ER02s26 03-12-04 12.33 Sida 26

Combinational services—The pragmatic first step toward all-IP

Ulf Olsson and Mats Nilsson

For a long time, it has been virtually impossible to read a telecommunications trade magazine without being told that the Internet protocol (IP) is the way to the future. Although this article is no exception, it introduces a slight twist: we examine the problems IP is designed to solve, and then go on to look at what is needed or missing in order to build telecommunications networks and products. With these limitations in mind, we describe how Ericsson intends to assist operators in evolving their networks. The evolution will take place in carefully orchestrated steps. In this article, we describe the first phase—combinational services; in a later issue, we will describe the remaining steps or phase in the evolution toward all-IP.

In general, the focus is on multimedia services. That is, in addition to the voice-only person-to-person communication capabilities of today, Ericsson is introducing the ability to add images, video, browser data, game board information, and so on. One notable voice complement already in existence is the tremendously successful short message service (SMS). When building the next generation of data-oriented services, Ericsson will glean all it can from its experience of this service.

Given that the final goal is a unified IP-based environment, the industry needs to approach this goal in a controlled and economically sound fashion, maintaining good business sense along the way. For a while, it seemed as if the players in the market upheld "IP nirvana" as a goal in itself, giving little concern to the actual challenges. The recent market slowdown, however, has given us all pause to step back and rethink the entire IP issue. Might it be possible to provide all, or most, of the promised benefits without discarding everything and starting over? Indeed! This article describes the near- to medium-term steps that the industry must take to reach all-IP.

Introduction

This article presents Ericsson's solution to helping operators provide a commercially viable all-IP network. The evolution toward all-IP begins with combinational services, which enable operators to begin earning revenue from multimedia services today. A later article will cover subsequent future steps in the evolution toward all-IP. To get started, let us consider two fundamental questions: What is IP, and why do we need it? The very name—Internet protocol—gives us a clue. The Internet protocol delivers datagrams created by a source in one network to a destination in another. Hence the term *internetwork*—that is, the treating of a set of networks built with different technologies as a single structure. The main purpose of the Internet protocol is to hide details from the link-level technology, effectively shielding anything on top of the IP layer from the mechanics below. This idea of abstraction is neither new nor restricted to communications. Indeed, this same principle has formed the backbone of modern software engineering.

For the moment, however, let us stay with the technical basics while we establish a few fundamental properties of IP: packet switching and best-effort service. We will not spend much time and ink on packet switching, since numerous articles have already covered that topic. However, we do need to point out that the most interesting property of packet switching is really the gaps between packets, which can be arbitrarily long. In other words, packet switching is well suited to information flows that are inherently bursty in nature, such as the typical traffic patterns you find between an end-user who is entering data on a computer and watching the results on a screen, and the server that performs the actual calculations. Data flows of a more stationary kind, such as voice and video, need more careful attention before they can be efficiently

BOX A, TERMS AND ABBREVIATIONS

3GPP	Third-generation Partnership Project
AMR	Advanced multirate
BGP	Border gateway protocol
BSC	Base station controller
BSS	Base station subsystem
CLNP	Connectionless network protocol
DTM	Dual transfer mode
EDGE	Enhanced data rates for global evolution
GPRS	General packet radio service
Host	A node (computer, mobile terminal, dishwasher, etc.) attached to a network link (for example, an Ethernet switch port). A pure host solely generates or absorbs packet flows. It does not forward packets between networks.
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
IMS	IP Multimedia System
IP	Internet protocol
IS-IS	Intermediate system-to-intermediate system
ISO	International Standards Organization
Inter-network	A set of networks interconnected by routers.
MMI	Man-machine interface
MMS	Multimedia messaging service
Network	A set of links interconnecting nodes (host and routers) within a single address space, using a single link technology.
OPEX	Operating expenditure
OSI	Open system interconnect
OSPF	Open shortest path first
PTT	Push to talk
RAB	Radio access bearer
RFC	Request for comment
RIP	Routing information protocol
Router	A special-purpose node that takes packets off their connected networks and retransmits them on other networks, preferably closer to the destination. Apart from forwarding packets, routers exchange network reachability information using several different protocols (RIP, OSPF, BGP, IS-IS, etc.) and use this information to build the routing tables employed by the forwarding function.
SCTP	Stream control transmission protocol
SIP	Session initiation protocol
SMS	Short message service
TCP	Transport control protocol
UDP	User datagram protocol
WCDMA	Wideband code-division multiple access

ER02s27 03-12-04 12.33 Sida 27

transported using IP. We refer to this as Problem 1: *The wrong flow pattern*.

