4

Wireless-Network Architecture

4.1 2G Wireless-Network Architecture

A simplified diagram of today's 2G wireless-network architecture is shown in Figure 4.1. Shown here are three RBSs, the BSC, and the MSC and connection with the PSTN. In addition to that, there are a number of other nodes not shown here, like the voice mail system (VMS), SMS, HLR/VLR, and so on. Regardless of whether the access transmission network (connection between BSC and RBS or backhaul) is leased or owned by the operator, RBS can be connected to the BSC via microwave links, fiber-optic, or wireline (usually copper) systems. The BSC provides the connectivity between the MSC/PSTN and the radio network. It performs the radio call management functions and radio network management functions and its capacity is usually given in Erlangs. BSC blocking probability is defined to be the probability that a new request for service is rejected at the BSC due to lack of resources. Resources are understood to be card processing, transmission link capacity, countable resources such as channel cards, and so on. BSC blocking probability does not include blocking due to limitations of the air interface or call admission control, and the BSC blocking probability is usually specified to be 0.5%. BSC and MSC are usually colocated and, therefore, shown as one node.

It is important to note that the physical layer of the transmission network, copper, microwave, or fiber optic will not change from 2G to 3G wireless networks. The only change will be more stringent requirements on its

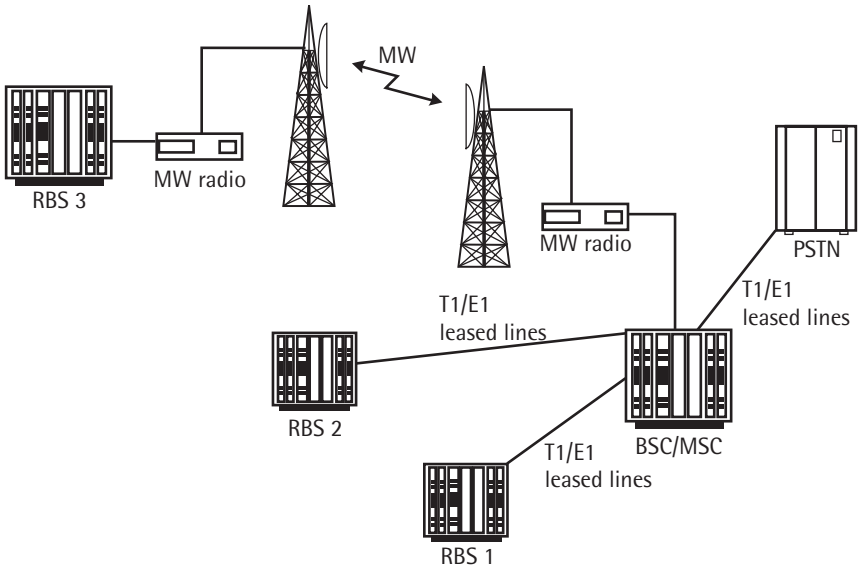

Figure 4.1 Example of 2G wireless network.

quality, availability, and reliability as well as higher capacities required to carry the traffic.

There are today several vertically oriented, single-service networks capable of delivering similar services. We have wireline and wireless networks as well as pure data/IP networks as well as cable TV/CATV networks. These are basically separate networks that build on different principles and practices to ensure the reliability of a single service with different approaches to network management, guaranteed service levels, and so on. For over 100 years, classic telephony networks have been optimized to carry real-time voice traffic between fixed points in the network. The classic telephony network supports an integrated service concept, involving the following:

- One service type—voice;
- One subscription;
- One user terminal.

This leads to a vertical industry orientation, where the operator offers everything from subscriber access to service creation and service delivery

across a wholly owned network infrastructure optimized for a particular service category. Each vertically integrated network incorporates its own protocols, nodes, and end-user equipment and terminals. This means that the telephony (voice) and data service domains are still more or less kept separate. The rapid convergence between telecommunications and datacommunications will lead to a convergence of these purpose-built networks into ATM/IP-based multiservice, or next-generation, networks that can provide reliable and real-time communications. This network convergence raises some fundamental issues of network characteristics and how to bridge the inherent value of reliable circuit-switching technology with more best-effort-oriented packet-switching technology.

4.2 3G Wireless-Network Architecture

4.2.1 Directions in 3G Developments

Multiservice switching is the foundation of next-generation networks. The need for wireline- and wireless-network infrastructure to support future data-communication and IP services is obvious. However, not only is the type of traffic changing, the quantity of traffic is also growing rapidly. Both of these new factors are of critical importance in the development of future networks. Another significant fact is that the major source of revenue for most operators around the world is still voice services. And it is likely to remain so for the coming years—while the main source of growth will be data-based services. This means that networks must be optimized to carry packet-oriented traffic at the same time that they are delivering a reliable and high-quality voice service; that is, the technologies are converging.

So far the networks have been shaped by the concept of one network, one service, one subscription, and one user terminal. This changed with the introduction of the Internet. The future networks will be multiservice networks, able to carry a full range of services, from voice communications and simple file transfers to high-speed Internet and real-time, broadband multimedia services. These services will be accessible via different access networks and a number of different terminals.

Using packet technology as the infrastructure for telephony services implies that packet networks must meet stringent requirements; QoS is a key issue. IP is today not a mature bearer technology for efficient large-scale telephony QoS solutions as regards delays and latency in voice streams. The performance available from today's IP environment cannot provide carrier-class telephony services, but that may change very soon.

The only technology available today that can meet those requirements is ATM. ATM seems to be the only appropriate technology for use in core multiservice networks, to carry both IP/data services and voice. ATM is able to act as a connectivity layer for both traditional voice and IP/data services. ATM was developed and standardized for both telephony and data, and the large incumbent telecom operators were involved in the standardization process. Therefore, most operators see ATM as the only safe migration path to new generation networks. In other words ATM has now found its role as a network infrastructure technology capable of carrying all the different transport needs of the future networks, including IP. The fundamental structure of the next generation multiservice network is based on a shared bearer network providing multiple services based on ATM. Figure 4.2 shows the layered backbone network model.

The bearer network is accessed through so-called media gateways (MGWs) or hybrid switches. The MGWs control the traffic in and out of the bearer network and handle interconnections to other networks, different access networks or to traditional switches and routers.

In this new network infrastructure, telephone calls will be converted to ATM switched virtual circuits while separate telephony servers control the setup of the calls through the network. The telephony servers thus provide the equivalent intelligence of today's telephony exchanges, but are not involved in the actual connections as such; the calls are set up end to end

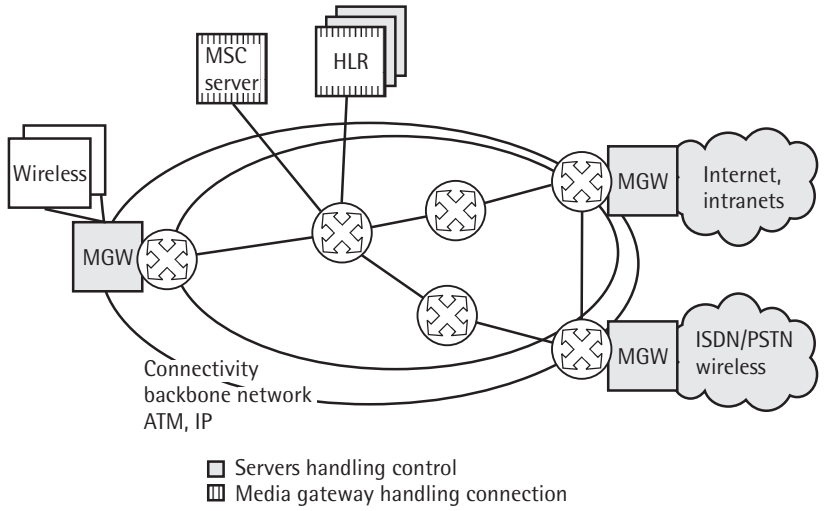

Figure 4.2 Layered backbone network model.

across the ATM network using a switched virtual circuit between the edge devices. This separation of the switching and connectivity functions in the network is a key to the evolution path toward a single bearer network based on ATM. Telephony subscribers may be connected to the ATM bearer network via the MGW from traditional switches, standard access nodes, or if it is an enterprise customer, PBX. The telephony server does not only control phone calls, it also contains all the required telephony intelligence for all MGWs in the domain.

The data access to the bearer network may range from LANs and ATM switched leased lines, to cellular networks and data transmitted from or through these. All access, voice as well as data, terminates in an MGW, which ensures that traffic gets onto the ATM bearer network. A single telephony server with several stand-alone MGWs (multiservice switches) would satisfy regulatory requirements on geographically spread points of presence. The operators would not have to deploy a traditional countrywide circuit-switched network, but would be able to simultaneously offer the full range of telephony and data services from day one. As network management is very expensive and complicated, another main benefit is the fact that only one network has to be managed instead of several different networks.

Today there are major developments in virtually all areas of network access and infrastructure. In access, new digital subscriber line (DSL) and wireless technologies are opening up bandwidth and giving users more and more network capacity, instant access, and on-line services. In the core network (CN), new switching and transport technologies such as ATM and WDM are expanding the capacity and the flexibility of the networks. While the focus for initial wireless 3G deployment is voice services, it is expected that wireless 3G will also become a key Internet access technology. Although expected 3G data rates of up to 2 Mbps will never compete with wireline services, the trigger that will fuel exceptional 3G growth is the promise of true mobility, Internet-connected data, and information services anywhere, any time, from a variety of handheld platforms. Global standards for 3G systems are still evolving, but high-level architectures and overall network characteristics are beginning to solidify, accelerating the need for new testing strategies and capabilities for a key component of 3G, the terrestrial radio access network (RAN). RAN infrastructure will support 3G functions including access, roaming, transparent connection to the PSTN and the Internet, and QoS management for data and Web connections. The concept of the 3G wireless network is shown in Figure 4.3.

The high-QoS characteristics of classic telephony networks must now migrate onto horizontally oriented next-generation networks that can support

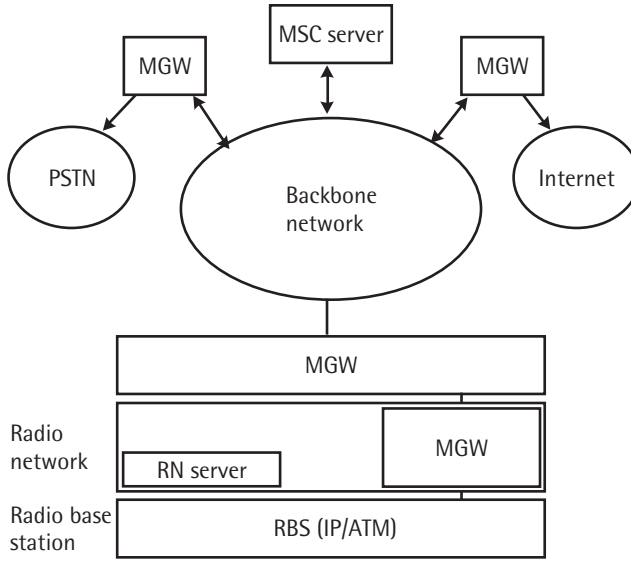

Figure 4.3 Concept of 3G wireless network.

multiple services based on ATM and the IP protocol. Similar developments have been under way in the enterprise market and took off around 1995 and 1996 with the widespread adoption of the TCP/IP protocol for intranet technology. For public network operators and service providers, the heavy increase in data traffic is leading to a bottleneck in narrowband networks—wireline or wireless or both. Operators have to offer circuit-switched and packet-switched services, and must expand even further to offer multiservice and multimedia networks (Figure 4.4). It is also essential for the operator to optimize the network resources—for example, to use one transmission network that will be suitable for all services. Operators need to ensure that their investments will address their transmission requirements well into the future. The future lies with packet-based transmission technologies.

Almost all networks of the past and today are vertically integrated. Vertically integrated networks are single-service networks where the operator offers anything from subscriber access to service creation and delivery. Services are offered across a wholly owned network infrastructure, optimized for a particular service category. Each vertically integrated network incorporates its own protocols, nodes, and end-user equipment and terminals. They operate different principles to ensure the reliability of a single service. Vertically

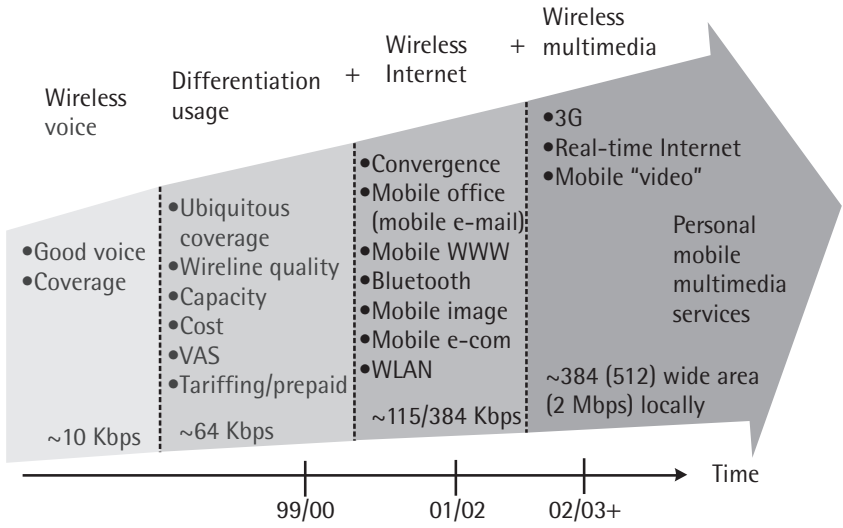

Figure 4.4 Wireless evolution.

integrated networks have different approaches to network management, guaranteed service levels, etc. Vertical integration means that telephony and data service domains are kept more or less separate.

