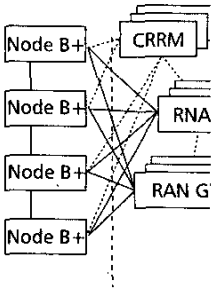
Key Concepts for Evolution Toward Beyond 3G Networks

Sami Uskela, Nokia



Now, when the first commercial 3G services based on 3GPP specifications have been launched around the world, it is the right time to start considering the evolution potential of the 3G systems.

Abstract

Now, when the first commercial 3G services based on 3GPP specifications have been launched around the world, is the right time to start considering the evolution potential of 3G systems. It is assumed that the majority of the traffic in future mobile networks will be generated by content consumption related services, which are realized with IP technologies. Thus, it is necessary to optimize the cellular networks for carrying IP traffic as efficiently as possible. In this article, we describe an evolution scenario for the 3G network architecture specified by 3GPP. The IP delivery part of the network architecture is first optimized within each subsystem, while maintaining interoperability with the legacy network. Later, the network is streamlined as a whole to provide the most efficient solution. We show how graceful evolution of the 3GPP system can benefit from possibilities of the new technologies, especially IP-based transport, while maintaining compatibility with existing user equipment and capitalizing on existing infrastructure investments.

Introduction

While the initial deployment of third-generation (3G) networks is ongoing and the first commercial 3G services have been launched, the development of mobile communications has not ceased. On the contrary, an increasing number of new initiatives and technologies are being introduced to complete or compete with 3G networks.

In this article, we discuss the evolution potential of the 3G networks defined by the Third Generation Partnership Project (3GPP). The 3GPP system is a combination of the new wideband code-division multiple access (WCDMA) air interface and related radio access network (RAN) architecture, and evolved Global System for Mobile Communications (GSM) and General Packet Radio Service (GPRS) core networks. The main improvement considered while specifying 3G networks was providing excellent support for simultaneous circuit- and packet-switched communications. However, the nature of the services anticipated to be offered over packet-switched bearers has changed from best effort services to IP multimedia. Thus, the 3G networks must evolve to meet new challenges.

The usage of mobile communication networks currently comprises mainly voice and messaging based person-to-person communications. It is anticipated that in the near future richer person-to-person and group communication models emerge. Furthermore, it is anticipated that media consumption via mobile networks will become a significant contributor to the traffic of the networks. The new usage patterns of mobile communications lead to an always-on society, where most, if not all, subscribers are continuously online; they expect to be able to initiate communication with their peers and access their favorite media at any time, without any delay.

The service creation for the always-on society will build on IP networking layer, which provides universal platform for development of multimedia services. Thus, it becomes pivotal for 3G networks to support efficient IP communications; even tough the circuit switched teleservices will be used also in the future, IP traffic will dominate.

When enhancing support for IP communications in a cellular environment, we should focus on increasing perceived end-user experience, which can be achieved mainly by:

- Higher bit rates
- Lower communication latency
- Ubiquitous service availability

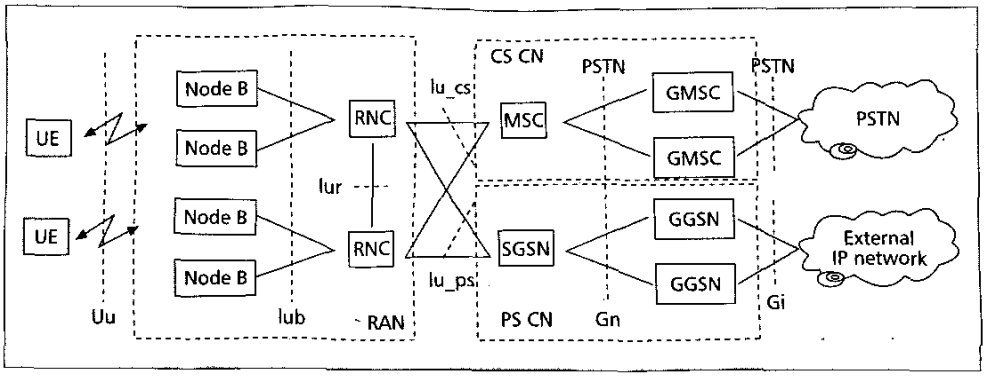
The development of 3G networks should aim to improve these aspects. However, the cost of service must also be reasonable. Therefore, we must also seek solutions that ensure the lowest cost for delivering required services.

