are performed within a transaction or no operations are performed. TelORB maintains locks on objects to prevent several processes from changing the same data simultaneously. Applications are also allowed to introduce their own integrity checks by implementing trigger functions called during certain operations.

Replication

Data in the database is stored entirely in the primary memory of the processors, which makes operation very fast. To survive crashes, data is replicated on at least two processors. An extended, optimized, two-phase commit protocol is employed to replicate changes quickly and safely.

Network redundancy

The mechanisms for network redundancy are closely related to the database, since they always synchronize the database contents of the geographically separate standby side with data in the active system.

Lookup

For an object to be found in the database, it must either have one attribute designated as a primary key, which uniquely identifies an instance of a particular type, or it must be referenced from a reference type attribute of some other database object.

Objects in the database may also be found by means of iterators, which retrieve every object of a particular type (which matches the selection criteria defined for the iterator). Iterators can be applied to the entire database or to a multivalue reference type attribute. When an iterator is applied to the

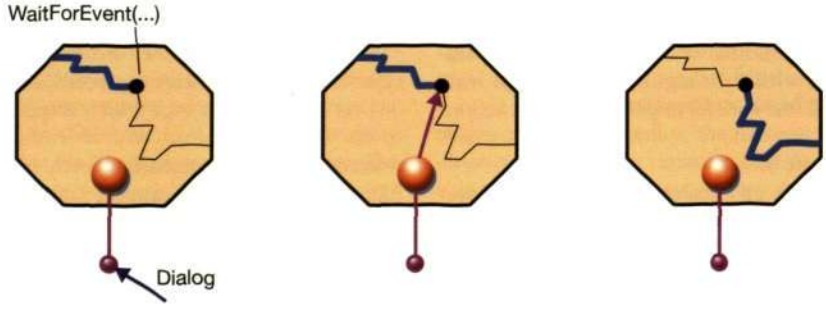


Figure 5
An example of the use of events. A method that receives a certain message in a dialog can post an event to a thread that is monitoring the event. When the thread calls the "WaitForEvent" method, it picks up any event that has already been posted or, if none, waits for an event to be posted. This renders a synchronous behavior to TelORB's otherwise inherently asynchronous programming model.

entire database, it can only retrieve types that have a primary key.

The O&M model

For O&M, TelORB requires that the specified object model be separate from the actual application implementation. This object model consists of CORBA objects, which have attributes, actions, and relations to other objects.

The managed objects of a system form a management information base (MIB), which contains information on object contents and on object types. This information can be displayed by a graphical application.

The O&M model also includes notifications, which the system sends to the operator, to inform him of particular events. A notification is always associated with a particular CORBA object. A special kind of notification is the alarm, which is sent when the system requires operator intervention (a typical example is when a processor board has failed and must be repaired).

Services

TelORB provides applications with several services, through

- the TelORB API—by means of C++ and Java classes (where standard Java classes do not suffice);
- Delos object types;
- Delos and the code generated from the specifications; and
- Corba IDL and the code generated from the specifications.

Threads and events

The lightweight threads in TelORB are designed for use with events specified in Delos. A Delos event is a type of message that is sent from an object to any thread that monitors the event. The intention has been to provide a way of adding or changing threads that execute control flows without having to change the implementation of objects that represent resources with more static behavior.

A new thread is created simply by instantiating a C++ class derived from a thread-base class in the TelORB API. The main program for the new thread consists of an overridden virtual member function. The program executed by the thread typically monitors several events from different objects and, where appropriate, waits for any of a select set of events, taking appropriate actions in response to the event received (Figure 5). These lightweight threads and events are only used for C++ code. The Java language provides its own threads and synchronization primitives.

Timers

A process can use one-time as well as periodic timers that can be set to expire in steps of five milliseconds, from ten milliseconds up to about 20 days. When the timer expires, TelORB executes a callback function. A timer can be cancelled at any time before the callback function has been executed.