The rapid convergence between telecommunications and data-communications will lead to a convergence of these purpose-built networks into ATM/IP-based multiservice, or next-generation, networks. This network convergence raises some fundamental issues with respect to network characteristics and how to bridge the gap between the inherent value of reliable circuit-switching technology and best-effort packet-switching technology. In order to survive in a converged communications marketplace, operators need to draw on truly open systems that invite competition in horizontal layers. The QoS characteristics of classic telephony networks must now migrate to horizontally oriented next-generation networks that can support multiple services based on ATM and IP, since the 3G network will ultimately be an open IP platform supporting a wide range of new global services [1]. IP telephony (IPT) has quickly emerged as a serious alternative to traditional circuit-switched telephony. Not just because it offers a more cost-effective service, but because the underlying technology offers a wealth of new business opportunities. The evolution of IPT will not only depend on its integration into successful business operations, but

also on the evolution of the underlying technology. One of the most limiting barriers for this type of service offering is, without a doubt, lack of interoperability. If different vendor offerings and different carrier networks cannot interoperate, this will limit the possibility for end-user connectivity. A number of all-IP wireless-network architectures have been proposed and planned for 2004 and beyond and it is too early to say exactly how the evolution of the wireless network will go from there. In 3G networks and beyond, bandwidth flexibility is a key issue and involves a flexible decentralized provision of bandwidth to a single user as the need varies, but also cost-effective bandwidth provision to a large number of users with different bandwidth requirements in the same network.

4.2.2 Horizontally Layered Network Architecture

In the horizontally layered network architecture (Figure 4.5), functionality and nodes are arranged in layers according to their specific areas of use. The layered concept of the network architecture of the 3GPP specifications comprises three distinct layers:

1. Application layer;
2. Network control layer;
3. Connectivity layer.

4.2.2.1 The Application Layer

The application layer is where the end-user applications reside. In modern networks, applications are implemented in mobile terminals and in dedicated application servers in the network. The application servers are often complemented with content servers, which host service-related databases or libraries, such as video-clip libraries or news history databases. Concepts such as the virtual home environment (VHE) and open service architecture (OSA) were developed in the 3GPP to allow operators to provide unique services. Operators benefit from being able to differentiate themselves from one another by providing unique services, thus securing for themselves a higher position in the value chain. They also have the option of developing these services themselves or of obtaining them from third-party software houses and they can even get external service providers to run them. This flexibility allows the operator to choose from a huge portfolio of services that it can offer its subscribers. The application layer interfaces with the

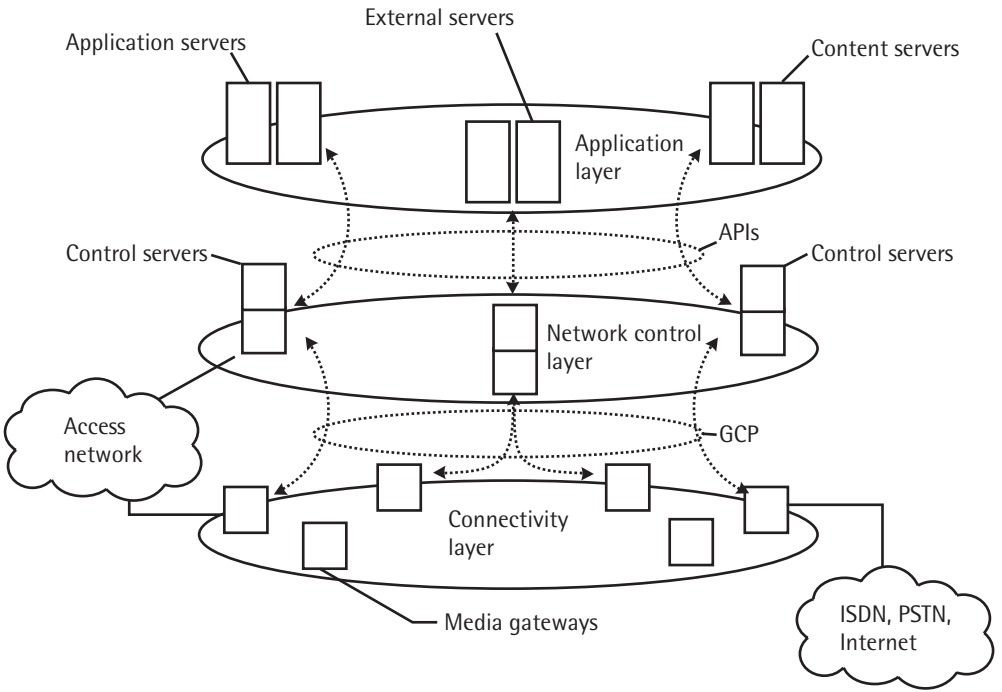

Figure 4.5 Horizontally layered network architecture.

network control layer via a defined set of open application program interfaces (API). By using open APIs, application developers can make use of the features of standardized service capabilities to design new services and applications.

4.2.2.2 The Network Control Layer

The network control layer incorporates all the functionality needed to provide seamless, high-quality services across different types of networks. The different networks can be seen as a set of domains, each of which houses control servers that are specific to a given network. Generally speaking, the network control layer houses several different kinds of network servers. The servers are responsible for controlling mobility management, the setup and release of calls and sessions requested by end users, circuit-mode supplementary services, security, and similar functions. These domains can be owned by various individual operators or by a single operator.

4.2.2.3 The Connectivity Layer

The connectivity layer is a pure transport mechanism that is capable of transporting any type of information via voice, data, and multimedia streams (Figure 4.6). Its backbone architecture incorporates core and edge equipment. The core equipment transports aggregated traffic streams between the different nodes at the edges of the backbone. As a rule, core equipment is a backbone router or backbone switch that handles traffic streams either according to very simple classification principles or to routes that the network operator has predefined by means of traffic engineering.

Edge equipment collects customer-specific data and statistics for accounting and billing purposes and provides the single bit-pipes that guarantee an appropriate QoS. The edge equipment is usually an MGW, which operates under the full control of the nodes in the network control layer. In addition, an MGW allows the bit streams to be processed, thus providing coding and decoding of speech streams, canceling echo, bridging multiple party calls, and converting transport protocols. The nodes in the network control layer also control these manipulations. This exertion of control down to the bit-stream level allows the variety of services and applications implemented by the different network control domains to be achieved via a common connectivity layer. At the same time, the services and application are independent of the transport technology applied, which may be mixed or

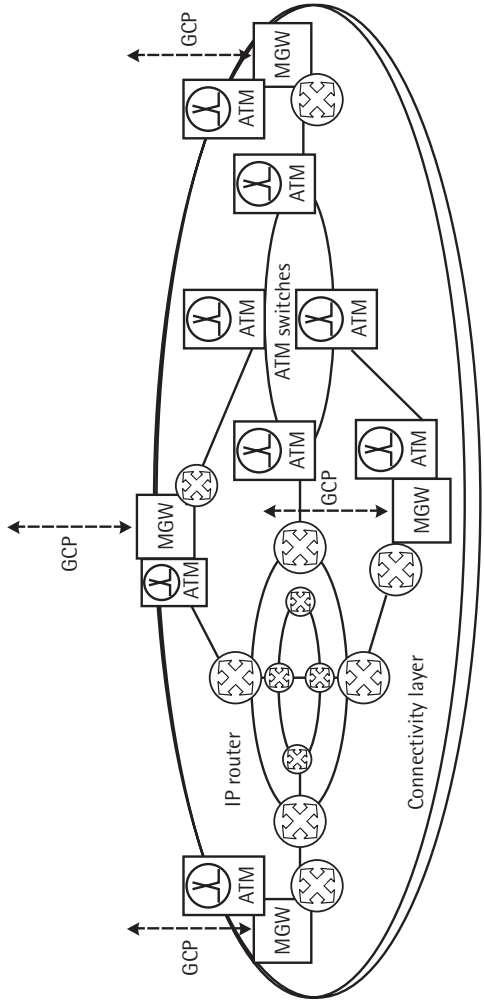

Figure 4.6 The connectivity layer.

vary over time as the network evolves. Connectivity-layer solutions can be based on either ATM transmission or IP transmission. It may also use a QoS-enabled IP-backbone network running IP over SDH, carrying packet- and circuit-mode communications.

4.2.3 3G Core Network

The 3G CN supports circuit- and packet-switched services. It contains the hardware and the software needed to provide end users with multimedia applications. The CN spans both the control and connectivity logical layers. One of the new nodes required in the 3G wireless network is the MGW. The MGW performs functions such as speech coding and decoding, echo cancellation, conference-call bridging, tone and announcement generation, setup and release of user data bearers, and QoS IP routing and switching, including QoS handling and packet retunneling. The MGW will also contain interface functions for different transport standards, for example between an IP- or ATM-based CN and an external STM network.

For volume-based charging support, the MGW keeps track of the volumes sent and received (for packet-based services), and it performs some security functions (for example, for packet mode services). Most MGW resources are shared between packet- and circuit-communication service, or can easily be reconfigured from one communication mode to the other.

The MSC server handles control-layer functions related to circuit-mode communication services at the WCDMA or CDMA2000 RAN and PSTN/ISDN borders, and performs MGW control, ISDN services control, mobility management, authentication, data collection and output, services switching function (SSF), Internet dial-in services (RAS), and element management. In addition to these functions, the MSC server houses the interworking and gateway functionality necessary to act as an SMS-IWMSC and SMS-GMSC for the Short Message Service.

The HLR is a network database for mobile telecommunications in general. The HLR holds all mobile-specific subscriber data and contains a number of functions for managing this data, controlling services and enabling subscribers to access and receive their services when roaming within and outside their home PLMNs. The HLR communicates with GSNs, MSCs, and other network elements via the MAP protocol. The HLR is a real-time mobile telecommunications node of GSM, CDMA2000, and UMTS systems. The HLR is vital for the operation of a network, as it holds all subscriber data, and it is equally important for the call setup in the network, as well as for the control of the roaming subscribers.

4.2.4 Universal Mobile Telephone System

UMTS is the accepted 3G standard for GSM operators. UMTS requires paired 5-MHz RF channels, four times as wide as the paired 1.25-MHz channels required for CDMA2000. For this reason, UMTS is sometimes referred to as WCDMA. By migrating to UMTS, operators will gain access to additional spectrum as well as the greater capacity and expanded functionality of the new technology. UMTS incorporates a more efficient variable vocoder (CODEC). In common with CDMA2000 1x, this vocoder will increase the voice capacity of a given amount of spectrum. Outside of the Americas, UMTS is being deployed on the 1,900-MHz (uplink) and 2,100-MHz (downlink) frequencies. Because of this, some operators, primarily those in the Americas who now use the 1,900-MHz frequencies for PCS, would be unable to migrate to UMTS. Allocation of other frequencies for UMTS may or may not be possible. The well-publicized failures of U.S. operators to acquire frequencies at 700 MHz (occupied by TV broadcasters), 1,700 MHz (occupied by the military), or 2,500–2,600 MHz (occupied by educational broadcasters) provide examples.

4.2.5 3G in GSM Networks

New 3G GSM networks will require new radio and CN elements as well as a new air interface. This will require new BSSs, which will include radio network controller (RNC) and Node B. The RNC will include support for connection to legacy systems and provide efficient packet connection with the CN packet devices (SSGN or equivalent). The RNC performs radio network control functions that include call establishment and release, handoff, radio resource management, power control, diversity combining, and soft handoff (handoff). A Node B is equivalent to a base station in the 2G network, but also incorporates support for the 3G air interfaces.

New cell-planning methods will be needed to support the new frequency allocations for 3G and the radio interface changes. More 3G base stations will be needed than are necessary in a comparable 2G coverage area. This gives an advantage to GSM 1800 and 1900 network operators whose cells already cover a smaller coverage area than those for GSM 900 networks. GSM 900 network operators will need to fill in coverage in between existing cell sites.

The 3G CN will be an evolution from GPRS or equivalent 2.5G CN systems. Upgrades to the mobile and transit switching systems to deliver packets will also be needed. A new piece of network infrastructure for 3G is also the MGW, which resides at the boundary between different networks to

process end-user data such as voice coding and decoding, convert protocols, and map quality of service. The connectivity layer also provides access to backbone switches and nonmobile networks such as cable television. In some vendor solutions, MGWs are controlled remotely by the MSC and GSN servers by means of the Gateway Control Protocol (GCP). The ITU is working to ensure that the GCP is an open standard protocol. Existing network operators can then upgrade their MSC and GSNs to implement 3G or alternatively to implement a new stand-alone MGW that is controlled from the server part of an upgraded 2G node.