We also mentioned best-effort service. What is this? Should not all communication be given the best possible effort by the involved equipment and operators? Certainly. But in this context, the term refers to the notion that packets are delivered from source to destination without any guarantees regarding the probability of successful transmission, delay, and so on. The reason for this seemingly odd restriction in the level of ambition is that the IP layer hides the transmission technology from the upper layers. Accordingly, the mechanisms available on a certain type of link layer might not be available on another. One mechanism might be optimized for throughput, whereas another might provide error detection, retransmission, guarantee sequences, or some other kind of service quality. On the IP layer, we cannot assume the existence of such help, since the packet might traverse all kinds of different links during its journey. The most important property of IP is that it is the point of convergence for every kind of link and use: "anything on IP, IP on anything." But this also means that IP is the least common denominator for all technologies concerned. We call this Problem 2: *The hidden bearer*.

Obviously, we are not interested in a world with no guarantees of delivery. Think, for instance, of file transfer: a single missing packet renders a 100 MB download useless. The protocols that guarantee safe and ordered transmission, or any other desired transport semantics, are typically built above IP, in the transport layer. Here we find, for example, the transport control protocol (TCP), user datagram protocol (UDP), stream control transmission protocol (SCTP), and literally hundreds of other more or less well-known members of the IP family. This is both bad and good news. Bad, because what looked deceptively simple—a standard, easy-to-understand packet-forwarding protocol—needs all kinds of support technology to make it useful. And good, because all these protocols rely on IP as the only common network protocol: if your network can transport IP packets, then the upper protocols require no additional support.

This ketchup-bottle effect (once you go with IP, you get a vast range of higher level protocols in the bargain) is actually the main reason why it is fundamentally a good idea to introduce IP as a cornerstone of telecommunications networks. Indeed, the decision would have been easy were it not for a few snags. In particular, the link technology in mobile networks has some unusual properties. Although we stated above that IP hides the link layers, one property that is usually taken for granted is that a link is either 100% up or 100% down. Radio links, on the other hand, are notoriously unreliable—bits get corrupted, packets get lost, and connectivity is frequently interrupted. Sometimes, this happens by design, for example, during general packet radio service (GPRS) cell reselection; sometimes, by accident, such as when the terminal temporarily leaves the coverage area. To make matters worse, the bit rates associated with radio links are much lower than what wireline technology can deliver, and the gap is widening. This, in summary, is Problem 3: *The narrow, leaky pipe*.

Market drivers

The IP family was created by a loosely organized set of enthusiasts (the Internet Engineering Task Force, IETF) who have resoundingly displaced all attempts by more traditional bodies to deal with Internet standardization. During the past few years, the family of protocols based on IP has grown organically, as problems and issues were stated and solved. At times, this growth came in response to an identified need in the community; at other times, it merely addressed an interesting aspect of technology. This has made it difficult to predict—and influence—where Internet technology is heading. The IP community embraces a minimalist attitude toward the process of creating protocols, which means that

- each piece of the IP puzzle is reasonably simple to understand;
- the puzzle is made up of many, many pieces, and
- the IETF creates the pieces (the protocols), not the puzzle (the network architecture).