Clock and calendar

TelORB provides real-time clock and calendar functions for converting the internal format into a standardized calendar format. Support for local calendars, including adjustments for daylight-saving time, has been prepared but is not fully implemented. The real-time clock is synchronized with different processors by virtue of a TelORB adaptation of the network time protocol (NTP).

Static process installation

As noted earlier, static process instances can be installed by application software through the TelORB API. Since these instances typically relate to the hardware configuration, applications must often check to see on which processor a process is to run. For this reason, the application installs a process instance for creation within a processor group. TelORB provides certain predefined processor groups, such as "all processors," but generally, each application must provide its own appropriate processor groups and register them through the TelORB API.

Handling software errors

Basically, any process in which an error has been detected is considered a bad process that should be terminated immediately—giving a crash dump for fault localization. TelORB automatically terminates processes with errors detected by hardware (dereferences of zero pointers, for example). But for errors detected by the application software, TelORB provides an interface through which information on the error can be attached to the crash dump. After termination, a static process instance is recreated in the usual way.

Processes that cooperate with a bad process are notified by means of dialog abortion indications. TelORB will not automatically terminate a process that receives a dialog abort indication, but leaves it to the

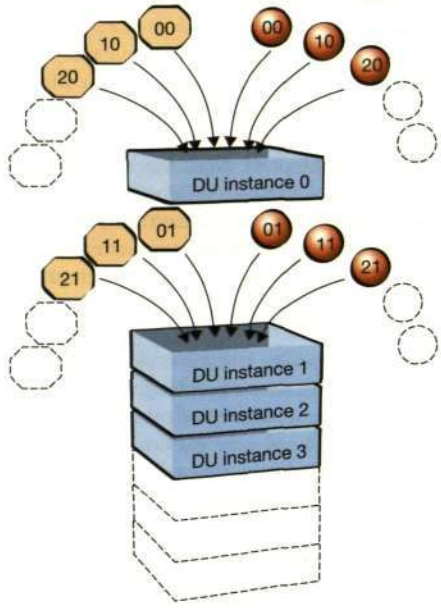


Figure 6
For dynamic processes that are identified by a key, the instances of several processes and corresponding database objects can be grouped into a distribution unit (DU) instance. TelORB guarantees that the contents of the DU are kept intact regardless of reconfigurations, thus ensuring that the processor always has local access from one of these processes to the corresponding database object.

application program to select an appropriate action or response.

Drivers

Drivers provide low-level control of hardware. They are programs that are executed in the processor's supervisor mode, which has the same privilege mode as the kernel. Applications can use their own drivers to control application-specific hardware. The TelORB API provides services for implementing drivers and for accessing them from processes.

TelORB gives drivers the following functionality: timers, installation of interrupt-service routines, interrupt level control, memory management, safe access to process data, the use of other drivers, and the ability to signal users of drivers in processes. Drivers can be opened, closed, and controlled from a process.

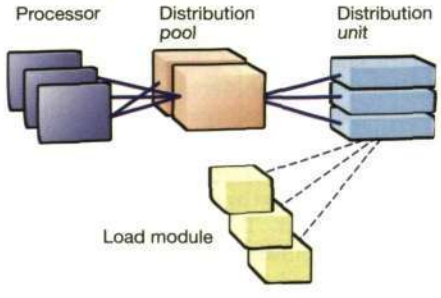


Figure 7
By attaching distribution units to distribution pools and assigning processors to the pools at configuration, TelORB can determine what code is needed to execute the distribution units on a processor in the pool.

Distribution

To make the most of the distributed processor platform, processes and data must be distributed in a way that balances processor load and minimizes communication between processors. As relates to TelORB, two simplistic approaches are either to distribute all process and database object instances arbitrarily, or to let the plant engineer specify the distribution of each individual instance. The first approach would probably result in bad performance, whereas the second approach would put unrealistic requirements on plant engineers. Moreover, each would fill memory with tables of the addresses of different instances.