4.2.6 3G CDMA Network

The ITU manages the 3G umbrella standard known as IMT-2000. This standard endorses five different modes of RF interface and three major types of terrestrial infrastructure known as RAN. The intention is for any of the RF modes to work with any of the RAN types. The two major types of RAN are UMTS WCDMA and IS-2000 (also known as CDMA2000). UMTS originated in Europe and WCDMA in Japan, but these are now almost identical, having been converged into a single specification under the 3GPP. IS-2000 is predominantly North American and is defined by the 3GPP2 organization. More recently, a third RAN concept has been added, providing direct access to and from IP-based networks. Third-generation systems are ultimately expected to migrate toward IP as part of the global trend toward carrying all traffic types over packet networks. However, the current UMTS WCDMA specification explicitly defines ATM as the transport layer in the RAN. While some IS-2000 RANs also use ATM, the 3GPP2 specifications allow the manufacturer to choose the underlying transport layer.

Given the planned migration paths of various 2G systems to 3G, 70% of 2G subscribers worldwide are expected to eventually migrate to some type of the CDMA version of 3G. The main network elements and interfaces referred to in the 3GPP specifications (see Figure 4.7) include the terrestrial components of the 3G system, referred to collectively as the RAN:

- User equipment, also called mobile station, subscriber unit, or simply handset, including mobile cellular telephones, handheld PDAs, and cellular modems connected to PCs;
- Node B, usually called the RBS, providing gateway services between the handset/RF interface and the RNC, via I_{ub} interface;

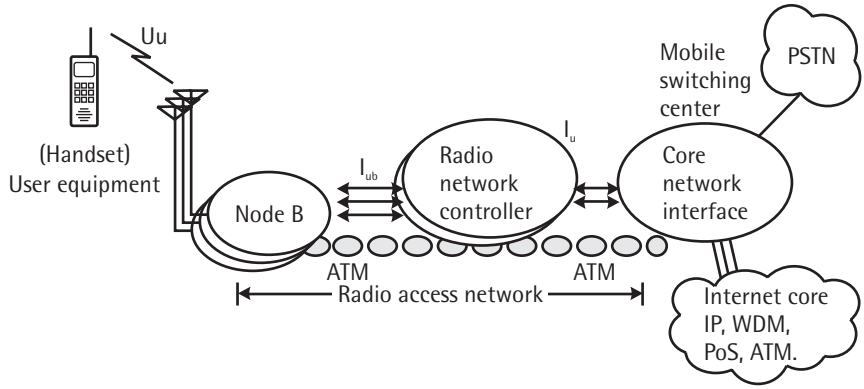

Figure 4.7 Main interfaces in 3G wireless network.

- RNC (or BSC), connecting to and coordinating as many as 150 base stations (the RNC manages activities such as handing over active calls between base stations);
- CN interface, referring to other terrestrial CN infrastructure connected to the RAN through I_u interface, such as the Internet and PSTN (the gateway device for this activity is usually called a mobile switching center or mobile multimedia switch).

An extensive set of protocols for communication within the RAN, to and from the user equipment, and between other networks has been developed by 3GPP. These protocols sit above AAL2 and AAL5. Together, they implement control-plane functions (e.g., signaling required to establish a call) and user-plane functions (e.g., voice or packet data). Wireless access in 3G network design constitutes one of the most important and driving requirements for the application of ATM extended with AAL2 switching (more information on ATM and AAL can be found in [2]).

The following description is the high-level data flow for voice and data call in CDMA2000 BSC and toward other parts of the RAN showing practical application of AAL2 and AAL5. New terminology used here aside from the well-known *backhaul* includes *sidehaul* and *fronthaul*. Fronthaul is a connection between BSC and MSC, and sidehaul is a connection between two BSCs.

Figure 4.8 shows the data flow for a land-to-mobile voice call in the 3G wireless system architecture. Voice traffic arrives at the BSC from the MSC (1). The fronthaul module receives PCM-encoded DS0s from the MSC and converts this traffic into AAL1 for transmission to the echo cancellation and vocoding module (2, 3). It also performs the required functions on the traffic channel, and converts the resulting data stream to the modified AAL2 format for transfer to the SEP module (4, 5). Modified AAL2 is a proprietary format used within the BSC. It is very similar to AAL2, with the exception that the channel identifier (CID) bit of the common part sublayer (CPS) packet is not used. This results in a simpler internal format, and therefore processing complexity and delays are both reduced. It should be noted that modified AAL2 is only used internally within the BSC, while all external AAL2 interfaces use standard AAL2. The selector element processing (SEP) module performs the selector element processing functions and sends the resulting modified AAL2 traffic channel to either the backhaul module for transmission to the appropriate RBS or to the sidehaul interface module for inter-BSC handoff (6, 7). The appropriate interface module converts the modified AAL2 data stream to the standard AAL2 format for transmission to the RBS or BSC (8). In the case of backhaul transmission to the RBS, the interface module also provides the necessary formatting. In the case of sidehaul transmission, the interface module converts the modified AAL2 format into AAL2.

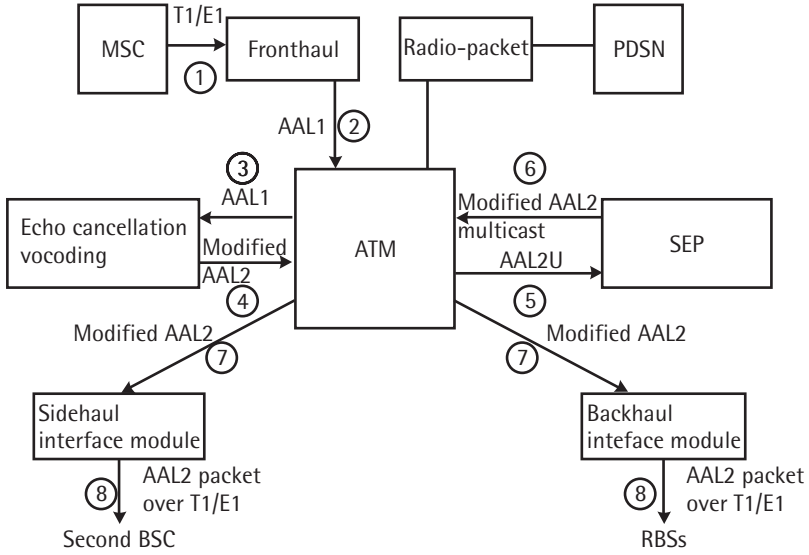

Figure 4.8 Voice call in 3G wireless network.

Figure 4.9 shows the data flow for a land-to-mobile packet call in the BSC architecture. Packet traffic arrives at the BSC from the PDSN (1). The interface module receives the IP/AAL5-formatted data stream over an OC-3 link. The interface module converts this traffic into AAL5 packets for transmission to the SEP module (2, 3). The SEP module performs the required selector element processing. This converts the output to the AAL5 format and sends the resulting AAL5 traffic channel to either the backhaul module for transmission to the appropriate RBS, or to the sidehaul module for inter-BSC handoff (4, 5). The appropriate backhaul module converts the AAL5 data stream to the standard AAL5 format for transmission to the RBS or BSC (6).

In the case of transmission to the RBS, the backhaul module also provides the necessary formatting. In the case of sidehaul transmission, the interface module converts the modified AAL2 format to standard AAL2.

4.2.7 3G Traffic Classes

When defining the UMTS QoS classes, also referred to as traffic classes, the restrictions and limitations of the air interface have to be taken into account. It is not reasonable to define complex mechanisms as they have been defined in fixed networks, due to different error characteristics of the air interface [3].

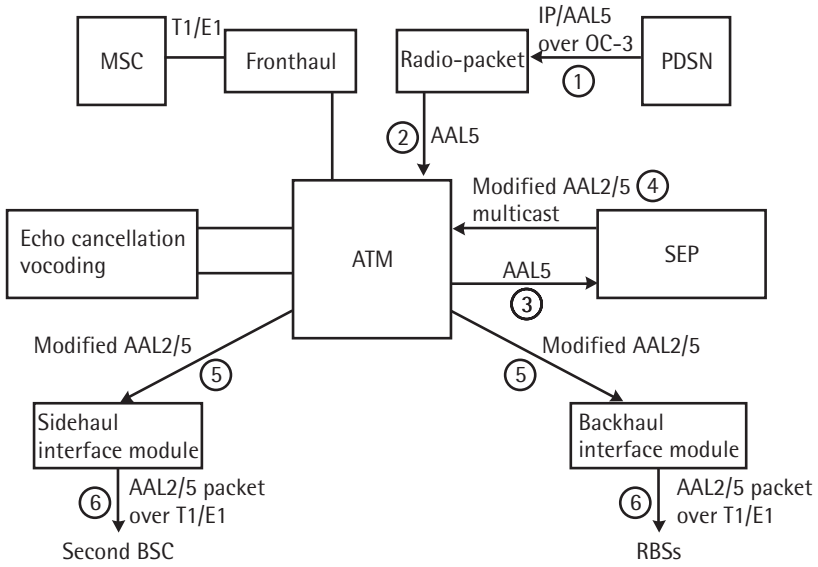

Figure 4.9 Packet call in 3G wireless network.

The QoS mechanisms provided in the wireless network must be robust and capable of providing reasonable QoS resolution.

There are four separate QoS classes:

1. Conversational class;
2. Streaming class;
3. Interactive class;
4. Background class.

The main distinguishing factor between these QoS classes is how delay sensitive the traffic is. Conversational class is meant for traffic that is very delay sensitive, while background class is the most delay-insensitive traffic class. Conversational and streaming classes are mainly intended to be used to carry real-time traffic flows. Conversational real-time services, like video telephony, are the most delay-sensitive applications, and those data streams should be carried in conversational class. The interactive background and classes are mainly meant to be used by traditional Internet applications like the Web, e-mail, Telnet, FTP, and news. Due to looser delay requirements as compared with conversational and streaming classes, both provide better error rate by means of channel coding and retransmission. The main difference between interactive and background class is that interactive class is mainly used by interactive applications (e.g., interactive e-mail or Web browsing), while background class is meant for background traffic (e.g., background download of e-mails or background file downloading). Responsiveness of the interactive applications is ensured by separating interactive and background applications. Traffic in the interactive class has higher priority in scheduling than background-class traffic, so background applications use transmission resources only when interactive applications do not need them. This is very important in wireless environments, where the bandwidth is low as compared with fixed networks.

4.2.7.1 Conversational Class

The most well-known use of this scheme is telephony voice. But with Internet and multimedia, a number of new applications will require this scheme, for example voice over IP and videoconferencing tools. Real-time conversation is always performed between peers (or groups) of live (human) end users. This is the only scheme where the required characteristics are strictly given by human perception.

A real-time conversation scheme is characterized by the transfer time being low because of the conversational nature of the scheme, and at the same time, the time relation (variation) between information entities of the stream must be preserved in the same way as for real-time streams. The maximum transfer delay is given by the human perception of video and audio conversation. Therefore, the limit for acceptable transfer delay is very strict, as failure to provide low enough transfer delay will result in unacceptable lack of quality. The transfer delay requirement is therefore both significantly lower and more stringent than the round-trip delay of the interactive traffic case. Real-time-conversation fundamental characteristics for QoS are as follows:

- Preserved time relation (variation) between information entities of the stream;
- Conversational pattern (stringent and low delay).

4.2.7.2 Streaming Class

When the user is looking at (listening to) real-time video (audio), the scheme of real-time streams applies. The real-time data flow is always aiming at a live (human) destination. It is a one-way transport. This scheme is one of the newcomers in data communication, raising a number of new requirements in both telecommunication and data communication systems. It is characterized by the fact that the time relations (variation) between information entities (i.e., samples, packets) within a flow must be preserved, although it does not have any requirements on low transfer delay.

The delay variation of the end-to-end flow must be limited to preserve the time relation (variation) between information entities of the stream. But as the stream normally is time aligned at the receiving end (in the user equipment), the highest acceptable delay variation over the transmission media is given by the capability of the time-alignment function of the application. Acceptable delay variation is thus much greater than the delay variation given by the limits of human perception. Real-time streamings fundamental characteristic for QoS is the preservation of time relation (variation) between information entities of the stream.

4.2.7.3 Interactive Class

When the end user (either a machine or a human) is on-line requesting data from remote equipment (e.g., a server), this scheme applies. Examples of human interaction with the remote equipment are Web browsing, database

retrieval, and server access. Examples of machines' interaction with remote equipment are polling for measurement records and automatic database inquiry. Interactive traffic is the other classic data-communication scheme that on an overall level is characterized by the request response pattern of the end user. At the message destination there is an entity expecting the message (response) within a certain time. Round-trip delay time is therefore one of the key attributes. Another characteristic is that the content of the packets will be transparently transferred (with low BER). Interactive traffic fundamental characteristics for QoS include the following:

- Request response pattern;
- Preserve payload content.

4.2.7.4 Background Class

When the end user, typically a computer, sends and receives data files in the background, this scheme applies. Examples are background delivery of e-mails, SMS, download of databases, and reception of measurement records. Background traffic is one of the classic data-communication schemes that on an overall level is characterized by the fact that the destination is not expecting the data within a certain time. The scheme is thus more or less delivery-time insensitive. Another characteristic is that the content of the packets is transparently transferred (with low BER). Background traffic fundamental characteristics for QoS include the following:

- Destination does not expect data within a certain time;
- Preserved payload content.

4.3 3G Transmission Networks

4.3.1 Replacing TDM with ATM in Transmission Networks

Wireless networks are leading the evolution of the information and communications society toward the mobile information society (MIS). This means that subscriber numbers are continuing to increase as mobile penetration reaches new heights.