The first target can be approached via two routes: enhancing the capabilities of the existing cellular air interfaces and utilizing complementary access technologies as part of the 3G networks [1]. Both approaches are currently studied actively in the 3G community. However, we discuss here only how to ensure that the network architecture is not the bottleneck when introducing more capable air interfaces.

The second target can be achieved by enhancing the air interfaces and rethinking the network architecture. The latter is the main topic of this article. In addition, the third target must be approached while designing the network architecture and especially in conjunction with accommodation of the complementary access technologies.

1536-1284/03/$17.00

The legacy of GSM networks and the desire for truly open interfaces between RAN and CN were the main drivers behind many design choices made during the development of the 3GPP network architecture.



■ **Figure 1.** *Simplified 3GPP Release 99 logical architecture (all interfaces and network elements not shown).*

The cost of service has several components, such as marketing cost, terminal cost, and license fees, on which the network architecture design does not have any significant impact. However, the network architecture definitely has an impact on the capital expenses related to network roll-out and the operational expenses related to running the network. In general, both capital and operational expenses tend to be proportional to the complexity of the network; thus, simpler network architecture may prove more economical.

The rest of the article is organized as follows. In the next section, we give the background required to understand the key concepts of the 3GPP network architecture. Then we describe the current 3GPP network architecture. Next we introduce an evolution scenario for the 3G network topology. Then we discuss further network architecture streamlining possibilities. Finally, we give our conclusions.

Background

Some historical background on the development of the 3GPP system is necessary to understand the design choices that have led to the current architecture and why the evolution scenarios presented later in this article have become topical only recently. Since the 3GPP network is strongly based on reuse of the GSM network architecture, we must start our examination from 2G networks.

The main objective of GSM network development was to create a mobile telephony network that would provide equivalent services to ISDN: voice calls, circuit switched data calls, fax transmission, and other predefined services. The currently important short messaging service (SMS) was more like a side product of mobile telephony. Time-division multiplex (TDM)-based transmission technology was predominant in fixed networks when GSM was designed. Thus, it was a natural choice to build GSM network architecture on top of TDM-based transmission.

In addition, open and well-defined multivendor interfaces between network subsystems were considered crucial for commercial success of the GSM system. In particular, openness of the interface between the base station subsystem (BSS) and core network (CN), the A-interface, was seen very important.

When packet data services became more important in fixed networks, the GSM community also introduced packet-based bearer services, GPRS, to augment GSM networks. The aim of the GPRS development was to provide efficient access to both IP and X.25 networks (later X.25 support has become obsolete and have been removed from specifications) while keeping the system compatible with existing terminals and minimizing changes needed in existing GSM infrastructure. Thus, a separate CN for GPRS was introduced, and its interaction with the circuit-switched CN was minimized. Since the GPRS was intended to carry packet-based services, IP-based transmission technology within the GPRS CN was selected.

When the development of 3G systems started, there were clear objectives to improve from GSM by increasing the bit rate over the air interface, providing different quality of service (QoS) classes for packet data, and enabling simultaneous usage of circuit- and packet-switched services (which has proven to be challenging for GPRS networks and terminals). On the other hand, backward compatibility with the existing GSM services and reuse of the GSM infrastructure were considered very important. Therefore, the GSM CN was adopted as the basis for the 3GPP CN, to which the new RAN was connected. Even though IP-based transmission was already adopted for the GPRS CN and the packet core of the 3G networks, it was not considered sufficient to provide real-time capabilities needed in RAN and circuit-switched services. Thus, an asynchronous transfer mode (ATM)-based transmission solution was adopted for the rest of the 3GPP network.

The openness of the interface between RAN and CN was also considered very important when 3G networks were designed; hence, an open and well-defined Iu interface was introduced. The Iu interface has two variants: Iu_cs between RAN and circuit-switched CN, and Iu_ps between RAN and packet-switched CN.