TelORB allows the application developer to group individual instances into a manageable number of distribution units (DU). This grouping is typically done per process type or database object type, by associating a Delos distribution unit type specification with the specification of the process or database object type. The distribution unit type specifies the number of distribution units over which the instances should be spread. When more than one distribution unit has been specified for the type, the application program must supply a function that maps instances onto distribution units. TelORB also allows process instances and database objects to be grouped or co-located into the same distribution unit. This minimizes the inter-processor communication needed for manipulating database objects (Figure 6).

When the plant engineer configures the plant, he should group processors into distribution pools and allocate distribution unit types to pools. TelORB configures (in run-time) individual distribution units from the types to processors within the pool, and reconfigures them as necessary; for example, when a processor crashes, when the system is extended through the addition of a new processor, and when the assignment of a processor to a pool is changed (Figure 7).

Thanks to the distribution units, table sizes stay manageable. Experience gained thus far indicates that the application designer must also decide which pools to use and which distribution unit types should be

allocated to them. Therefore, this information belongs to the internal delivery package (IDP).

Memory model

TelORB divides the processor's logical addressing space into three main parts:

- instruction memory space (holds the code to be executed);
- kernel data memory space (holds data for the TelORB kernel and any drivers); and
- process data memory space (holds data for the process currently executing).

All processes share the code in the instruction memory space and can read (but not write) data in the kernel data memory space. This space is used to accelerate access to the database, and could also be exploited by drivers.

The process data space is mapped to different physical memory when different process instances are dispatched. This way, processes become isolated from each other, so that an error in one process cannot affect another process (Figure 8).

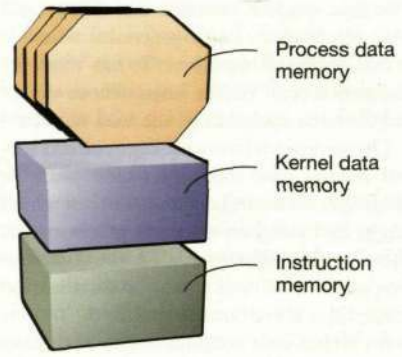


Figure 8
Although each process instance has its own memory space, all process instances share the instruction memory and the kernel data memory. Kernel data can only be written through OS calls.

Separating load modules

Code is loaded into load modules that are not directly related to process types. Although a process is implemented in a load module, it can also execute code in any number of other load modules by means of a mechanism called linked procedure call (LPC). The LPC mechanism is used for implementing database object type methods and triggers (Figure 9).

External communication

For communication with the outside world, TelORB provides common Internet protocols and the means of implementing application-specific protocols, as needed.

Internet protocols

TelORB implements TCP/IP and UDP/IP for communicating with other systems.

CORBA protocols

TelORB supports the Internet interoperability protocol (IIOP), which is transported over TCP/IP.

Development environment

TelORB comes with a development environment that runs on a UNIX workstation and includes a software structure model, build support, and tools for configuring a plant. Application programs, which can be run on the host on simulated processors, use a source-level debugger and other tools for pinpointing faults in the code.

Software structure model

The software structure model is based on managed items (MI), which have attributes (for instance, an identity and version) and relationships to other MIs. A file structure is also associated with each MI type.

Basic managed items

The most basic MI types are

- the load module (LM);
- the internal delivery package (IDP);
- the object unit (OU);
- the source code component (SCC); and
- the software interface (SWI).

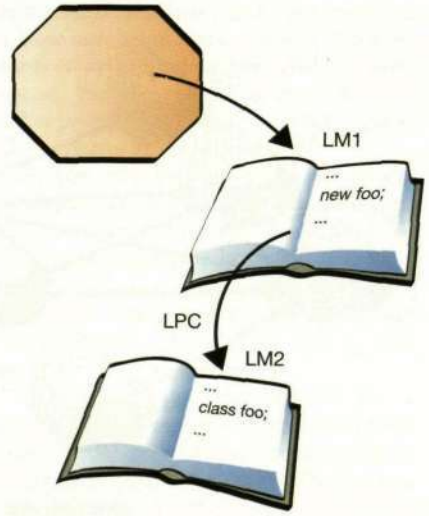


Figure 9
By means of the linked procedure call (LPC) mechanism, a process that executes code in one load module (LM) can create and use objects defined in another LM. If the new LM is updated, the new version will be used without having to relink (offline) the code that uses it.