Also, multimedia communications and other packet-based traffic will gradually increase their role and finally predominate in mobile networks. This development has already started with modest data volumes over current

mobile (and wireless in general) networks; a rapid increase in data applications and traffic is expected soon. New technologies and technical solutions enable higher data volumes right now in existing networks. In GSM networks, HSCSD and GPRS greatly expand these networks' capabilities to handle data traffic; they also enable new and user-friendlier applications thanks to the higher bit rates available. This development will continue with still higher bit rates over the air interface in the new 3G WCDMA- and CDMA2000-based networks and 1xEV-DO (also called HDR). These increasing data traffic volumes mean that the share of the packet-based traffic in the total traffic mix in the mobile network is increasing, that the same time as total traffic volumes are also rising rapidly.

Evolution of the circuit-switched networks into packet-based networks will take some time, and should be done in well-planned and managed steps, so that the efficiency of the mobile network is preserved during the change over phase. In many cases, basic mobile voice services are also growing quickly due to growth in the number of subscribers, which also contributes to the overall traffic increase and continues to require economic solutions for this type of traffic. Therefore, the well-planned steps are vital to manage mobile operators' cash flows and to make full use of existing investments. It is in the interest of a mobile-network operator to direct his future transmission-network strategy toward this expected increase in the penetration of advanced data services.

Transmission is an important element in any wireless network, affecting both the services and service quality offered, as well as the costs of the wireless operator. Optimization of transmission solutions is thus certainly worthwhile from the operator's business point of view. In current mobile networks, transmission has been optimized for the narrowband circuit-switched traffic and this type of traffic will continue to dominate for some years. However, as stated above, packet-based information over the mobile network will show rapid growth and any reasonable network development plan must take this into account and plan for a smooth and economic transition and evolution path for the transmission network. So, in broad terms, the transmission network must continue to provide well-engineered and economically optimized solutions for the growing volumes of circuit-based traffic, while at the same time develop the strategy and the readiness to cope with the even faster growing data traffic of the future. This type of transmission solution is needed in all parts of the mobile network, both in access networks with many points and low-capacity links, as well as in CNs with high traffic volumes. This means, for example, that in a wireless network, a transmission solution is needed that provides for efficient transport of large number of

voice channels and that can evolve to also carry packet-based traffic, either ATM or IP or both.

The solutions might be similar or different in different parts of the network and even the role and share of the different traffic types (TDM, ATM, and IP) might be different, but the transmission network must support them all in a planned and managed way. In any mobile network, there are different transmission needs, typically divided into two main application areas with their own characteristics:

1. The access transmission network, which connects the base stations to the closest network control or network hub point, and called here the access transmission network or backhaul. It would include the radio base station, BSC (or RNC), and BSC/MSC.
2. The core transmission network, which connects the control (or hub) points to the mobile network switching centers, and called here the core transmission network.

The radio network will be connected to the CN by a backbone network (access and core transmission network), allowing wideband access and interconnection of subscribers. The 3G backbone network can use any transport technology, but is certain to be based on packet technologies, such as ATM and IP. The backbone network is built as a mesh of IP routing or ATM switching nodes interconnected by point-to-point links. Technologies such as IP over ATM may be used that uses ATM switching to multiplex IP traffic. This IP over ATM architecture supports voice traffic alongside IP. Many vendors prefer a pure end-to-end IP approach, whereas others prefer an ATM/IP hybrid to guarantee quality of service. Alternatively, IP over SONET/SDH could be a different backbone network solution that would eliminate the ATM layer by establishing point-to-point links between IP routers directly over SONET/SDH rings that run over a dense-wavelength-division multiplexing (DWDM) layer. This enables terabits per second (Tbps) of aggregate network bandwidth.

When transporting voice over a packet-based network, overhead is added to each voice packet. The amount of overhead depends very much on what the protocol stacks for the user data look like. Figure 4.10 sketches the different possible protocol stacks. So if speech is transported over an AAL2/ATM connection, 5 octets are needed for the ATM cell header and 3 octets for the AAL2 minicell header. Assuming the AMR CODEC is used (WCDMA systems), the speech frame size can vary between 10 and 40

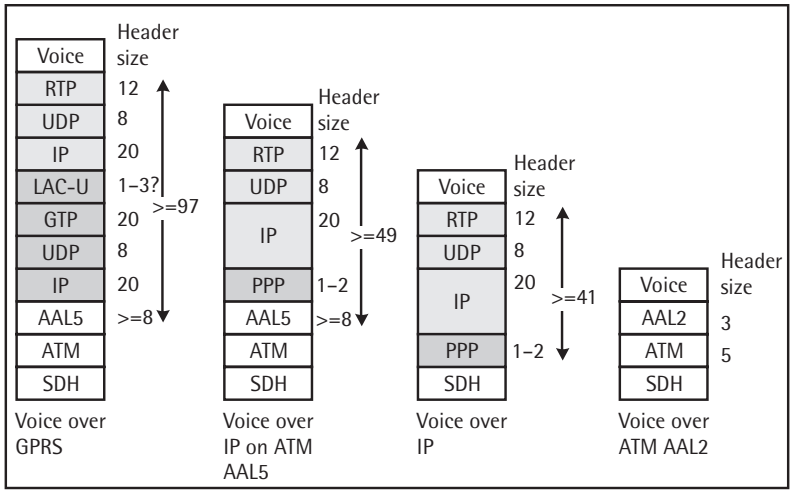

Figure 4.10 User-plane protocol stack alternatives.

octets, depending on the quality of the radio link. If for an average connection the speech frame size is about 21 octets, two speech frames will fit into an ATM cell. So the overhead is 5.5 octets per speech frame in average. Bandwidth increase factor is 1.26.

If IP is used directly on SDH, the overhead is calculated as about 41 octets per speech frame. By applying IP header compression, the overhead can be reduced to about 13 octets per speech frame. Bandwidth increase factor is 2.95 (1.62 for header compression). If IP is transported over ATM, the overhead is about 49 octets per speech frame plus the ATM cell header. Assuming again AMR CODEC, the IP datagram varies between 59 (49 + 10) and 89 (49 + 40) octets in size. This means that two ATM cells always have to be used, leaving empty space in the second ATM cell. So 106 (2 × 53) octets are always sent per speech frame. Bandwidth increase factor is 5.05. If the speech is sent via GPRS access, the calculation is even worse.

When the transmission is packet based, silence suppression can be applied. This means that typically only 40% of the time does data need to be transferred. The rest is silence, and no bandwidth is used. This is common for ATM and IP transmission. IP is a best-effort transport, so IP as such does not guarantee any QoS.

Of course, platforms and systems such as the value-added service centers, gateways, billing systems, customer service elements, IN systems, and the

like will also need to be upgraded. High-bandwidth over-the-air applications (data, video, etc.) also require high-capacity transmission systems, much more so than 1G and 2G networks. Increasing subscriber numbers and share of data traffic creates significant growth in the transmission capacity needed for long-term network evolution. The implication of this is an increasing role of fiber and high-capacity microwave systems (SDH/SONET) in future transmission networks and their physical-layer implementation.

There are more and more data traffic and applications that create bursty traffic. Thus, in order to provide cost-effective transmission solutions for data traffic, packet-switching solutions introducing dynamics to traffic handling must be provided. Also, investments made in fiber- and microwave-radio-based transmission links are very important as new evolution phases can be built on top of existing infrastructure. Wireless networks of the next generation will also require new, more advanced solutions for the core and access transmission networks.

Data traffic is inherently variable, and transporting it over the TDM network is inefficient. In a TDM-based approach, timeslots are dedicated for connections regardless of whether information is actually being sent. In a multiservice network, the underlying network can be physically subdivided into multiple networks, one for each service (i.e., voice, data, private lines, etc.). Using an ATM-based infrastructure, much more efficient use of transmission network is possible, since ATM allocates bandwidth on demand based on immediate user needs.

In 2G wireless networks, deterministic multiplexing is applied, when each connection is characterized by a constant bandwidth (e.g., one timeslot). The minimum needed bandwidth over the physical link is then simply the sum of the constant bandwidths of the connections. Since the traffic characterization is not probabilistic, statistical gain is not available. Third-generation wireless networks use packet-switched (ATM) systems and statistical multiplexing. When several connections from VBR sources are multiplexed together, a statistical multiplexing gain is obtained (Figure 4.11), because there is a certain probability that traffic bursts on different connections do not appear at the same time.

It is possible to maintain the same blocking probability with less bandwidth if statistical multiplexing is used instead of deterministic multiplexing. The price for it is that the QoS (packet delay and loss) will not be ensured in a deterministic, but in a probabilistic, fashion.

Statistical multiplexing of data traffic can occur side by side with the transmission of the delay and loss-sensitive traffic such as voice and video [4]. Like voice telephony, ATM is fundamentally a connection-oriented tele-

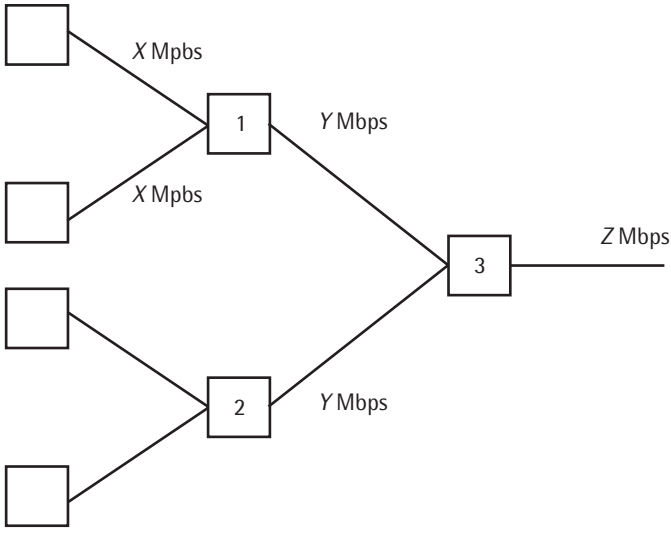

Figure 4.11 Multiplexing gain.

communications system. That means that a connection must be established between two points before data can be transferred between them. An ATM connection specifies the transmission path, allowing ATM cells to self-route through an ATM network. Being connection-oriented also allows ATM to specify a guaranteed QoS for each connection.

Since voice telephony is a real-time application, delay, among other quality measurements, is the most important factor that affects the quality of voice. According to ITU-T Recommendation G114, an end-to-end delay of 0 to 150 ms is acceptable for most applications. A delay of 150 to 400 ms is acceptable assuming that the administrators are aware of the transmission time impact of the transmission quality of user application, but any delay over 400 ms is unacceptable for general network planning purpose [5]. AAL2 has been designed and used to reduce the packing delay for narrowband services. The idea is to multiplex voice packets from several sources onto one ATM cell so that the time to fill a cell can be reduced significantly.

IP over ATM and use of IP routers is a basis for 3G wireless network transmissions. An IP router is a packet-switching device used to connect several different networks to form one common network based on IP

networking technology. Based on its understanding of the network of which it is a part, the router decides how each packet is going to be forwarded, but it also must be able to differentiate between high-priority packets and low-priority packets and make the right decision and avoid internal congestion at the same time. IP routers for wireless networks must efficiently be able to handle small packets of data, low-speed links, delay-sensitive traffic, synchronization, a large number of nodes, and continuous on-line connections. These demands come from the nature of wireless traffic, where low-priority packets cannot block the way for high-priority voice packets. Routers in wireless networks must also be able to provide radio base stations with a high-quality synchronization signal that is distributed via transmission links between routers and base stations, assuming that a global positioning system (GPS) is not used. Routers in 3G wireless networks are an integral part of products, such as RNC for WCDMA, BSC for CDMA2000, and MGWs for packet data services.

4.3.2 Importance of AAL2

Third-generation wireless networks will integrate multimedia services, 2G voice services, and TCP/IP networks [6]. As mentioned earlier, data traffic is inherently variable, and transporting this type of traffic over an underlying TDM network is inefficient and thus expensive. An ATM-based approach can take advantage of the statistical nature of data traffic in addition to the constant rate of voice in order to provide a more bandwidth-efficient solution. Of course, ATM over the fiber-optic media has to have the same traditional transmission-network functionality like multiplexing, grooming, add-drop multiplexing, and protection while offering additional services like guaranteed performance, virtual private networks, prioritized rerouting, adjustable statistical multiplexing levels, security screening, customer network management, and so on. With an ATM-based transmission infrastructure, provisioning circuits, changing their bandwidth, and monitoring them is possible through the use of a standard network management system.

In deterministic multiplexing, each connection is allocated its peak bandwidth. In ATM, statistical multiplexing is used where the amount of bandwidth allocated in the network to the VBR source is less than its peak, but greater than the average bit rate. So, the sum of the peak rates of connections multiplexed can be greater than the link bandwidth as long as the sum of their statistical bandwidths is less than or equal to the provisioned link bandwidth. The bandwidth efficiency due to statistical multiplexing increases as the statistical bandwidths of connections get closer to their

average bit rates and decreases as they approach their peak rates. In general, though, statistical multiplexing allows more connections to be multiplexed in the network than deterministic multiplexing, therefore allowing better utilization of network resources.

Generally speaking, efficiency gain due to statistical multiplexing is a factor of a number of different connection characteristics and network parameters. For example, depending how bursty the sources are and the length of those bursts, the efficiency gain due to statistical multiplexing may or may not be significant. In 3G wireless networks, core and access transmission networks are ATM based, so the calculations of the required transmission links will be a lot different from 1G and 2G circuit-switched wireless networks.

