

White Paper: Core Network Design and Dimensioning for GPRS and UMTS

By Moe Rahnema

Senior Principal Consultant, LCC International

ABSTRACT

This paper discusses the protocol architecture, and the design methodology for GPRS and UMTS core networks. The topics discussed include the new elements used for packet routing and external network interfacing, an overview of the core network protocol architecture, transport technology and interconnect topology options and their pros and cons, as well as the dimensioning parameters for network elements and links.

I. INTRODUCTION

The core network in GPRS (General Packet Radio Services) referred to as the 2.5G wireless packet data networks will provide interoperability between the radio access network (RANs) and the external networks[1]. It will also provide and support the feature servers operators will deploy for increasing the number and variety of subscriber services. The core network extends internet protocol based packet data services to the radio access infrastructure and hence is capable of routing IP packets within the context of mobility. The basic GPRS core network infrastructure and elements will also be used in the release 99 specifications for the 3G UMTS core networks which use the European wideband CDMA air interface standard. Therefore a discussion of the GPRS core network should cover the basic core network infrastructure and function in the 3G UMTS networks, as well.

2. Core Network Architecture

2.1 GPRS

The GPRS core network implements IP based packet switching capability into the existing GSM core network infrastructure to achieve interconnectivity with packet data networks and the internet. This data overlay network provides packet data transport at rates from 9.6 to 171 kbps. Additionally, multiple users can share the same air-interface resources.

GPRS attempts to reuse the existing GSM network elements[4] as much as possible, but in order to effectively build a packet-based mobile cellular network, some new network elements, interfaces, and protocols that handle packet traffic are required. Therefore, GPRS requires modifications to numerous network elements to include the core network elements such as the HLR data bases. The GPRS system architecture and the various network elements and interfaces are shown in figure 1[6-7]. The main Core Network elements in GPRS consist of the Serving GPRS Support Node (SGSN), Gateway GPRS Support Node (GGSN), and upgraded location registers. The SGSNs, and the GGSNs may be connected either directly and/or through intermediate routers and switches to form what is referred to as the packet switched core network. The core network is used as the interface between the radio access network and the external data networks such as PDNs, and the internet. The point of contact to the external networks is realized through the GGSN, using the Gi interface. SGSN interfaces with the radio

access network (base station sub-system, for instance) through the Gb interface and provides mobility management and call signaling functions. The SGSN maintains signaling connections with the HLR, and MSC/VLR through the Gr, and Gs interfaces, respectively. The GGSN maintains signaling connections with the HLR through the Gc interface. The interconnections between the SGSNs, and the SGSN and GGSN are implemented through the Gn, and the Gp interfaces, respectively.

Figure 1 GPRS System Architecture

2.1.1 The New Network Elements

The packet domain Core Network functionality is logically implemented on the two network nodes, the Serving GPRS Support Node and the Gateway GPRS Support Node. The GPRS support nodes originating from GPRS evolve to UMTS network nodes, in the release 99 specifications. Therefore, the name GPRS Support Node is used even if the node may provide UMTS functionality only.

The SGSN represents the mobile's point of attachment to the core network and provides the following specific functions for the data services:

- Handles call control signaling with data services location registers
- Provides mobility management such as tracking of mobile's routing area and serving cell.
- User authentication and verification
- Billing data collection
- Handling of the actual user's traffic and conversion between the IP core and radio network
- Standard interfaces to the HLR for management of end user subscriber data

Each SGSN in the network serves mobiles in a limited area, referred to as the SGSN area. An SGSN area is thus the part of the radio access network served by an SGSN, and may consist of one or several routing areas. An SGSN area may also consist of one or several BSC areas. There is not necessarily a one to one relationship between SGSN area and MSC/VLR area.

GGSNs are used as interfaces to external IP networks such as the public Internet, other mobile service providers' GPRS services, X.25 networks, and enterprise intranets. GGSN contains routing information for attached GPRS users. The routing information is used to tunnel the protocol data units (PDUs) to the MS's current point of attachment, i.e., the SGSN node. The GGSN may request location information from the HLR via the optional Gc interface.