The load module contains the object code file to be loaded onto a processor, where it is relocated and executed. It has relationships to several object units whose object code is to be included in the load module.

The internal delivery package collects several related load modules which must be used together in the target system but which might be loaded on different processors. It also contains information on the distribution pools to be used when the distribution of the DUs it contains is specified.

An object unit contains source code that is to be included in only one load module. By contrast, a source code component contains library source code whose object code is to be linked to any number of load modules. Object units and source code components provide and use software interfaces.

A software interface contains interface specifications (Delos and IDL specifications and C/C++ header files) that are shared by several managed items. When provided by an object unit, a software interface can be used by any number of object units, source code components, or other software interfaces (Figure 10).

Build support

When the source code is structured according to the model of managed items, it can automatically be built by the build support that ships with TelORB. In particular, the build support

- generates C++ and Java code from Delos and CORBA IDL specifications and files it into the appropriate directories;
- compiles generated code and application source code;
- assembles the load modules (for instance, it determines which LMs of an SCC are to be included); and
- gathers information needed by the plant-configuration tool to assign load modules to the processors.

Languages and compilers

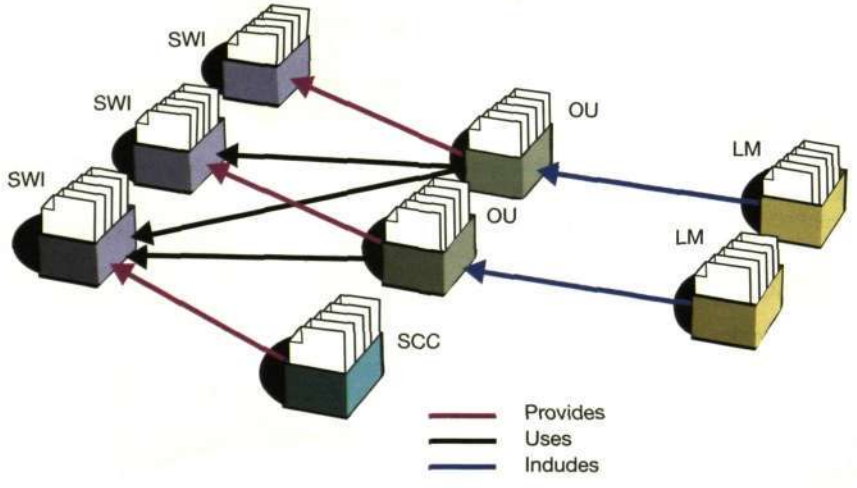
C/C++

Programs in a TelORB system are mostly written in standard C or C++. The current compiler implementation does not support C++ exceptions. The use of templates is discouraged until an acceptable way is found for dealing with them in the build support. C and C++ code is compiled with the GNU C compiler, version 2.7.

Java

TelORB provides a run-time environment for Java programs. TelORB supports a subset of Java with APIs added to enable Java programs to make use of the TelORB ser-

Figure 10
Relationships between the most common types of managed items.



vices. In essence, the subset is the equivalent of standard Java 2 with graphical support removed.

CORBA IDL

The CORBA IDL can be used to specify interfaces within the TelORB system and between TelORB and other systems.

Delos

Delos is a proprietary specification language used for specifying processes and database objects for which C++ has no constructs. Delos was originally a language family used to express interfaces, behavior (coding language), software structure, and distribution. Thus, TelORB is able to express interfaces and distribution. With Delos, developers can express data types, different categories of object types, dialogs, notifications, process types, and distribution unit types.

Delos specifications are compiled with the Delos compiler, *delux*. The compiler is divided into a front end, which parses the specifications and produces an intermediary format, and a back end, which takes the intermediary format and generates C++ and Java code plus some other information required by the target system.