The legacy of GSM networks and the desire for truly open interfaces between RAN and CN were the main drivers behind many design choices made during the development of the 3GPP network architecture that are further elaborated in the next section.

3GPP Network Architecture

Figure 1 depicts simplified 3GPP Release 99 network architecture with only the network elements and interfaces relevant for our discussion. There are three main subsystems in the architecture:

- The RAN, which contains all radio-access-specific functionalities and related network elements
- The circuit-switched CN (CS CN) that manages circuit-switched sessions and interconnects the cellular network to the public switched telephone network (PSTN)
- The packet-switched CN (PS CN) that manages packet-switched sessions and interconnects the cellular network to external IP networks such as the Internet

Naturally, the system also contains user equipment (UE). In addition to the entities shown in the figure, the architecture also contains registers and databases, the most important being the home location register (HLR), which contains subscriber profiles and cryptographic keys required for user authentication.

RAN functionality is divided into two network elements: Node B and the radio network controller (RNC), which are connected over the Iub interface. The former is merely a radio modem, which mostly takes care of the WCDMA layer 1 processing. The latter handles most of the RAN functionality: the RNC manages radio resources, terminates air interface layer 2 and 3 protocols, performs macro diversity combining (MDC), schedules downlink radio frames, gives power control commands, and so on.

There is a strict one-to-many mapping between Node Bs and RNCs; each Node B can be connected to only one RNC, while one RNC can manage hundreds of Node Bs. Furthermore, there is vertical connection between RNCs, which is needed when a UE is in soft handover with Node Bs controlled by different RNCs or a user moves from the coverage area of one RNC to that of another RNC.

This functional split imposes specific requirements for the transmission technology used to connect RNCs and Node Bs. Since the MDC and radio frame scheduling are performed in the RNC, the latency and jitter the radio frames experience while being transmitted between the nodes must be controlled with an accuracy of microseconds. In addition, the Node Bs need an external reference clock signal to maintain the air interface frequency accuracy; the reference clock is usually distributed over the transmission network physical layer.

The logical architecture of the 3GPP Release 99 CS CN is the same as in GSM: MSCs take care of routing and managing circuit switched sessions. There are also gateway MSCs (GMSC) used to route mobile terminating calls from the PSTN to the MSC serving the called subscriber. Both user and control planes are handled in the MSCs. The CS CN is connected to a RAN via the Iu_cs interface, which connects the MSC to RNCs. There is strict one-to-many mapping in the Iu_cs interface: each RNC can be connected to only one MSC, while one MSC can manage hundreds of RNCs. There is no strict mapping between MSCs: any GMSC can connect to any MSC.

PS CN functionality is distributed into two network elements: SGSN and GGSN. The former takes care of most of the session management (including QoS), mobility management, and AAA functionality of the PS CN. The latter is merely a gateway between external IP networks and a cellular system, which, however, have important functionality such as QoS mapping between the networks, mobility anchoring, packet filtering, and so on. The PS CN is connected to a RAN via the Iu_ps interface, which connects the SGSN to RNCs. There is strict one-to-many mapping in the Iu_ps interface: each RNC can be connected to only one SGSN, while one SGSN can manage hundreds of RNCs, but any SGSN can connect to any GGSN and vice versa.

The logical architecture assumes that both user and control planes are handled in the same network elements throughout the network; that is, all network elements have integrated user and control plane functionality. However, the specifications have clear logical separation of the user and control plane functionality; the protocols conveying control plane messages and user plane traffic are completely independent in all interfaces.

Some characteristics of the 3GPP Release 99 logical architecture are not optimal for delivering IP-based services. There is a strict tree hierarchy in the network; handling of user and control planes is tightly coupled; and there are strict delay and timing requirements for transmission links within the RAN.