The process described above has the advantage of being productive and focused on individual tasks. Also, given that IP networks started out providing best-effort services only, some of the swiftness of development can be attributed to the fact that solely the endpoints needed to be involved when a new protocol was added. The routers along the path of the packets were not affected. Now, however, as IP technology is being applied to more challenging problems, such as the three specifically noted above, other parts of

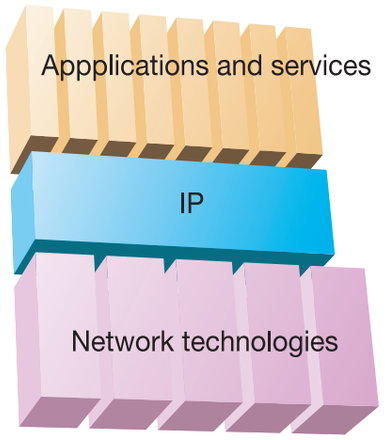


Figure 1
IP over everything, everything over IP.

BOX B, THREE FUNCTIONAL CATEGORIES OF IP TECHNOLOGY

IP technology is not a homogenous field. In general, there are three functional categories:

- host functions—that is, all the applications and upper-level protocols involved in creating or absorbing packet streams;
- packet forwarding—that is, the technology that allows packets to be moved from an ingress to an egress link with minimum delay; and
- routing—that is, the protocols and implementation strategies that efficiently and robustly move topology and reachability information between routers.

The first two categories are core Ericsson technology, since they directly influence efficiency and quality in end-user services. Routing technology is different: a typical telecommunications node needs to participate in routing to the level that flows can be established and maintained. Typically, the nodes in a telecommunications network are housed in sites (as determined by end-user demographics and the relative costs of transmission and equipment) that are interconnected with a backbone network. The process of interconnecting backbones that belong to different organizations for the purpose of interchanging traffic is known as peering. The routing mechanisms used to interconnect sites and implement peering across points-of-interconnect tend to be much more complex than what is needed within a site. Before a router can be added to the core of the Internet, it must be tested and trusted. Accordingly, the router market tends to be conservative.

the network must be redesigned. The relative rate of innovation is thus hard to maintain. In particular, in those areas of interest to telecommunications and mobile communication. Ericsson can play a useful role by serving as the bridge between the 3GPP (and other classical standards bodies) and the IETF.

Another argument for IP has been the notion that IP networks are inherently less expensive to build and maintain. This is basically true for networks that solve a straightforward best-effort problem. And in many instances, best-effort service truly is good enough, since the availability requirements for a typical enterprise network are far from the 99.999% availability expected of a telecommunications system. Consequently, it is not surprising that IP networks to date have been relatively inexpensive: they solved a simpler problem. Interestingly, some case studies in the press provide anecdotal evidence that the deployment of IP phones is at least as costly as building and maintaining a classical switch-based voice network. In the long run this will not be so, since the technology is maturing, and in particular, the operating expenditures (OPEX) part of the equation has the potential to be much lower.

The target vision

Ericsson's long-term interest in IP extends far beyond that of a mere technology shift in existing networks. Indeed, Ericsson envisions using IP to provide a mobile communication system that yields true multimedia services to its end-users. The structure of this vision has already been outlined in the form of the IP Multimedia System (IMS) defined by 3GPP (Releases R5, R6 and beyond).

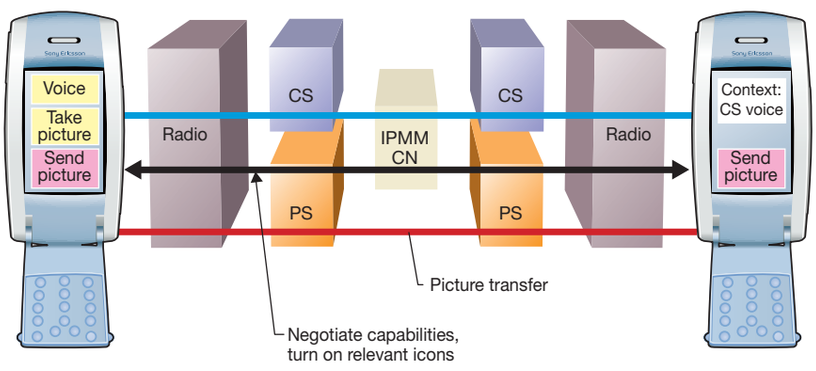
In this scenario, applications (server-side in the network; client-side in the mobile terminal) will have a unified, IP-based interface to the underlying transport machine. This cannot be the simple, best-effort interface of today. Instead, mechanisms must be added to signal the needs of the application to the lower layers, in order to make optimum use of the narrow, comparatively unreliable radio link. Ideally, these mechanisms will allow applications to adapt themselves to the properties of the radio link without expressing the properties in a way that is specific to radio technology. Servers deliver content to many kinds of client; therefore, the less they have to adapt to specific clients, the broader the market they can cover. This constitutes a major step toward reducing OPEX, by stemming the constantly escalating cost of hiring and retraining competent network personnel.