The AAL performs functions required by the user, control and management planes, and supports the mapping between the ATM layer and the next higher layer. The functions performed in the AAL depend upon the higher-layer requirements. In short, the AAL supports all of the functions required to map information between the ATM network and the non-ATM application that may be using it.

The transport of voice traffic over ATM networks has been a fundamental principle of its design from the start. However, its deployment has been problematic, and telephony over ATM is still an area that is being developed. Normally we regard voice over ATM as the transport of voice over emulated circuits (a replacement for a PRI or T1 carrying voice calls). ATM has always been able to carry voice and data over the same wires, since that is what it was designed to do. In an uncompressed format (standard 64-Kbps PCM), the traffic is CBR and is presented to the network over a circuit using AAL1. When compression is deployed it is possible to use AAL5, or the AAL2, which is specifically designed to work with compression hardware [7]. But until recently, voice transport via ATM relied primarily on AAL1 circuit emulation, which adds 12% to 15% overhead to every voice circuit. Moreover, AAL1 lacks bandwidth-saving features like voice compression and silence suppression. With AAL1, an emulated T1 (1.544-Mbps) circuit requires 1.74 Mbps of ATM bandwidth and is not really applicable for cost-sensitive real-world applications.

The new AAL2, on the other hand, was designed specifically for cost-effective voice transport. AAL2 is used in 3G wireless networks as a backhaul connection between RBSs and the BSC. A new adaptation layer was required to provide the flexibility for network operators to control delay on voice services and to overcome the excessive bandwidth needed by using structured circuit emulation. AAL1 simply cannot be extended to meet these new ATM

networking requirements. AAL2, as specified in ITU-T Recommendations I.363.2 (1997), I.366.1 (1998), and I.366.2 (1999), carries the specific mandate to provide efficient voice-over-ATM services. Developed by the ITU and adapted by the ATM Forum in February 1999, AAL2 as defined in ITU-T Recommendation I.363.2 includes the following capabilities:

- *VBR-real time (VBR-rt) support.* While AAL1 supports only CBR transmission, AAL2 supports both CBR and a traffic class called VBR-rt, which is a better fit for voice calls and other applications that send information at a variable rate.
- *Statistical multiplexing.* Unlike AAL1 circuit emulation, which reserves a fixed amount of bandwidth for each circuit, AAL2 can allocate unused bandwidth to other traffic on demand.
- *Cell sharing.* AAL2 can pack several short packets from different sources into one ATM cell, letting multiple connections share the same bandwidth with less overhead.
- *Variable packet fill delay.* To let service providers balance delay against efficiency, AAL2 supports variable settings for packet fill delay, the time allotted to stuff packets into cells before putting them on the wire.
- *Voice optimization.* AAL2 includes specific bandwidth-saving features like voice compression, silence detection and suppression, and idle voice channel suppression.

Voice communication by nature is half duplex; one person is silent while the other speaks. There are also pauses between sentences and words with no speech in either direction. By taking advantage of these two characteristics, it is possible to save bandwidth by halting the transmission of packets during these silent periods. This is known as silence suppression or digital speech interpolation (DSI). The extra bandwidth saved from the silent period of one voice channel can be used by other connections if using AAL5- or AAL2-based connections. With AAL1 (circuit emulation-based services), these savings can only effectively be used by ABR or UBR services since the connection admission control mechanism will have allocated the bandwidth required for CBR QoS. This technique can improve bandwidth utilization by as much as 40%. In order to create a natural-sounding conversation, background noise can be generated at the far end to recreate a realistic environment.

AAL2's VBR service handles voice-over-ATM far more efficiently than the CBR service of AAL1, with its inefficiently utilized, permanently allocated bandwidth. Before AAL2, users wanting to implement voice-over-ATM had to live with AAL1's limitations or adopt a proprietary solution (increasing network efficiency but negating interoperability). The new standard means that ATM switches can extend to voice the benefits of ATM's statistical gain. Access connections using this scheme can transport voice circuits over the same facilities as data circuits, minimizing the use of precious bandwidth. AAL2 is designed to make use of the VBR ATM traffic classes (with higher multiplexing gains), providing bandwidth-efficient transmission of low-rate, short and variable packets for delay-sensitive applications. AAL2's structure lets network administrators take traffic variations into account in the design of an ATM network optimized to match traffic conditions. AAL2 also enables multiple user channels on a single ATM virtual circuit and varying traffic conditions for each individual user or channel. Its structure also provides for the packing of short-length packets into one (or more) ATM cell and the mechanisms to recover from transmission errors. Compared to AAL1 and its fixed payload, AAL2 handles a variable payload within cells and across cells. This provides a dramatic improvement in bandwidth efficiency over either structured or unstructured circuit emulation using AAL1.

It has been proven that AAL2 provides good mechanisms for fine adjustment of packetization delay and for achieving high bandwidth efficiency for low bit rate and VBR application. The cost is some overhead in the form of headers. This overhead is not important at all when compared with that for techniques, such as partial filling of cells (AAL2 behaves at least as well as partial filling), but may be significant when users have a choice among several AALs. For example, for transmitting long packets (longer that 45 octets), the user may chose either AAL2 (through it segmentation SSCS) or AAL5 (explicitly specified for long packet data). It is known that for small packets (up to a few hundred octets) AAL2 is more efficient than AAL5, while for long packets AAL5 is definitely recommended. In this case the user should select the AAL based on its data-generation patterns.

A very realistic application for AAL2 switching has been identified for 3G of wireless networks, in the frame of IMT-2000 standardization, to support the functionality known as soft handoff.

The ATM using AAL2 for narrowband services described in this specification fulfills an urgent market need for an efficient transport mechanism to carry voice, voice-band data, circuit-mode data, frame-mode data, and fax traffic. Voice transport will include support for compressed voice and

noncompressed voice together with silence removal. Reference [8] describes the procedures and signaling required to support the efficient transport of narrowband services across an ATM network between two interworking functions (IWFs) to interconnect pairs of non-ATM trunks. It specifies the use of ATM virtual circuits with AAL2 to transport bearer information and ATM virtual circuits with AAL2 or AAL5 to transport CCS. The virtual circuits used may be PVCs, SPVCs, or SVCs. The specification supports the transport of common channel signaling (CCS) information as well as channel-associated signaling (CAS) information. ATM trunking using AAL2 provides both switched and nonswitched services to the narrowband network.

4.3.3 QoS Concept

From an engineering planning prospective, the initial choice of network topology can seriously impact initial investment and future flexibility. Although much has been learned in terms of network topology (system design), LOS and path-loss (coverage) issues, and QoS optimization issues, much is still in the process of being developed, tested, deployed, and tested again in the field. It is already proven that the delivery of raw bandwidth over wireless media without acceptable QoS will not result in market acceptance. Without clearly understanding QoS in the context of a wireless broadband access system, it is unlikely that the underlying architecture and the resulting hardware and software design will result in a successful system.

The issues of quality delivery are somewhat more complex for wireless broadband access systems than for wireline systems. Data-delivery problems include slow peripheral access, data errors, dropouts, unnecessary retransmissions, traffic congestion, out-of-sequence data packets, latency, and jitter. In addition, wireless access introduces high inherent BER, limited bandwidth, user contention, radio interference, and TCP traffic-rate management. QoS mechanisms must address all of these concerns.

In data networking, quality usually implies the process of delivering data in a reliable and timely manner. The definition of reliable and timely is dependent upon the nature of the traffic being addressed. These terms may include references to limitations in data loss, data retransmission, and packet order inversions, as well as data accuracy expectations and latency variations (jitter). QoS is a complex concept, requiring a complex mechanism for implementation. A casual user doing occasional Web browsing, but no FTP file downloads or real-time multimedia sessions, may have a different definition of QoS than a power user of large databases or financial files, frequent

H.323 video conferencing and IP telephony. For example, in a wireless system, QoS mechanisms must cope not only with considerations particular to the wireless environment, but with wireline-networking considerations as well. In ATM networks, traffic descriptors are usually used only as a rough guide, and many service providers systematically practice overprovisioning and allocating a lot more bandwidth than necessary [9]. Although not a very efficient system, the perceived QoS is satisfactory in most cases.

The migration from circuit-switched to ATM and a packet-switched network has also affected QoS mechanisms. An IP-centric wireless system for packet-switched network traffic requires a new approach to provide optimal QoS performance. The use of QoS as the underlying guide to system architecture and design constitutes the fundamental differentiation between wireless broadband access systems designed with traditional circuit-centric or ATM-centric approaches and IP-centric wireless broadband access systems. Queuing is the commonly accepted tool for manipulating data communications flows. Data packets must be queued for packet headers to be examined or modified, for routing decisions to be made, or for data flows to be output on appropriate ports. However, queuing introduces a delay in traffic streams that can be detrimental and can totally defeat the intent of queuing. Excessive queuing can delay time-sensitive packets beyond their useful time frames or increase the round trip time, producing unacceptable jitter or causing the data transport mechanisms to time-out. Therefore, queuing must be used intelligently and sparingly, without introducing undue delay in delay-sensitive traffic.

To achieve high-quality (often referred to as toll-quality in circuit-switched networks) voice transmission, the absolute amount of transmission delay, as well as the variation in that delay, called jitter, must be kept low. In a wireless environment where TDMA, FEC, and other such techniques are necessary, queuing must be used only to enable packet and radio-frame processing. However, in the case of real-time flows, the overall added delay in real-time traffic should be held to below about 20 to 25 ms. This amount of delay is not perceptible to the human ear. Several additional factors contribute to delay in data networks, including cell-packing delay, coding and compression delay, queuing delay, and freeze-out delay. Cell-packing delay, or packetization delay, is simply the amount of time it takes for the sending device to fill packets before they can be sent. Coding and compression delay represents the amount of computational time that is needed for the sampling, quantization, and compression of the signal. This delay can become significant if a large amount of compression is being done to the signal. Queuing delay occurs in networks experiencing congestion and represents

how long packets must wait in queue at a bottleneck node in the network. The use of queue management as the primary QoS mechanism in providing QoS-based differentiated services is a simple and straightforward method for wireless broadband systems. Wireless systems are usually more bandwidth-constrained, however and are therefore more sensitive to delay than their wireline counterparts. So QoS'-based differentiated services must be provided with mechanisms that go beyond simple queuing.

In a network not experiencing congestion, the most significant type of delay should be freeze-out, or serialization delay. This delay is due to the fact that packets take a finite amount of time to transmit, and during this time, the channel is not available to transmit any other information. For example, when two voice channels are packetized and sent over the same data link, each has to wait to send its packets while the link is busy transmitting the other channel's packets. Freeze-out delay can be roughly quantified as the size of packets divided by the speed of the channel (in bps). Freeze-out delay by itself provides a minimum estimate of one-way delay for packetized voice traffic; in reality, actual delay would be higher due to the additional delays mentioned before. TCP controls transmission rates by sensing when packet loss occurs. Because TCP/IP was created primarily for the wireline environment with an extremely low inherent BER (today it is on the order of 1×10^{-9} or better for fiber optics), TCP assumes any packet loss is due to network congestion, not error. Therefore, TCP assumes that the transmission rate exceeds the capacity of the network and slows the rate of transmission; however, packet loss in the wireless link segment is due primarily to the high inherent BER, not congestion.

With a range of data flows, each having different bandwidth, latency, and jitter requirements, the IP-centric wireless system must be able to manage QoS mechanism parameters over a wide range in real time. The QoS mechanism must be able to alter system behavior to the extent that one or more data flows corresponding to specific applications be transparently switched on and off from the appropriate end users. This approach is in contrast to other QoS mechanisms that seek to achieve high QoS by establishing circuit-centric connections from end to end without regard for the underlying application's actual QoS requirements. By providing an application-specific QoS mechanism, scarce wireless bandwidth can be conserved and dynamically allocated where needed by the QoS mechanisms associated with each application type.

Mobile quality of service (M-QoS) has been defined in order to augment traditional QoS requirements found in wired ATM networks [10]. This wireless QoS refers to the QoS parameters associated with the wireless

links, such as link delay, bit error rate, and channel reservation, as well as the performance parameters associated with the handoff blocking probability and cell loss.

4.3.4 ATM Physical Layer

Bandwidth is the key resource in both circuit- and packet-based networks. In circuit-based networks, a fixed amount of bandwidth is dedicated to a call in progress. Since 55% to 60% of the call consists of silence, circuit-based networks do not make optimum use of bandwidth. The principal advantage of packet-based networks is that they use bandwidth much more efficiently. Unlike circuit-based networks, packets from many different sources share a circuit, allowing for efficient use of fixed capacity. ATM seems to be the technology of choice for many different 3G wireless networks.

The physical layer is at the lowest level of the ATM stack. It takes the full cells from the mid-layer and transmits them over the physical medium. The ITU-T originally defined only two speeds that should be supported by ATM (i.e., 155.52 Mbps and 622.08 Mbps); however, over time a number of additional speeds and interfaces have evolved, going as low as E1/T1 and as high as the Gbps range. The physical layer itself is subdivided into two sublayers: the transmission convergence (TC) sublayer and the physical medium dependent (PMD) sublayer. These two sublayers work together to ensure that the optical or copper interfaces receive and transmit the cells efficiently, with the appropriate timing structure in place. ATM, being an international transmission technology, has to be able to work with a variety of formats, speeds, transmission media, and distances that may vary from country to country. The standardization of the physical layer interfaces enabled just such connectivity. Single-mode fiber, multimode fiber, coaxial pairs, and shielded and unshielded twisted pairs are today all standardized for use in the ATM environment.