The strict hierarchy combined with coupling of the control and user planes has major implications for the logical architecture:

- The user plane traffic must be processed in every hierarchy level, which naturally is a source of additional delay and jitter.
- Since the user traffic has to always go via the highest level in the hierarchy, the routing of traffic might be nonoptimal, especially when two terminals within the coverage of the same Node B are communicating with each other.
- When the user and control planes are always handled in the same network elements, scaling the capacity of the user and control planes independently is difficult.
- Failure of an upper level node usually impacts all the nodes below the failed node in the logical tree; for example, if an RNC crashes, all Node Bs connected to that RNC are disconnected from the network.

However, there are also some reasons why hierarchy is needed in cellular networks:

- It would be challenging to realize real-time requirements of mobility management operations (especially handover) without a hierarchical solution.
- Aggregation of network topology information is required in order to have scalable networks.
- Common sense says that local events should not have global effects; for example, handover between adjacent BTSs should not require global signaling.

Therefore, we should find solutions to avoid the cons of the hierarchy while keeping the benefits of it. This could be achieved by transforming one-to-many bindings to many-to-many bindings between the hierarchy levels and decoupling user plane handling from control plane

Some characteristics of the 3GPP Release 99 logical architecture are not optimal for delivering IP based services: There is a strict tree hierarchy in network; handling of user and control plane is tightly coupled; and there are strict delay and timing requirements for transmission links within RAN.

Due to inherent any-to-any connectivity between hosts, IP networks enable more distributed functional architectures and more versatile network topologies than what is practical in TDM or ATM-based networks.

hierarchy. Next, we will discuss evolutionary solutions that aim to enhance the logical architecture of each subsystem separately.

Evolution of the Subsystems

When IP-based transport networks have become commonplace in fixed networks and have proven to provide robust solutions, it has become clear that benefits of the IP based networking technologies should also be utilized in cellular networks. Due to inherent any-to-any connectivity between hosts, IP networks enable more distributed functional architectures and more versatile network topologies than what is practical in TDM or ATM-based networks. In the following, we discuss how the strengths of IP-based transport can be leveraged in 3G networks.

RAN Evolution

A new RAN architecture has been proposed [2] to address the issues discussed above; RNC functionality of the conventional RAN architecture is distributed into smaller pieces. First, radio-interface-related processing (e.g., MDC, power control, and radio frame scheduling) is relocated into base stations. To distinguish these advanced base stations from Node Bs of the R'99 architecture, they are called Node B+ in this article. Second, remaining non-radio-interface-specific user plane handling is located in a new network element called a RAN gateway (RAN GW), with the main responsibility to provide standard Iu_cs and Iu_ps interfaces toward the CS CN and PS CN. Third, the networkwide radio resource management related functionality is isolated into a new network element called a common radio resource manager (CRRM). The remaining RNC functionality is handled by the RAN access server (RNAS).

While the main reason for relocating the radio-specific processing into the Node B+ is improved radio performance due to decreased delay in the terrestrial network, it also removes the strict delay and jitter requirements set to the transmission links within the RAN. However, the Node B+ also has to get an external reference clock in order to maintain the frequency accuracy of the air interface, which may impose some requirements on the transmission technology if the clock reference is distributed over the terrestrial transmission network to the Node B+s.

When the radio-specific processing of the user plane is performed in the Node B+, the functionality of the RAN GW is reduced to a user plane routing point, which hides the internal structure of the RAN architecture from CS CN and PS CN. The same applies to the RNAS; when the radio resource management and other radio-specific functionalities are handled by other network elements, its main task is to hide the internal structure of the RAN architecture from CS CN and PS CN. Therefore, the RNAS and RAN GW are mainly needed to provide compatibility with existing infrastructure and are not needed if the CN is up to date.

The isolation of networkwide radio resource management functionality into the CRRM clarifies the roles of the network elements and improves the overall performance of multiradio networks. The radio resource management common to all radio technologies allows optimized utilization of the radio spectrum and capabilities.

There is no strict tree hierarchy in the proposed architecture; any Node B+ can connect to any RNAS and RAN GW and vice versa. Moreover, it has been proposed to increase the flexibility of the Iu_cs and Iu_ps interfaces [3] by allowing each RNC (or RNAS and MGW) to be connected to several MSCs and SGSNs. This breaks the rest of the strict tree hierarchy of the logical architecture. In this architecture, horizontal connections between Node B+s are needed when UE is engaged in handover between two (or more) Node B+s.