Other functions include network and subscriber screening and address mapping. One (or more) GGSNs may be provided to support multiple SGSNs. The GGSN can deliver to the MS IP-based data services that are consistent with the look and feel of the Internet or other IP-based networks the subscriber may be accustomed to accessing via a wireline environment. As regards the external IP networks, GGSN is viewed as performing common IP router functions. This may include firewall and packet filtering functions.

A GGSN may serve more than one SGSN nodes, just as an SGSN may be served by more than one GGSN for redundancy purposes or for connecting to multiple ISPs, and external data networks.

In addition to the two new core network elements discussed in the above, GPRS will require an upgrade of the GSM HLR software to incorporate the packet switched user service profile.

2.1.2 The protocol Architecture

The core network in GPRS and UMTS use the internet protocol, IP, as the protocol in the network layer. Options for the underlying data link layer to transport the IP protocol data units between the core network elements include ATM, or protocols running on private lines such as PPP, and HDLC. The protocols used within the transport layer are UDP for IP services, and TCP for services which require delivery guarantee such as X.25 services.

Data transfer between two GPRS support nodes (GSN) such as a SGSN and a GGSN, or between two SGSNs takes place through a tunneling protocol referred to as the GPRS Tunneling Protocol (GTP). The GTP encapsulates and carries the protocol data units (PDU) that need to be transported between two GSNs. GTP uses a Tunneling ID (TID) within its header to indicate which tunnel a particular PDU belongs to. In this manner, packets from different mobiles are multiplexed and de-multiplexed by GTP between a given GSN pair. The UDP/TCP destination port reserved for GTP is 3386.

The GTP tunnel ID value is established by a procedure which is referred to as the Packet Data protocol (PDP) context establishment which takes place on the signaling plane. PDP context activation specifies a two way GTP tunnel between the SGSN and the GGSN used to carry encapsulated user data packets between a mobile and the external packet data network. The PDP context activation process also leads to the establishment of an SNDCP (Sub-Network Data Convergence Protocol) tunnel between the mobile station and the SGSN serving the mobile. The SNDCP is used to transfer the PDP data between the SGSN and the mobile station.

The PDP context specifies two important parameters which are defined in the following:

- The PDP Type: Defines the end user protocol used between the external packet data network and the mobile station (MS). The PDP type is divided into an organization and a number field. The organization field specifies the organization responsible for the PDP Type number field, and the PDP address format. For X.25, for instance, the organization is ETSI, and the PDP Type number is 0. The PDP address is the X.121 format.
- The PDP Address: This is the address that the PDP context of the mobile station is identified with from the external packet data network. This address is assigned dynamically to the MS at PDP context activation by the GGSN for that service. In case the MS is assigned a permanent IP address at subscription time, the address is activated by the PDP context establishment.

The PDP context activation will also specify the quality of service profile information elements which represent the QOS values negotiated between the mobile station and the SGSN.

The protocol architecture for GPRS is shown in figure 2.

Figure 2-GPRS protocol architecture

2.2 Core Network in UMTS

The release 99 specifications for the 3G networks of UMTS which are based on the European wideband CDMA air interface standard are based on the same core network infrastructure used in GPRS. Therefore, from the core network point of view, there is basically no functional differences between GPRS, and the UMTS 3G networks, other than the fact that the 3G core networks will have to be sized to accommodate higher traffic handling capacity, and might use different algorithms for authentication, and/or ciphering.

3. Core Network Transport Technologies

The transport technology options which may be used to provide the connectivity between the core network nodes (such as the SGSNs, GGSNs, etc) at the physical and data link layers consist of the following:

- Dedicated Private Lines
- ATM virtual circuits
- Frame relayed PVCs
- VPNs
- MPLS

3.1 Dedicated Private Lines

Dedicated private lines include the T and their European version E carriers and their fractionals, Sonnet based optical links, and possibly line of site digital microwave links.