The TC sublayer takes care of header error check (HEC) generation and verification, cell scrambling and descrambling, cell delineation, and decoupling. The HEC is a one-byte field in the ATM cell header, which protects the header from errors. The PMD sublayer covers bit timing, line coding, the physical connectors, and signal characteristics.

The UNI documents detail the physical media types allowed at the user interface and the details differ for the public and private UNIs. For example, Category 5 twisted pair is permitted at the private UNI, but not at the public UNI. The original objective for the physical layer was operation over the SDH and SONET only. When the ATM Forum V3.0 and V3.1 specifications were

ratified, however, other interfaces were included. These interfaces were the DS3 and a 100-Mbps interface based on the transparent asynchronous transmitter/receiver interface (TAXI) fiber distributed data interface (FDDI) standard. DS1 operates at 1.544 Mbps and is approved for ATM over twisted pair at a distance of up to 3,000 feet. DS3 operates at 44.736 Mbps on coaxial cable up to 900 feet. STS-1 (51.84 Mbps), STS-3c (155.52 Mbps), and STS-12 (622.08 Mbps) operate over single-mode fiber up to 15 km. E1 (2.048 Mbps) and E3 (34.368 Mbps), together with the Japanese standard J2 (6.312 Mbps), are standardized for coaxial cable with no distance specified. Recently, a definition for a 2.5-Gbps physical interface (an SDH interface) was completed. This definition describes how cells are mapped to this higher-speed transport.

4.3.4.1 ATM in SONET/SDH Fiber-Optic Networks

The integration of SONET/SDH and ATM involves more than offering interfaces that allow connectivity between the technology networks. The high-speed transmission attributes of SONET/SDH and the switching and bandwidth-management capabilities of ATM complement each other to form the foundation of broadband networking. SONET/SDH provides ATM with access to a high-speed infrastructure; conversely, ATM offers the high-speed traffic that takes full advantages of a SONET/SDH infrastructure.

Key to integration of ATM and SONET/SDH is the ability to monitor and react to failures, ensuring survivability. SONET/SDH rings provide protection against the failures within the core transmission network, but should also extend to protect broadband services against link and node failures. The integrated ring provides facility-layer reliability (link) complementing ATM reroute, which provides ATM-layer reliability (node) in addition to link reliability. The ATM service platform should be able to intelligently monitor the SONET/SDH payload overhead. Upon the detection of a failure, the ATM services should be able to provide SONET/SDH 1+1 automatic protection switching as defined by ITU and Bellcore specifications. The ATM switch needs to provide one-for-one redundancy on the ports, and the complete switchover must take place within the specified 50-ms QoS parameter defined by service providers. If SONET/SDH switchover is unable to overcome the network failure, ATM reroute will reestablish the failed connection over a new route. These inherent SONET/SDH switchover-protection capabilities and ATM reroute capabilities are critical to new broadband services.

Both SONET/SDH and ATM have operations and maintenance information in their headers, but use the information in complementary ways, with

SONET/SDH checking for errors on a span-by-span basis, while ATM considers performance on an end-to-end basis between different switches. ATM will send 50 cells per SONET/ATM frame, each with its own operations, administration, and maintenance functionality, making it possible for the ATM layer to detect a fiber cut much sooner than the SONET/SDH layer. This way, ATM can enhance the speed with which SONET/SDH detects problems.

As per ITU-T I.630, ATM Protection Switching, the individual VP/VC protection-switching concept was developed to apply primarily to the situations where server-layer protection-switching does not exist. It is useful to protect only a part of VPs/VCs that need high reliability. The rest of the VPs/VCs remain unprotected. This helps to reduce the necessary bandwidth for protection and can be used for protection against ATM-layer defects as well as physical-layer defects.

The ATM protection-switching architecture can be a 1+1 type or an m:n type. In the 1+1 architecture type, a protection entity is dedicated to each working entity with the working entity bridged onto the protection entity at the source of the protected domain. The traffic on working and protection entities is transmitted simultaneously to the sink of the protected domain, where a selection between the working and protection entity is made based on some predetermined criteria, such as server defect indication. In the m:n architecture type, m dedicated protection entities are shared by n working entities, where $m \leq n$ typically. The bandwidth of each protection entity should be allocated in such a way that it may be possible to protect any of the n working entities in case at least one of the m protection entities is available. When a working entity is determined to be impaired, it first must be assigned to an available protection entity followed by transition from the working to protection entity at both the source and sink of the protected domain. It is noted that when more than m working entities are impaired, only m working entities can be protected.

In general, if lower-layer (e.g., SDH or optical) protection mechanisms are being utilized in conjunction with ATM-layer protection mechanisms, then the lower layers should have a chance to restore working traffic before the ATM layer initiates protection actions. The objective here is to avoid unnecessary protection actions and any issues of contention.

4.3.4.2 ATM in Microwave Radio Networks

ATM was originally developed for use in high-reliability fiber-optic SONET/SDH networks and not for more difficult media like radio. Over

the last few years more and more transmission systems, especially those in wireless networks, are using ATM over the microwave networks. These radio systems carrying packetized traffic such as ATM (or frame-relay) have to be designed in a way that takes into account behavior of this kind of traffic. Because ATM is primarily designed for an essentially error-free environment, in the wireless arena the sources of errors and their consequences on ATM traffic and its QoS are being studied today [11]. Although important in any network, error bursts are expected to be very significant sources of degradation in the microwave network.

ATM is designed for low-BER links, and the radio links with just a moderate BER can cause unacceptably high cell-loss and misinsertion rates [12]. By definition, a misinserted cell is a received cell that has no corresponding transmitted cell on the considered connection. Cell misinsertion on a particular connection is caused by defects on the physical layer affecting any cells not previously associated with this connection. Since the mechanisms that cause misinserted cells have nothing to do with the number of cells transmitted on the observed connection, this performance parameter cannot be expressed as a ratio, only as a rate. In a process of dimensioning microwave point-to-point systems for ATM traffic, there are a number of issues to be considered. Since bit errors in the microwave system typically appear in multiples and spread less than the ATM header length, the single-bit-header correction feature may not improve cell-loss rate as much as predicted and intended. The latest research shows that the BER is degraded approximately one decade from the microwave radio system to the ATM CBR virtual circuit due to the cell loss.

In wireless networks, broadband terminals as well as smaller and denser cells will increase the total capacity needs enormously. Many of today's 1E1/T1 links will be increased to STM-1/OC-3 and higher capacities and will require high-capacity SDH/SONET microwave radios even at spur links to the last cell site in the network. Aside from the capacity, these microwave radios need a very sophisticated error-correction technique to satisfy ATM transport layer requirements.

Normally, in a fiber-optic system, BER should be 10^{-9} measured at the ATM CBR virtual circuit. The same quality corresponds to BER = 10^{-10} in the microwave radio system. Systems with BERs worse than 10^{-6} are considered unavailable.

Although still under research, the following facts should be kept in mind:

- To achieve 1×10^{-9} user BER from the ATM network, 1×10^{-10} is required on the SDH radio link.

- When BER is above 1×10^{-6} (low BER) end users' quality is very low (1 cell lost per second). AIS generation at low BER should be considered both for the purpose of rerouting and disconnecting.
- For ATM NMS operation, 8×10^{-5} is required on the SDH microwave radio link.

4.3.5 Traffic Modeling and Simulation Tools

The transmission network consists of the elements required for the transport of call and signaling information between the major nodes in the service network. The core transmission network covers elements required for the transport of calls and signaling between nodes in the CN. Typically, the major nodes are POIs, MSCs, BSCs, TSCs, STPs, HLRs, and ESNs. The access transmission network consists of the elements that are required for the transport of call and signaling information between the RBSs and BSCs in the service network.

Due to recent and ongoing efforts of both regional and global standardization and research processes, there is a wide consensus regarding the basic architectural aspects of 3G mobile communications systems, including applying ATM as switching and multiplexing technology [11]. In order to utilize transmission facilities efficiently, traffic simulation models and tools will be used in the process of wireless-network planning. All three aspects of network planning (RF, core, and access) will be based on statistical methods and traffic-simulation tools.

Multiplexing of different traffic streams has consequences on the dimensioning, since there will in many cases be a nonnegligible gain compared to just adding the capacity demands of individual traffic streams. The reason is that the variation of the individual streams is usually smoothed out in the aggregated traffic stream. It is often convenient to split up this so-called multiplexing gain into traffic levels:

- Multiplexing gain on a call level for all services (only call arrivals are considered, often modeled as a Poisson process);
- Multiplexing on a transaction level (for services with discontinuous transmission, such as interactive services, voice with DTX);
- Multiplexing on a packet level (as with VBR packet services).

For constant bit rate services with continuous transmission, multiplexing gain can be obtained on a call level only. In order to include multiplexing

gain on subcall levels (for example, with VBR services) when dimensioning, it must be considered that arrivals of transactions or packets (depending on the level considered) often exhibit a pattern that deviates significantly from the Poisson process.

Inverse multiplexing for ATM is a method that makes it possible for several physical links to carry a single ATM stream. The main advantage is increased robustness. The traffic is distributed on all physical links and in case of a failure on one physical link, the traffic is distributed over the remaining physical links. No traffic will be lost if the remaining capacity is sufficient. Another important factor is that larger links result in an increased potential for statistical multiplexing gain. In order to perform the calculation, the following data is needed:

- Service characteristic description;
- Traffic-related data;
- GoS requirement.

The service description contains parameters representing the basic behavior of the service. The traffic mix is responsible for setting how intensive the given services are used in the RBS. This data is most likely based on the outcome of the cell-planning procedure. The GoS requirement reflects the percentage of the users that will not get service because of the limited resource on the transmission and switching network.

There are two types of services that can be defined and characterized: circuit- and packet-type services. Circuit-type services are for services with definite bandwidth demand and holding time. These types of services are delay sensitive and demonstrate CBR or VBR with well-defined characteristics (like voice). The parameters that describe the circuit-type services are equivalent bandwidth and overhead.

The equivalent bandwidth for CBR is the bit rate itself. For VBR it is somewhere between the peak and the average bit rate and it is very dependent on the traffic characteristics. The overhead of the service takes into account the amount of extra data volume that must be transferred on the transport network excluding the ATM overhead, which is included later on, when the logical links are mapped onto the physical ones. The overhead can consist of protocol overheads like frame protocol, MAC, RLC, and so on. Calculation of the overhead should take into consideration what kinds of overheads are already added in the equivalent bandwidth.

Packet-type services can be used to describe best-effort services that are not delay-sensitive services. For best-effort services, well-defined resources are not guaranteed within a certain time, and only an average throughput over a longer period of time (at least an hour) can be guaranteed. Mean BHT, burstiness factor, and overhead characterize this type of service. The average throughput is given by the mean BHT in bytes per hour. The burstiness factor describes how bursty the traffic is; that is, what portion of a given bandwidth can be utilized by the traffic (in an average level) to be able to handle the deviations from the average bit rate. Its value is between 0 and 1, where burstiness equal to 1 means nonbursty traffic (a typical value is 0.6 or 0.7 in the case of traffic generated by IP applications). The smaller the burstiness factor, the more bursty the traffic, requiring higher bandwidth. The overhead parameter has the same role as in the circuit-service type.

The other group of information is traffic data and GoS requirement specific to an RBS. The traffic data is given by the traffic mix parameter, which is a list of services offered by the RBS and their level of usage. In case of circuit service, the latter one is an Erlang value describing the traffic volume. For packet service it is the average number of simultaneously attached users (as defined in WCDMA). In both cases they are meant to be a busy-hour value. The GoS requirement is responsible for adjusting the percentage of the calls that cannot be served because of transmission resource shortage. The corresponding parameter name is blocking. It can be set for each RBS and affects only the circuit-type services.

Trunk calculation for RBS-BSC backhaul interface is based on RBS capacity and can be defined in terms of the number of sectors and the number of simultaneous calls it can support. Let us assume that under full configuration, the CDMA2000 RBS will support 4 RF carriers, equating to a maximum of 12 RF sectors. This requires that 8 E1 (or 2 E1s per RF carrier) backhaul connections between the RBS and BSC. The actual number of E1/T1 spans, however, depends on traffic demand. Due to the high efficiency of ATM-based backhaul, the maximum number of 8K traffic channel per E1 span is 180 while the number of 13K traffic channels per E1 span is 125. These are typical CDMA2000 numbers that may differ from supplier to supplier.

The RBS traffic load can be acquired from RF planning in terms of Erlangs. Then, the traffic is converted to the number of channels by Erlang B formula. Based on above maximum number of channels per E1 span, the number of E1 spans of RBS-BSC can be derived accordingly. For example, under mobility environment, 1 carrier and 3-sector RBS may require 1T1 or 1E1 backhaul to the BSC.

The interface of BSC-BSC is to support inter-BSC soft handoff which carries packet-based voice traffic channels. The inter-BSC handoff traffic Erlangs can be determined by the formula below:

$$\text{Handoff traffic (in Erlangs)} = \alpha \times \lambda \times T$$

where

λ = the handoff arrival rate at border cell, which is the function of mobile speed, border cell radius, and the number of calls at border cell (air capacity);

T = the mean usage duration, which is the average handoff call duration;

α = the ratio of the rate of system border crossings and the rate of cell border crossing.