CS CN Evolution

The coupling of the user and control plane handling within the CS CN have already been addressed in 3GPP Release 4 by introducing the MSC server concept [4]. The MSC (and GMSC) functionality is divided into two network elements: MSC server and MGW. The former takes care of all the control plane functionality within the CS CN and terminates the control plane protocols of the Iu_cs interface. The latter handles user plane processing and switching, which includes termination of user plane protocols of the Iu_cs interface, speech codec processing, and switching between the circuits. The implementation of the MSC server concept is straightforward from a specification point of view, since the user and control plane protocols are already independent in the Iu_cs interface and the CS CN internal interfaces. With these enhancements, the CS CN logical architecture can be evolved according to the requirements set earlier in this article.

PS CN Evolution

A similar concept for dividing the user and control plane functionality into separate network elements has been proposed for the PS CN [5]; the SGSN would be divided into an SGSN server and a packet-switched media gateway (PGW). The SGSN server takes care of all the control plane functionality within the PS CN and terminates the control plane protocols of the Iu_ps interface. The PGW terminates user plane protocols of the Iu_ps interface and performs forwarding of the user plane datagrams. The GGSN functionality remains the same as in the R'99 architecture; since there is very little, if any, true control plane functionality in the GGSN, there is not much benefit to be gained by splitting it in two. The implementation of the SGSN server concept is straightforward from a specification point of view, since the user and control plane protocols are already independent in the Iu_ps interface and the PS CN internal interfaces.

It has been noted that in fact the PGW is a more or less redundant network element [1]. Since the user plane protocol stack between RNC and PGW, and PGW and GGSN are exactly the same, the PGW only forwards datagrams. Thus, the user plane from the RNC could connect directly to the GGSN and vice versa. However, we must remember that the PGW is assumed to perform functions like traffic volume

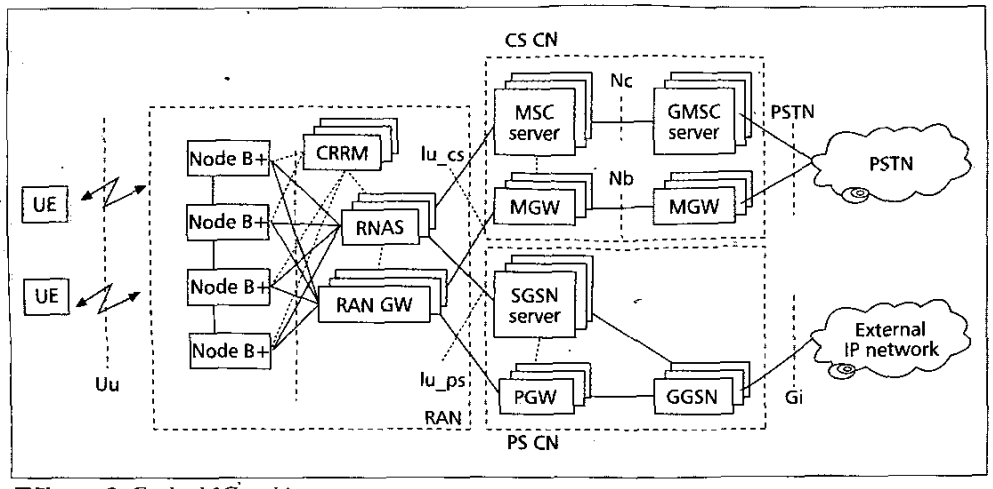


Figure 2. *Evolved 3G architecture.*

accounting and legal interception, which has to be organized in alternative ways, if the PGW is removed from the architecture.

A Summary of Subsystem Evolution

When all the enhancements to the subsystems discussed above are applied, the resulting architecture is depicted in Fig. 2. Some of the requirements set earlier in this article are well addressed:

- User plane handling is almost completely decoupled from control plane handling.
- The strict tree hierarchy is replaced by many-to-many bindings between the hierarchy layers.
- The requirements for transmission links within RAN are relaxed.