3.1.1 Advantages and Disadvantages of Private Lines

Advantages

- QOS and congestion not an issue (only within routers)
- Easier to engineer (size to handle peak traffic)
- Provided in more areas

Disadvantages:

- Reduced reliability (usually no alternate-rerouting)
- Costs more and costs are usually distance sensitive
- Inefficiencies involved in sizing to handle peak traffic flow (not efficient for sporadic bursty data)

3.1.2 Sizing Criteria for Private Lines

The bandwidth (size) of a private line between any two nodes should be based on the expected peak traffic between the two nodes. This means any excess bandwidth not used at times of light traffic is simply wasted.

3.2 ATM Virtual Circuits

ATM virtual circuits can be provided on either a permanently allocated or switched basis through an ATM network. The network providing the ATM service can be either a carrier provided public network, or a network of privately owned equipment serving a single wireless operator[5]

3.2.1 ATM advantages and disadvantages (against leased lines)

Advantages:

- More cost effective, can provide bandwidth on demand
- Costs are not distance sensitive in most cases
- Enhanced reliability through built in-rerouting capabilities
- Allows multiple virtual connections through a single physical port
- Dynamic bandwidth sharing
- Multi-class services to efficiently handle each traffic class and required performance.

Disadvantages:

- More complex engineering and configuration to handle QOS
- Involves some protocol overhead
- Can incur further packet losses and delays within the ATM switches

3.2.2 Sizing parameters and issues

ATM provides virtual circuits with different bandwidth and delay characteristics to suite different classes of services. Each virtual circuit class involves parameters which need to be specified according to the bandwidth characteristics of the traffic class to be handled. ATM offers the following virtual circuits:

UBR (unspecified bit rate)-This is for handling best effort delivery service suited mostly for handling non-real time data services such as email, Usenet, file transfer, LAN internetworking, and etc, and does not guarantee their delivery (a higher layer protocol such as TCP would be needed to ensure end-to-end delivery of the service). UBR circuits do not guarantee any amount of bandwidth, and are normally policed by the network not to exceed a specified maximum bit rate.

ABR (available bit rate)---This is similar to UBR except that the network does guarantee a specified minimum bit rate for the service, and hence achieves better performance in terms of a more stable throughput. ABR circuits are more suited to handling Web Browsing and transaction services which are more delay constraint than other data services.

CBR (constant bit rate)---The CBR virtual circuits are used to provide circuit emulation services for handling highly delay critical traffic such as real time un-coded voice and video services. The CBR can also be used to emulate private line services through an ATM network. There is only one parameter required to specify a CBR circuit and that is the peak bit rate for the service.

VBR-rt (variable bit rate, real time)---is specified by three parameters consisting of the expected peak cell rate, sustained cell rate, and the maximum burst size. The network tries to minimize the end-to-end delay as well as the delay variations from cell to cell. The VBR-rt virtual circuits are suited to handling compressed voice and video services which are delay critical.

VBR-nrt (variable bit rate, not real time)---is specified by same parameters as the VBR-rt, but is suited to transporting bursty data services which are less delay critical such as banking transactions, airline reservations, process monitoring, and frame relay interworking.

3.3 Frame Relay

Frame relaying services are currently offered through PVCs (permanent virtual circuits), which are set up in advance by the network operator. All of the current frame relay providers offer fixed rate services of 56 Kbps. Many offer higher rates of 128 Kbps, 384 Kbps, and up to 1.536 Mbps. Some of the carriers may offer a 64 Kbps service in which case the pricing will be the same as the 56 kbps circuits.

PVCs can be either simplex or duplex: Simplex allows only one-way traffic; two simplex circuits can be set up with different throughput. Duplex circuits allow bidirectional transmissions; they use the same information rate in each direction.

3.3.1 Frame Relay Advantages and Disadvantages (compared to ATM)

Advantages:

- Uses variable frame lengths and hence incurs less header overhead for handling certain data services that involve lengthy transmissions
- Is often available in more regions than ATM, at the present time
- Easier to engineer and configure in compared to ATM
- Frame relay equipment and interface cards are less expensive in compared to ATM equipments
- Like ATM, offers distance insensitive pricing for long distance services
- Like ATM, allows multiple virtual connections through a single physical port

Disadvantages:

- Runs at slower speeds compared to ATM, currently is offered at maximum speed of T1; though some vendors are developing or may already offer equipment to run at multiple T1 and up to DS3 rate.
- Basically treats all traffic the same and no extensive service quality features are offered
- Not best in handling delay critical traffic such as conversational voice

Frame relay's relatively low overhead and variable packet size makes it an efficient protocol for transmitting data[5]. With frame relay, the default overhead on a 128 byte frame is approximately 4 percent of the frame. As the frame size increases, the overhead percentage decreases dramatically. For example, with a 1600 byte frame, a size common with LAN protocols, the overhead drops to less than one-third of a percent.