Even if all traffic is successfully moved onto a single network, will the network be able to support the many different kinds of traffic with widely varying delay, bandwidth, privacy, and delivery requirements? Most likely. But new protocols will have to be added to the mixture, introducing a number of network engineering challenges. In summary, although the base technology is homogenous, it still presents a multidimensional problem for operations personnel.

Security is also important. Privacy is becoming a key requirement through legislation and strong public demand. To be as trustworthy as traditional, closed, circuit-switched networks, the IP-based networks must be protected against all kinds of attacks, externally and from within.

In addition, the cost of producing a phone call will not disappear entirely, not even for multimedia calls. Studies show that people are more willing to pay for person-to-person calls than for person-to-content calls. The cost of distance is all but gone thanks to excess capacity in dark fiber, but the cost of delivering quality, priority service remains, adding to the end-user and operator requirements for secure and reliable systems.

Figure 2
An example of combinational services: sharing pictures during conversation.



The starting point

At the moment, the mobile communications industry has two main technologies at its disposal, giving rise to a sound business

that provides for a genuine public need. These are

- circuit-switched voice and video telephony, which provide connectivity to any phone at surprisingly low cost; and
- packet-switched access, in the form of GPRS or similar technology. Although the standard allows for some level of quality-of-service control, in practice, this has not been broadly implemented.

As we have already discussed, the current state of the art is not only about possibilities but also comes with a few significant limitations:

- current packet-switched radio characteristics make it difficult to use IP over the air for voice. In particular, handover delays are much higher than for circuit-switched telephony. This is not a major issue when the end-user is browsing, but it is very noticeable during conversations.
- packet-switched links tend to be asymmetric, with much better bandwidth (and slightly better delay properties) in the downlink. However, if the data stream is voice from a mobile terminal, then the uplink characteristics become the limiting factor, and excess downlink capacity is of no consequence.
- current GSM/GPRS terminals can handle circuit- and packet-switched traffic, but not both at once. Although the standard defines Class-A mobile terminals that overcome this limitation, implementing them is prohibitively complex and expensive.

This final limitation will be the first to be resolved. In WCDMA, we are already able to set up multiple, parallel bearers over the air interface (multiple radio access bearers, multi-RAB) and use them to provide simultaneous real-time flows over circuit-switched connections, and interactive flows over packet-switched connections. And now, GSM has a standardized mechanism—dual transfer mode (DTM)—that yields similar possibilities.

The technology building blocks

The ability to simultaneously handle circuit-switched and packet-switched traffic permits us to create what Ericsson calls combinational services. In many use cases, the end-user experience is every bit as good when the real-time part travels over a circuit-switched link as when implemented on an ideal conversational IP bearer.