However, there are still several network elements processing the user plane.

It is important to note that all the discussed architectural enhancements can be realized without any modification to the air interface protocols. Thus, this evolved 3G architecture inherently provides support for already deployed 3G terminals that are implemented according to R'99 specifications. Furthermore, the evolution steps discussed above can be deployed independently and only in selected parts of the network; for example, the whole RAN architecture does not have to change overnight, but the new architecture can be deployed gracefully when new equipment is installed.

Streamlining the Architecture

When examining the evolved architecture depicted in Fig. 2, we note that the evolution of each subsystem independently leads into a system with multiple gateways, which actually have very little functionality; their main task is to provide a standard interface toward the other subsystems. While maintaining the standard intersubsystem interfaces enables different evolution speeds for each subsystem, streamlining the whole network architecture should also be considered. Next, we will briefly describe two potential scenarios for the next steps in the evolution of the logical architecture: first with the assumption that the conventional circuit-switched paradigm must also be supported with the next phase of the architecture, and second with the assumption that IP multimedia is a huge success and eventually phases out the CS CN.

In the first option, further streamlining the user plane of the packet data network should be considered. The redundancy of the PGW was already addressed above, but also the user plane protocol between the RAN GW and the Node B+ could be exactly the same as between the RAN GW and PWG (or GGSN). Thus, it would be attractive to pursue the possibility of streamlining the user plane handling in the extreme; that is, introducing a direct user plane connection from Node B+ to GGSN and vice versa.

Streamlining the CS CN any further may prove to be challenging without serious interoperability problems with existing phones and telecom infrastructure. However, it might be worthwhile to pursue the possibilities of also eliminating the RAN GW from the CS user plane; that is, connecting the CS user plane directly from a Node B+ to an MGW.

The streamlined architecture resulting from the evolution steps introduced above is depicted in Fig. 3. It is notable that this architecture provides inherent support for all 3G terminals, including the 3GPP R'99 terminals already deployed.

In the second scenario, there is no need to support circuit-switched bearers, when the functionality of RNAS is very limited; hence, it would make sense to also streamline the control plane architecture. An attractive option to consider would be combining the functionalities of the RNAS and SGSN server into one network element, which we call herethe control server.

If the further evolution step introduced above is implemented, we get an extremely streamlined packet data architecture, illustrated in Fig. 4. It is notable that even though this architecture does not resemble the original 3G architecture too much, with careful design it is possible to realize the extremely streamlined architecture without significant changes to the air interface

It is important to note that all the discussed architectural enhancements can be realized without any modification to the air interface protocols. Thus, this evolved 3G architecture inherently provides support for already deployed 3G terminals that are implemented according to R'99 specifications.