Programming libraries

To simplify application development, TelORB includes the following programming libraries:

- a standard C library (minus a few functions that did not fit into the TelORB environment);
- a C++ class library, which provides lists, queues, collections, and random numbers;
- library classes that contain implementations of Delos data types, such as the string and octet string types and other support for code generated from Delos specifications.

Plant configuration

Engineers configure the plant by running a program called epct, which takes an input file that describes the configuration and outputs a file structure that can be transferred to the TelORB file system. The input file

- lists the internal delivery packages with software to be run in the system;
- defines the distribution pools;
- allocates distribution unit types to the distribution pools;
- creates representations of the processors in the system; and
- allocates the processors to distribution pools.

The output file structure contains the LMs to be loaded, a boot-load table for the first processor to be loaded, and files that specify which managed object operations are needed to take the system into operation.

Debugging

Vega—a simulated processor environment—is the main tool for starting applications. It is a UNIX process with a "hardware" adaptation layer that makes it behave like an ordinary processor. Multiple *Vega* processes can be interconnected to verify distribution aspects.

Several tools are used for debugging, including the following:

- *gdb*. The GNU debugger for high-level debugging of C or C++ code. The tool can be extended with graphics wrappers, such as *xemacs*. At present, a distributed, multithreaded, high-level debugger that supports C++ and Java is being developed for TelORB.
- *sysview*. Another graphics application with which processes can be examined as they run, and from which traces can be initiated and displayed.
- a Telnet-based inspection tool. This tool enables developers to examine the contents of the database from a remote location.

Conclusion

The market for intelligent network solutions is growing rapidly, which means that there will be an increased need for zero-downtime platforms that support a massive amount of transaction-oriented processing. Nodes in intelligent networks often need ultra-fast databases to handle requests from a large number of users. The TelORB platform is well-suited to meet these requirements. Furthermore, its exceptional scalability permits operators to gradually expand their systems as the need arises.

Because the system uses commercially available, "off-the-shelf" processor boards, operators can always take advantage of the latest achievements in hardware design.

Applications are built using well-known languages, such as C++ and Java, and interoperability is provided through the built-in object request broker.

TelORB is a truly open platform whose characteristics, in terms of robustness and flexible configurations, are unparalleled. It is currently deployed as the base for the Jambala platform.

Previous issues

No. 2, 1999

Cello—An ATM transport and control platform
WCDMA evaluation system—Evaluating the radio access technology of third-generation systems
Tigris—A gateway between circuit-switched and IP networks
GPRS—General packet radio services
Jambala Mobility Gateway—Convergence and inter-system roaming
Professional Services—Meeting the changing needs of network operators

No. 1, 1999

Ericsson's *Pro* products—Adapting mass-market technology to fit specialized needs
Open communication devices using the EPOC operating system
Mobile Advantage Wireless Office—A digital wireless office system for TDMA/136 networks
Edge—Enhanced data rates for GSM and TDMA/136 evolution
Ericsson's *e-box* system—An electronic services enabler

No. 4, 1998

WAP—Wireless application protocol
ADSL Lite—The broadband enabler for the mass market
Competitive broadband access via microwave technology
Integrating spectrum in D-AMPS/IS-136 wireless networks
Wireless LAN systems
GSM on the Net
The SDH interface in AXE

No. 3, 1998

BLUETOOTH—The universal radio interface for ad hoc, wireless connectivity
Internet directory services with click-to-dial
Jambala—Intelligence beyond digital wireless
ERION—Ericsson optical networking using WDM technology
Access 910 system


Since 1924, Ericsson Review has covered important steps in the evolution of Ericsson and the telecommunications industry.
In this interactive cavalcade of articles, pictures, movies and facts you can experience many of the most important events and milestones during the past 75 years.
1924
75
1999
Ericsson
REVIEW
ERICSSON

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

ISSN 0014-0171
Ljungföretagen, Örebro 1999