Building block 1: multi-RAB

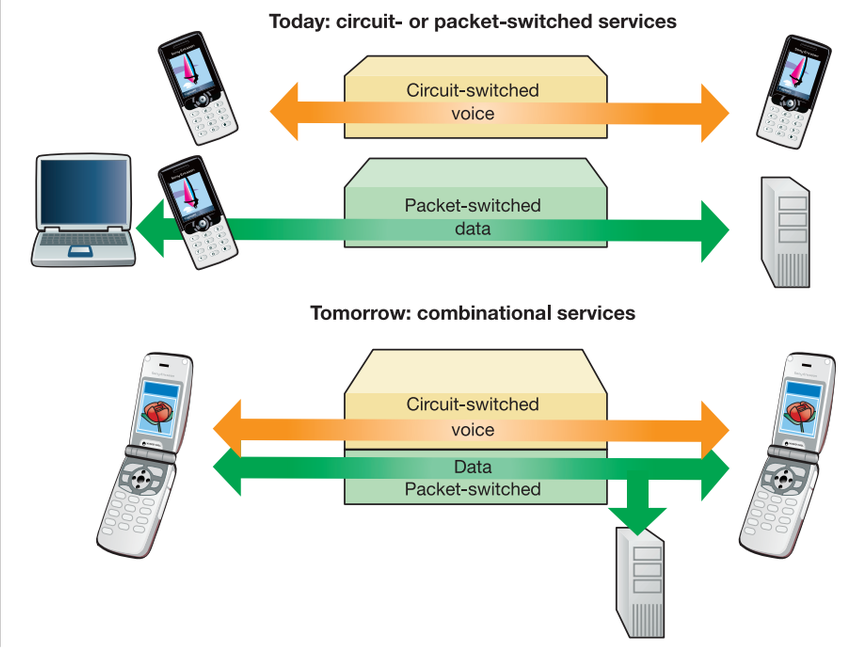
We do not describe multiple radio access bearers (multi-RAB) in this article, since WCDMA technology has been thoroughly covered elsewhere. However, we should mention that the recent focus on combinational services has stressed the importance of thorough interoperability testing of mobile terminals and networks.

Building block 2: DTM

Thanks to DTM, combinational services can also be implemented on GSM. DTM technology is a clever way of allocating radio resources (frequencies and timeslots) to circuit- and packet-switched traffic, so that terminals can be built using a radio subsystem that is only slightly more complex than a standard GSM/EDGE terminal.

Although the GSM/GPRS standard allows for simultaneous circuit- and packet-switched traffic in a GSM terminal, the general solution calls for a Class-A terminal, which requires two complete radio frequency and signal-processing sections. The timeslot used for circuit-switched traffic

Figure 3
One network for multiple uses: person-to-person and person-to-content communication.



ER02s30 03-12-04 12.34 Sida 30

and the timeslots carrying packet-switched traffic (packets) can be on different frequencies, making two radios a necessity. This translates into prohibitively expensive equipment and short battery life. Class-B terminals (what we find on the market today) temporarily suspend the packet-switched session when handling a circuit-switched call. The packet-switched session can be re-established without loss of state (same IP address), but packets cannot be transferred during the circuit-switched call. Therefore, if you want to send a multimedia message (MMS) while you are conversing on the phone, you must first hang up, transfer the MMS, and then call back. Compare this to SMS: it is possible to send and receive SMS during a call, so why not MMS? By means of careful slot allocation, DTM enables a single radio to have parallel, simultaneous circuit-switched (CS) and packet-switched (PS) capabilities, allowing you to send MMS without interrupting an ongoing phone call. This is more convenient for the end-user and will in all likelihood prolong the voice call, which translates into increased operator revenue. Several classes of terminals have been defined for DTM. Table 1 shows those most likely to be selected in practical use.

The advantage of Class 1 is that it facilitates scheduling by the base station controller (BSC). Classes 5 and 9, on the other hand, are easier to implement in the radio network and terminals, since they reuse existing protocols and procedures. The emerging industry trend is for early implementations that support Classes 5 and 9, followed by Class 1. The Ericsson GSM roadmap is well aligned with this scenario.

Building block 3: reachability and capability negotiation

Unfortunately, successfully crossing the air interface is not enough. We sometimes forget that the average end-user is not especially interested in the intricacies of channel coding and wave propagation. Instead, end-users want a mobile terminal that is reliable, simple to use, and well adapted to the current context. In other words, some entity in the mobile terminal must interpret what the end-user is trying to do and translate that into a sequence of operations. Let us assume, for example, that a woman and her husband are having a phone conversation about a striking garden exhibition. During the course of the conversation, the woman decides to show her husband what she has been describing. Ideally, the man-machine interface (MMI) should be simple enough that she need only press a camera trigger. The mobile terminal should contain enough intelligence to figure out how it is to reach the other party over a packet-switched connection and send images. For this to happen, the following building blocks are needed:

- a coordinating function in the mobile terminal (at very least, an extended address book with reachability information for all relevant networks);
- a reachability mechanism on the packet-switching side. Different options are being discussed in this domain. The long-term solution will probably be based on IMS using the session initiation protocol (SIP) to find the other party and negotiate session parameters. Note, however, that MMS technology already contains the reachability mechanism, and might therefore serve as a useful technology for launching the market. In the end, IMS provides a more generic base for future development, so all early solution deployments should have a clear path toward the final state; and
- a mechanism for distributing capability information that allows terminal- and network-based applications to make intelligent use of information on the subscription, session state, bearer states, end-user preferences, and so on. Many alternatives are being considered in this area: among other things, SIP, HTTP, and XML Web services.

Application designers should not be forced to re-invent these basic mechanisms. Indeed, in crossing the chasm from interesting technology to market growth, the roles of the service layer enablers, IMS, and associated service creation environments (programming environments) are just as important as the bearers. Note also that services are implemented both in the terminal and network servers. The terminal side is growing steadily in importance, which means that Ericsson is paying careful attention to the timely introduction of terminal as well as network solution components.

In addition, no feature or service may be difficult for end-users to install or configure. Experience of GPRS rollout has shown that if a service is not simple and intuitive, end-users will probably not even try to make it work. Autoconfiguration and over-the-air downloads can improve market uptake of basic features and promote the growth of

TABLE 1, EXAMPLE DTM CLASS DEFINITIONS

Class	Circuit	Packet
5	One (or one-half) timeslot	One timeslot uplink, one downlink
9	One timeslot	One timeslot uplink, two downlink
1	One-half timeslot (using AMR half-rate)	One-half timeslot

ER02s31 03-12-04 12.34 Sida 31

new services. Imagine, for example, a new offer each morning: "Care to try our new "Gardening World" service? Download and test it for free for five days!"

In summary, the pieces that are needed to complete this puzzle are basic air interface capabilities, service coordination, reachability and negotiation, distribution of capability information, full support of these items in the terminal, and appropriate development tools.

The steps

Given the basic strategy of providing operators with the tools they need to get to market early, Ericsson's approach is to combine the building blocks in a step-by-step fashion.

Initially, one might consider current opportunities (without WCDMA or DTM). One example is to add push-to-talk (PTT) communications capabilities to, say, multiplayer games built around packet-switched technology.

The main step, however, is to deploy simple forms of combinational services using multiple radio access bearers in WCDMA and DTM in GSM.

The third step is to introduce basic IMS reachability mechanisms and application support, thereby enabling a rich set of compelling multimedia applications.

In the long term, as standardization and technology evolves, improved packet radio bearers will become viable alternatives to circuit-switched bearers, eventually facilitating the move to a pure IP environment.

Wrapping it up

If we follow these steps will we have come any closer to solving the three problems associated with IP in mobile communications? Let's see:

- Problem 1, *The wrong flow pattern*: By allowing real-time flows to use circuit bearers we ensure that the right tools get used for the right jobs.
- Problem 2, *The hidden bearer*: By introducing mechanisms for distributing capability information, we enable applications to act on bearer status (as well as on any other relevant piece of information).
- Problem 3: *The narrow, leaky pipe*: By separating the flows (real-time critical on the circuit side, and best-effort on the packet side), we enable recovery mechanisms (of, say, TCP) to handle packet loss and variable delays.

Figure 4
The power of "show-and-tell": talking is useful, but human communication is so much more effective when we are also able to share what we see.

Conclusion

This article tells the story of a carefully considered evolution toward all-IP. Ericsson firmly believes that the step-by-step approach described in this article constitutes the best way of creating sustainable and profitable business. It represents a responsible approach to growth and risk management, focusing on what can be done in the near term to drive the development of revenue-generating services. Furthermore, it introduces key elements, such as the IMS subsystem, that are cornerstones of the full all-IP solution. The remaining parts of this target architecture will be the subject of a later article.

First and foremost, the combinational approach and the long-term solution are founded on end-user wants and needs, as well as on careful evolution of wireless operators' current and future assets. This way, Ericsson contributes toward building on the strength of key technology, the installed base, and fruitful cooperation with operators.