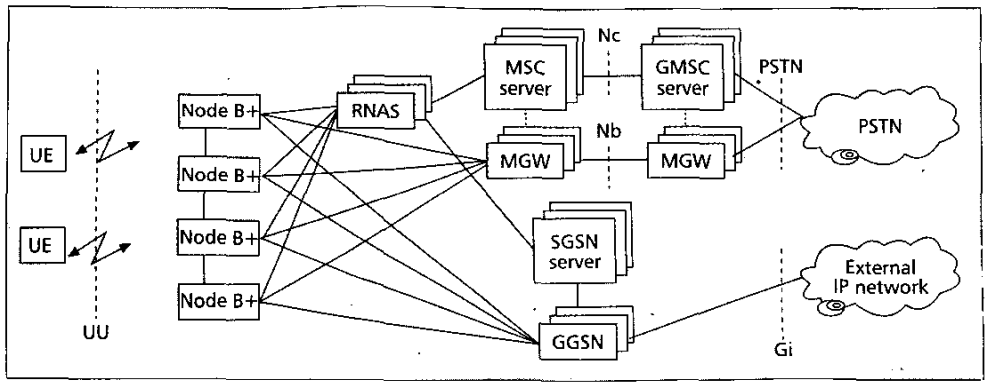


Figure 3. *Streamlined 3G architecture.*

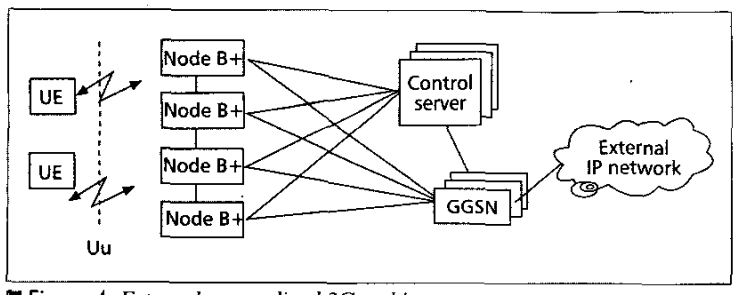


Figure 4. *Extremely streamlined 3G architecture.*

protocols, ensuring backward compatibility with older-generation packet-based UE.

Conclusions

The designers of the 3G networks did not start from a clean table; on the contrary, the design choices of 3G networks were limited by requirements for open multivendor interfaces, backward compatibility and reuse of existing 2G infrastructure, and availability of technologies. The notable disadvantages of the current 3G network architecture are strict hierarchy and integrated user and control plane handling.

Now, when IP-based transport has become a viable option for the terrestrial part of 3G networks, there are new possibilities for network architecture design. However, the requirements for open multivendor interfaces, backward compatibility, and reuse of existing infrastructure are still valid. Thus, it makes sense to optimize architecture within each subsystem while maintaining the standard interfaces between the subsystems. That would allow coexistence of the current and new architecture within the same network, which is essential for graceful network evolution.

IP-based terrestrial transport enables more distributed functionality and more flexible data routing than current ATM-based transport. The distributed functionality can be utilized to define a new functional split in the RAN, where most of the radio related functionality could be located in a BTS, which enhances radio performance and relaxes transport requirements in the access network. The flexible data routing can be utilized to remove strict hierarchy by enabling many-to-many bindings between network elements and to decouple user and control plane handling within each subsystem.

However, maintaining existing standard interfaces between the subsystems limits the possibilities for streamlining the architecture. Thus, the network architecture resulting from optimizing each subsystem separately can be further streamlined when the network is studied as a whole.

The graceful evolution of 3G networks can lead to streamlined and competitive network architecture that takes full advantage of the capabilities of IP transport. The evolution discussed in this article allows coexistence of new and old architectures, and facilitates reuse of infrastructure investments. Furthermore, evolutionary development of 3G networks enables inherent backward compatibility with existing 3G terminals.

Acknowledgments

The author would like to express sincere gratitude to several colleagues within Nokia who have contributed to the concepts described in this article.

References

[1] H. Honkasalo *et al.*, "WCDMA and WLAN for 3G and Beyond," *IEEE Wireless Commun.*, Apr. 2002.
[2] J. Kempf, "Open RAN Architecture in 3rd Generation Mobile Systems," MWIF MTR-007, Rel. 1.0.0, Sept. 2001.
[3] 3GPP TS 23.236, "Intradomain Connection of Radio Access Network (RAN) Nodes to Multiple Core Network (CN) Nodes," v. 5.2.0, Mar. 2002.
[4] 3GPP TS 23.205, "Bearer-Independent Circuit-Switched Core Network, Stage 2," v. 4.4.0, Mar. 2002.
[5] 3GPP TR 23.873, "Feasibility Study for Transport and Control Separation in the PS CN Domain," v. 4.0.0, Mar. 2001.

Biography

SAMI USKELA (sami.uskela@nokia.com) received his M.Sc. degree in electrical engineering from Helsinki University of Technology in 1999. He has been with Nokia since 1997. He has been involved in research on network architectures and service platforms ever since. He has been actively involved in standardization of 3G systems: during 1999 he contributed to network architecture and service platform standardization in TTC (Japan) and during 1999–2000 to ALL IP standardization in 3GPP. In May 2001 He was nominated senior specialist within Nokia Networks responsible for beyond 3G network research. He has published about five articles in international conferences and journals, and holds over 20 patents and patent applications covering different areas of network architectures and service platforms.