Based on the above equation, we are able to calculate handoff traffic in Erlangs. Again, the number of E1/T1 spans is calculated in the same way as the backhaul interface (RBS-BSC).

The interface of BSC-MSC is based on the IOS[1] standard to support telephony service. The trunk group size depends on the voice and circuit-switched data traffic load on the BSC. Basically, the BSC traffic Erlang can be calculated from Erlang/sub and the amount of subscribers within the BSC. Then, it is converted to the number of trunks and E1s/T1s by Erlang B formula. Usually, trunk size is dimensioned under the load of 70% to protect the system from overloading.

4.3.6 2G and 3G Coexistence

The ATM Forum specification for circuit emulation service (CES) defines the means for ATM-based networks to employ AAL1 to emulate, or simulate, synchronous TDM circuits over the asynchronous infrastructure of ATM networks [2]. The circuit emulation (CE) function enables existing TDM circuits to be mapped over ATM. CE thus enables operators to migrate an existing TDM network to ATM while preserving the investment in TDM equipment. The reader should note that CE products are available for all major circuits, for example the American T1 standard and the

1. International Organization for Standarization.

European E1 circuit standard. CE is broken into two versions, structured and unstructured, but both versions of CE use AAL1 CBR connections.

In s*tructured CE,* the ATM network recognizes the internal structure of the circuit and is able to recover this structure at the receiving end. Circuit structure refers to the timeslots. Structured T1/E1 supports N × 64 Kbps (fractional T1/E1). This means that particular timeslots may be mapped to different virtual circuits and hence to different destinations. Several timeslots from a source circuit may be mapped to one virtual circuit. E1 contains 32 timeslots per frame, and the first timeslot, timeslot 0, is used for framing. As framing is irrelevant within the ATM network this timeslot (timeslot 0) is often terminated within the first ATM switch and regenerated at the destination ATM switch, which produces an E1 output. Additionally, some means of recovering the original signal clocking must be available. Which method is to be used at the destination is communicated across the network in a CE call setup request, or is set by the administrator. All timeslot 1s from each frame are mapped to an ATM cell and, hence, to an ATM VC. In mapping the source information into cells in this manner, an important issue is encountered, that of latency. An AAL1 1-byte header will be included in each cell payload; thus, to use the ATM bandwidth as efficiently as possible we should wait for 47 timeslot 1s. In waiting this long to fill the ATM cell we may reach a point where the QoS of the application is compromised.

To get around the latency issue we may choose to pad out the cell, which is to send a cell only partially filled with voice samples. The level of padding will depend on several factors, including how far the call has to go overall. Different strategies may be employed for different types of calls. For example, a call that is recognized as a long-distance or international call may be heavily padded by the CE function in order to minimize latency. That call will inherently have a long end-to-end delay. Similarly, it may be acceptable to map local calls to ATM cells without any padding to maximize bandwidth usage. Structured CE also specifies support CAS, commonly used by PBXs to indicate off-hook and on-hook conditions. Since the structured mapping of the individual timeslots does not convey TDM framing information end-to-end, the CAS information is encoded and transported separately, requiring additional overhead.

In *unstructured CE,* the network does not attempt to recognize the internal circuit structure. Rather, it simply transmits the entire circuit across the network. This unstructured service emulates a 1.5- or 2-Mbps data leased line. On E1 circuit, a 376-bit (or 47-byte) chunk of the source signal is taken. These 47 bytes have the AAL1 one-byte header added to make up the full 48-byte payload. An ATM cell header is added, making a full ATM cell.

These cells are then sent on a CBR connection. CE IWF provides timing to the TDM equipment in a synchronous mode, or accepts timing in an asynchronous mode. Timing transfer is critical for many legacy TDM networks, especially T1/E1 multiplexer networks.

CE is most likely the method that will be used to combine existing TDMA networks with the new ATM-based 3G networks.

On the other hand, existing transmission networks are based on PDH as defined in Recommendation G.702. ATM is considered the suitable technique to support B-ISDN. The SDH will form the basis of transport of the ATM cells, but during the transition period, there is the need to transport ATM cells using existing PDH transmission networks. Recommendation ITU-T G.804 provides the mapping to be used for this transport of ATM cells on the different PDH bit rates for both 1.544- and 2.048-Mbps hierarchies [13]. These mappings cover both the 1.544-Mbps and 2.048-Mbps-based hierarchies and are used in conjunction with the frame structures defined in Recommendation G.832.

The detailed requirements on how to map ATM on a fractional physical link will be in accordance with ATM Forum Document af-phy-0130.000, "ATM on Fractional E1/T1" (September 1999).

4.3.7 Transmission-Network Architecture

4.3.7.1 Transmission-Network Objectives

The main objectives of the transmission network are to connect all the points of interest, satisfy the capacity demands and provide reliable service using microwave, copper, fiber-optics or satellites. During the wireless-network build-out, it is important to establish a transmission plan that will include all the present requirements as well as future expansion (number of cell sites, RBS type, and future capacity requirements). Transmission network design typically involves a trade-off between network reliability and speed of deployment and price. An example of the small, three cell-site, RF network where mixed backhaul (transmission) media are used—leased T1/E1 lines and microwave—is shown in Figure 4.1. In wireless networks, the term backhaul (and sometimes access transmission or access transport network) is used to describe RBS-BCS connectivity exclusively. The terms *core transport* and *core transmission* are usually used to describe network connectivity between other network nodes. The core transmission network is the connection between both MSCs and BSCs and MSCs and the PSTN.

BSC-MSC connectivity refers to the ability of the BSC to support the reliable transfer of signaling messages with the MSC for call (e.g., voice, fax,

and packet data) setup and teardown, mobility management, radio resource management, and transmission facilities (terrestrial circuit) management. BSC-BSC connectivity involves the ability of the source BSC to support signaling messages with the target BSC for direct RBS to RBS soft or softer handoff, access handoff, access probe handoff, and channel assignment into soft or softer handoff in CDMA networks. The supported signaling protocol will allow for efficient resource allocation and deallocation (inter-RBS connection setup and teardown) and call connection control between a source and a target RBS during soft handoff.

Various types of transmission-network topologies are shown in Figure 4.12. Star and tree formats are examples of linear transmission-network architecture and used for small- to medium-size wireless networks. The size of the network is assessed based on the number of cell sites and the required backhaul capacity. The small wireless network shown in Figure 4.1 is an example of this type of transmission-network architecture. There is no network protection in this case and a problem on any of the E1/T1 links will affect one or more (in case of daisy-chain sites) cell sites. Added service reliability can be achieved with automatic rerouting. Many successful mobile operators protect transmission by using automatic traffic rerouting, assuring additional reliability in normal situations, such as when access microwave radio links suffer cutoff due to poor weather conditions or possible fiber-

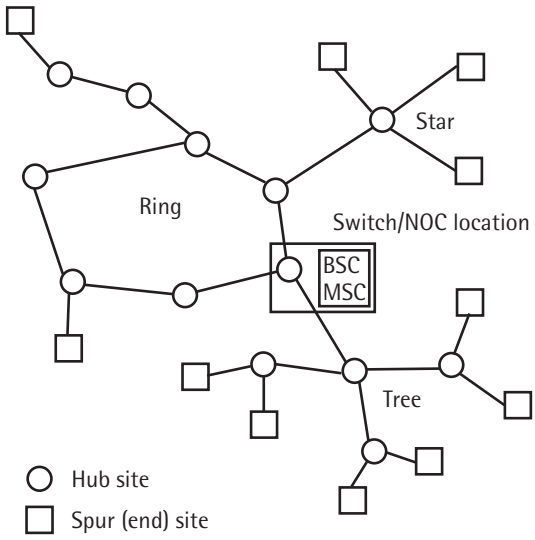

Figure 4.12 Transmission network topology.

optic cable cuts or any other human error. With a flexible rerouting transmission system such as T1/E1 trunk rerouting, backup capacity can pass via physically separate routes, as the problem is not likely to interrupt both routes simultaneously.

The base station trunk is the entire physical transmission link between two base stations or sites or between a base station and its base station controller, typically T1/E1 or nxT1/E1 links. In case of traffic failure, trunk rerouting switches all traffic in the main trunk simultaneously to the backup trunk. Large base stations comprising a number of circuits are switched simultaneously for minimum service downtime.

Rerouting can be arranged for all sites or only critical sites, such as base stations that are labeled as higher priority—for example, hub sites. Hub sites are those sites that collect traffic from more than one other site (typically three to four other sites) and carry that traffic toward the BCS location or fiber-optic ring hub site.

For a larger transmission-network it is recommended that a ring configuration be used as a high-capacity backbone carrying traffic to the switch location. Additional fibers in the fiber-optic network or cross-polarization in the case of a microwave network can be used to further increase (double) the capacity of the ring. The ring configuration shown in Figure 4.13 has the BSC and MSC not colocated but interconnected through two SDH high-capacity networks. It is usually recommended that the BSC and MSC be colocated and placed close to the point of presence of the PSTN to simplify interconnect. Ring architecture is considered a reliable communication facility since it provides automatic protection from the following:

- Site hardware (batteries, towers, antenna system) failures;
- Radio or MUX equipment failures;
- Propagation failures in the microwave network;
- Cable cuts in the fiber-optic network.

A ring topology also provides basic user features such as simple operation, fault location, and maintenance, and it provides alternate routing of E1/T1 traffic automatically and no loss of E1/T1 traffic due to signal failure. Each E1/T1 circuit must be dedicated completely around the ring, and reuse of the same E1/T1 in the opposite direction is not possible. For ultimate reliability, both directions can be 1+1 hardware protected. In microwave systems

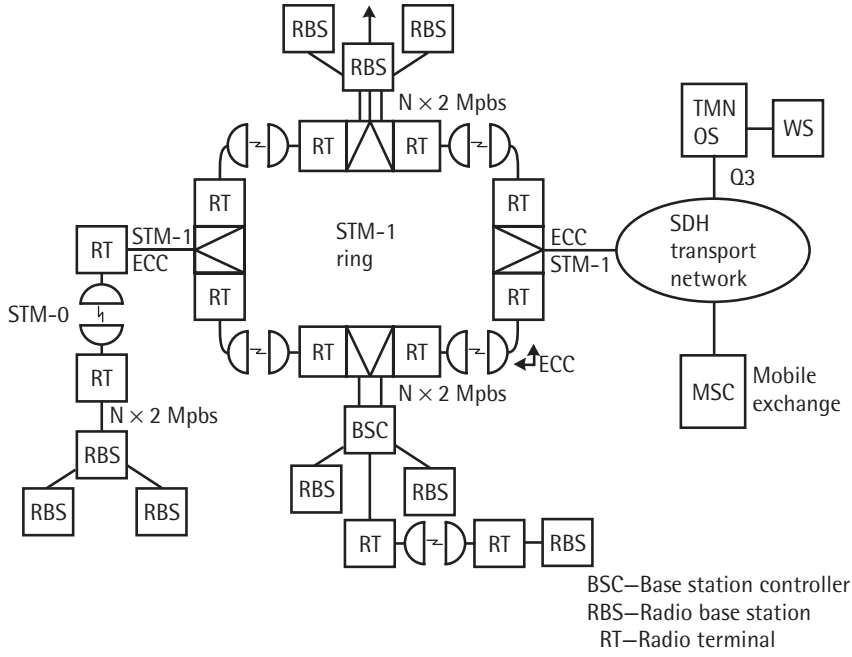

Figure 4.13 Microwave ring topology.

additional protection (e.g., space or frequency diversity) at lower frequencies may be required against short-term multipath outages.

All the sites that belong to the ring are considered hub sites and have to be planned so that during the deployment stage they are completed first in order to provide connectivity and protection for the rest of the network (spur links). In PDH networks, additional hardware with built-in intelligence to assess the T1/E1 quality and switch circuits, if needed, will be required. This hardware has to be added at every site and it is useful for small networks.

SONET/SDH have incorporated several protection and switching techniques from their inception. These include linear APS, path-switched rings, line-switched rings, and virtual rings. These techniques provide the ability for a network to detect the problem (under 10 ms) and heal itself automatically in the case of failure with a restoration time under 50 ms. Self-healing schemes use fully duplicated transmission systems and capacity for alternate routing of today's TDM or STM circuit facilities. The restoration capacity and the associated transmission systems are essentially unused, except in the rare occasions of network failure.

Although expensive and relatively complex to implement, the dual-homed ring architecture is the choice for high-capacity digital service providers. This architecture uses a drop-and-continue feature that ensures that traffic is available to pass between adjacent rings at two separate nodes or offices. If an entire office is lost, the receiving ring equipment will select traffic from the other office or node. Although this architecture looks expensive, due to network survivability it offers a high potential for cost reduction in the long run.

ATM features dynamic bandwidth allocation and ATM switches which together with SONET/SDH transmission will provide support for the emerging broadband multimedia services and existing legacy low-bandwidth telephony and data services. ATM can provide fault detection and traffic rerouting much faster than the existing 50-ms switching time requirement in the SONET/SDH networks. ATM-based capacity management and dynamic reconfiguration have the potential to significantly reduce the transmission facilities required for network survivability, providing economic justification for ATM deployments in large networks.

4.3.7.2 Cluster Topology

The cluster topology is a time-saving, cost-efficient, flexible, reliable, and future-proof way of building and implementing the transmission network for wireless systems. This topology is applicable to large wireless networks. In this context, the transmission network refers to the access network connections from the BSC to the RBSs.

The basic idea is to group the RBSs into several independent subnetworks, or clusters. Each cluster has a separate connection to the BSC or to an intercity transfer point. Grouping the RBSs into several clusters creates the initial cluster topology network. The number of clusters and the number of RBSs in each cluster are dependent upon the total number of RBSs and their configuration. Some of the more important factors governing the design are listed below:

- The total number of timeslots generated by the RBSs in a cluster must not exceed the capacity of the link connecting the cluster to the BSC.
- The cluster size must support a flexible, uncomplicated, and efficient network topology inside the cluster. A too large cluster will result in a large and inflexible topology.

- In a microwave link network, the number of clusters connected to one BSC must not be too high, as this will result in high concentration of microwave equipment at the BSC site, which may cause interference problems as well as tower stability problems (too many antennas).
- In general, a cluster size of 10 to 25 RBSs is recommended.

The RBSs in each cluster can be connected to a cross-connect node placed at a central hub-site in each cluster. The cross-connect node consolidates the traffic from the RBSs on the link to the BSC in order to minimize the required link capacity toward the BSC. Figure 4.14 gives an example of a cluster.

The topology inside the clusters can be of any type: cascade star, tree, ring, or a combination of these. Each cluster is then directly connected to a hub site in the central cluster, which is either the BSC site or an intercity (backbone) transfer point connecting to the BSC cluster in another city, as shown in Figure 4.15. The RBSs in this central cluster can be connected directly to the BSC or the intercity transfer point. They can also be connected through a cross-connect node to the BSC or the intercity transfer

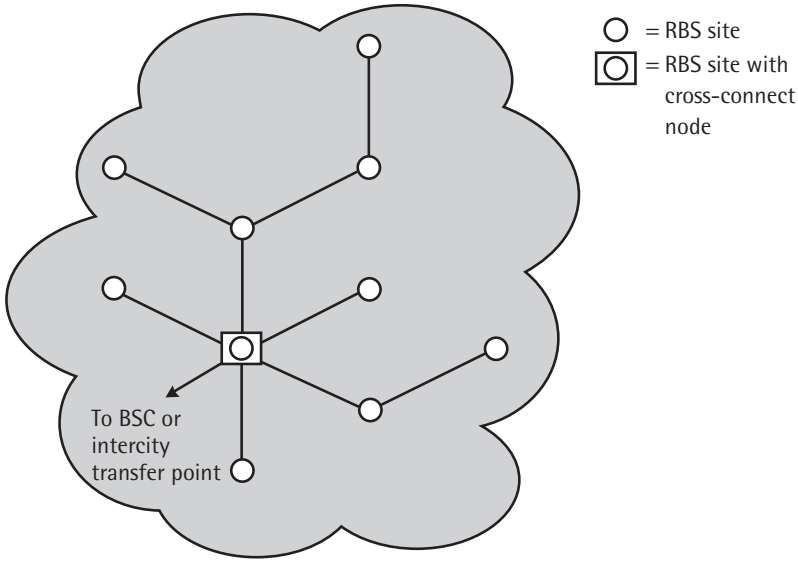

Figure 4.14 RBS cluster.

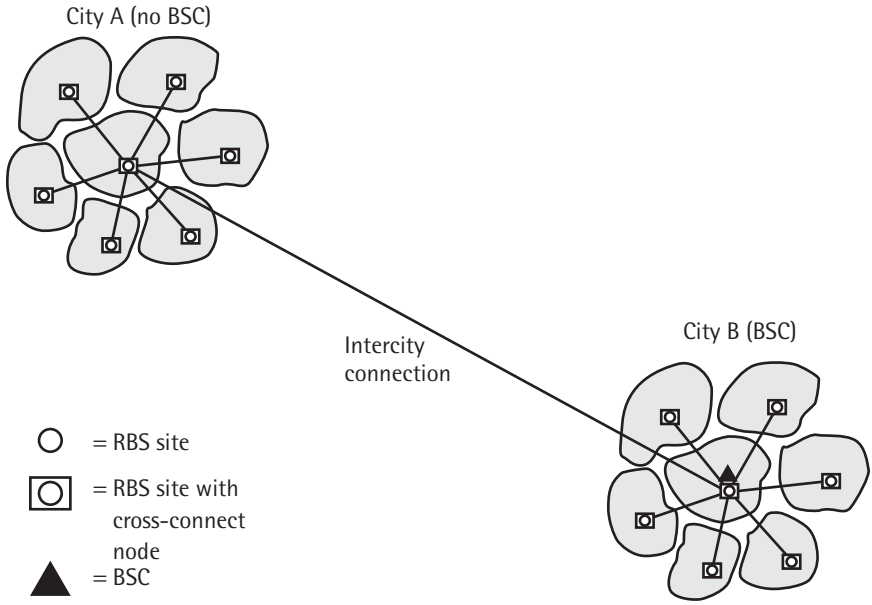

Figure 4.15 Intercity connection.

point. In a microwave link scenario, the links from the clusters toward the central cluster should be hardware protected as 1+1 to guarantee high availability. If required, considering the hop length the links should operate in the lower frequency bands to reduce fading due to rain. Operating in the bands below 10 GHz completely eliminates fading problems caused by rain.

In the cluster topology, it is both easy and cost efficient to create a high level of redundancy. Redundancy is achieved by connecting several clusters in ring structures (Figure 4.16). A ring is built simply by adding a link between the cross-connect nodes in two of the outer clusters. All links in a microwave ring should be nonprotected 1+0 links. However, in the initial network the links from the outer clusters to the central cluster were 1+1 protected. Therefore the standby radios from the 1+1 links can be used for the new connections between the outer clusters required to create the ring, but also for expansion in other parts of the network, thus minimizing investment cost for new equipment.

As shown in Figure 4.16, there are two possible paths from each cluster. Should the primary path be down due to a link failure, there is always a secondary path to the BSC. In a ring structure, all links in the ring must be able

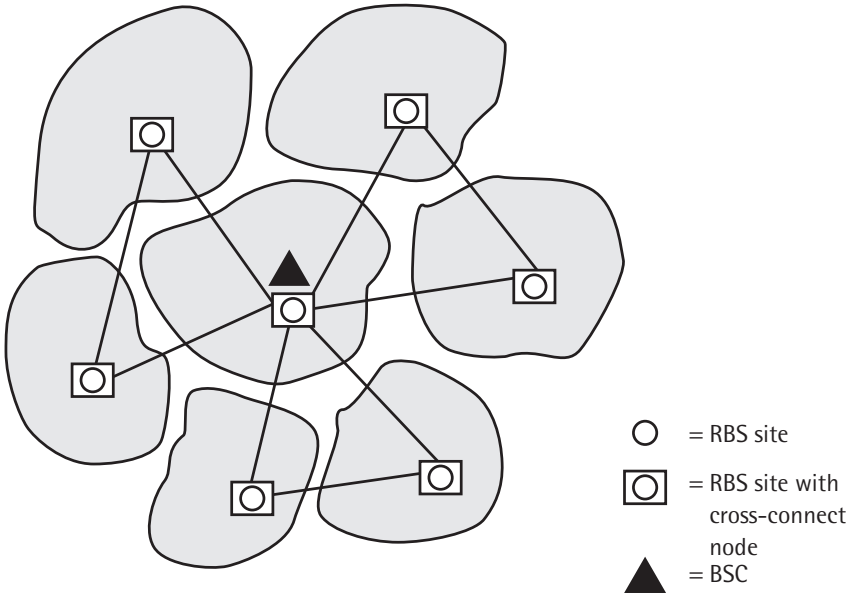

Figure 4.16 Redundancy configuration.

to transport traffic from the other cluster(s) in the ring. Therefore, the existing links may have to be upgraded to higher capacity, unless the links were dimensioned for a future ring structure already in the design of the initial network. More complex redundancy schemes can be constructed by connecting rings to each other to form a meshed network. These connections should, as before, be made between the cross-connect nodes in the clusters. The cross-connect nodes in each cluster are also excellent points of connection to a fiber-optic network.

The cluster topology is an ideal topology when introducing one or more remote BSCs. A remote BSC is a small and compact stand-alone node that can be introduced to add BSC capacity to high-density traffic areas. A remote BSC should, however, not be introduced initially but at a later stage when the traffic distribution becomes clear and when the need for such a node can be analyzed. The idea is to invest as the network grows. In the cluster topology, a remote BSC is introduced simply by replacing one of the cross-connect nodes with a remote BSC. Furthermore, by introducing a remote BSC in a cluster, the required capacity on the link from the cluster in question to the central cluster (MSC) is substantially reduced. The reduction

is a consequence of the trunking efficiency of the remote BSC. Thus, there is no need for large rearrangement of the transmission network when a remote BSC is introduced.

4.3.7.3 Quality, Performance, and Availability

In today's wireless networks, with converged voice and data, performance degradation may be as dangerous and costly as hardware failures. A degraded transmission network can result in unacceptable signal transmission quality, loss of information, and dropped connections. High availability does not mean just preventing catastrophic failures but also preventing quality and performance degradation.

BER, errored seconds (ESs), and one-way delay are usually parameters of interest that will define the quality of the transmission network. Some wireless technologies are more sensitive to delays in the T1/E1 links than others. For example, cdmaOne and CDMA2000 due to the soft handoff have much more stringent requirements on the network delays and synchronization than TDMA-based wireless technologies.

High availability of the wireless network is an end-to-end network goal—the network management system (NMS) can help identify critical resources, traffic patterns, and performance levels. NMS can also be used to configure error thresholds, set corporate policies, and provide reports showing end-to-end results. Transmission-network survivability is usually measured by its long-term availability or average network uptime. Most operators expect their network to be continuously available (or at least as little downtime as possible) to minimize potential loss of revenue. The survivable wireless network has an infrastructure of transmission facilities and reliable network elements that are used to manage them. High network availability at the transport level may be achieved using millisecond restoration schemes provided by self-healing network configurations such as SONET/SDH rings or fast facility protection (FFP). DACS in combination with SONET/SDH ring configurations will ensure network availability and survivability. An FFP network comprises two physically diverse routes with identical transmission systems (route diversity). Each route carries half of the working traffic and half of the restoration traffic. The restoration traffic on each route is the duplicate of working traffic on the other route. If the media on these routes are different (i.e., one is fiber optic and the other is, for example, microwave), we talk about the media diversity. These highly reliable solutions don't come cheap, and in many cases a compromise must be made between the cost of the network and its deployment time and network reliability.

Regardless of the transmission-network medium and topology, hardware redundancy is an option when designing the transmission network. Protection types usually used are 1+1, where one card or module serves as a protection for another one, or 1+*N,* where one card or module protects *N* other units. Linear 1+1 protection switching means that identical payloads will be transmitted on the working and protect fibers or working and protect frequencies in case of the microwave system. Liner 1+*N* protection switching assumes that there is one protect fiber or frequency for *N* working fibers or frequencies. A rule of thumb is that if all the hardware is protected with 1+1 and 1+*N* configuration, fewer spare parts are needed. In case of hardware failure, protection will kick in and the operator will have sufficient time to order replacement parts from the supplier. A ring configuration could provide protection against hardware failures as well, so additional hardware protection might not be required. This is something that transmission engineers must decide, and it is a decision that will be based not only on technical but also budgetary requirements.

References

[1] Bos, L. and S. Leroy, "Toward an All-IP-Based UMTS System Architecture," *IEEE Network*, January/February 2001, pp. 36–45.

[2] McDysan D., and D. Spohn, *ATM Theory and Applications*, New York: McGraw Hill, 1998.

[3] ETSI, UMTS-QoS Concept and Architecture, ETSI TS 123 107 V4.0.0, December 2000.

[4] Malis, A.G., "Reconstructing Transmission Networks Using ATM and DWDM," *IEEE Communications Magazine,* Vol 37, No. 6, June 1999, pp. 140–145.

[5] Liu, C. et al, "Packing Density of Voice Trunking Using AAL2," Globecom '99 General Conference, 1999.

[6] Eneroth, G., et al, "Applying ATM/AAL2 as a Switching Technology in 3G Mobile Access Networks," *IEEE Communications Magazine*, June 1999, Vol 37, No. 6, pp. 112–122.

[7] McDaysan, D., and D. Spohn, *ATM Theory and Applications*, New York: McGraw-Hill, 1998.

[8] ATM Forum, "ATM Trunking Using AAL2 for Narrowband Services," AF-VTOA-0113.00, February 2000.

[9] Roberts, Jim W., "Traffic Theory and the Internet," *IEEE Communications Magazine,* January 2001, pp. 94–99.

[10] Hac, A., *Multimedia Applications Support for Wireless ATM Networks*, Upper Saddle River, NJ: Prentice Hall, 2000.

[11] Zorzi, M., R.R. Rao, "On the Impact of Burst Errors on Wireless ATM," *IEEEE Personal Communications,* Vol.6, No. 4, August 1999, pp. 65–76.

[12] Lankl, B., and M. Salerno, "ATM Traffic and Its Impact on Radio System Design," Sixth European Conference on Fixed Radio Systems and Networks, Bergen, Norway, June 1998.

[13] ITU-T G.804, ATM Cell Mapping into Plesiochronous Digital Hierarchy (PDH), February 1998.