ATM's 53 byte fixed cell is optimized for handling multimedia traffic at high speeds. However, with ATM, approximately 10 percent of each full cell is overhead. In very short transactions, where cells are not full, an even greater percentage of bandwidth carries no user data.

As a general rule, the lower the speed (e.g., T1 and sub T1), the more should be the concern with overhead for data applications. At levels of T3 and above, the benefits of ATM outweigh the efficiency considerations. At relatively lower speeds (T1 and below), frame relay more efficiently transports data. However, ATM has been recommended as the transport protocol for the release 99 of the UMTS core network[2-3].

3.3.2 Sizing Parameters and Issues

Frame relay allows for a fixed or for a variable rate information. The customer asks for a specific committed information rate (CIR) which the service provider guarantees within a certain probability. CIR is a parameter which specifies the maximum average rate at which the network may transmit without discarding data. This rate is averaged over a specified interval T_c . The user can transmit bursts of data at higher rates, but only at the risk that the extra information will be the first dropped should congestion occur in the network (by setting the Discard Eligibility, DE, bit). The parameters involved here are the Committed Burst Size (B_c), and Excess Burst size (B_e). Committed Burst Size (B_c) is the largest number of consecutive bits that the network agrees to carry within a specified time interval T_c without discarding data. The Excess Burst Size (B_e) parameter specifies the excess over the Committed Burst Size, which the network agrees to carry (within the time interval T_c) with a greater likelihood that some data will be discarded.

Some frame relay carriers now offer at a greatly reduced price a 0 CIR service in which no data rate is guaranteed to reach the destination--in effect all data is considered to be a burst. In a network not experiencing congestion, 0 CIR provides the same level of service provided by higher CIRs since no packets are actually dropped, and even in the face of congestion a large portion of burst packets will make it through the network. It is not yet clear if 0 CIR service will remain a practical alternative as more users begin to use frame relay and the frame relay networks become more utilized.

3.4 VPNs

Virtual Private Networks (VPNs) are IP tunnels which are set up through either the internet or a private IP network to transport the TCP or UDP protocol data units between the SGSN and GGSN in GPRS or UMTS. However, this approach has the drawbacks of increased tunneling overheads, as well as potential performance bottlenecks. IP is a connectionless technology, that has no absolute QoS mechanisms. The only QOS mechanisms offered by IP are packet prioritization and scheduling mechanisms which helps to differentiate among multiple services and provide service quality on a relative basis (no absolute guarantees of delays and throughputs). This relative QOS mechanisms is not currently implemented in the internet, but can be implemented in a private IP network to provide the interconnect between the core network elements.

3.5 MPLS

Multiprotocol label switching (MPLS) refers to a hybrid routing and switching technique in which the the IP routing capability is combined with the fast switching aspect of ATM. The IP routing function is incorporated within edge devices which use the IP header information to

determine the routing for the packets. The packets are then tagged with a swappable label designating the route of the packets within the intermediate nodes which are ATM like switches performing the switching function on the packets label. Packets can be labeled according to the packets source/destination TCP ports and/or source-destination IP addresses as well as other parameters such as packets priority and QOS related parameters.

The MPLS approach can be classified as a flow-based resource reservation mechanism where packets are classified and scheduled according to their flow affiliation for resource reservation. Resource reservation has the potential for end-to-end QoS due to the concept of flows; a flow originating from a defined end point and destined for another. Thus, MPLS provides the QoS benefits of the connection oriented data transfer while maintaining the benefits of IP interface and routing functionality at the edge of the network. MPLS is considered to be the replacement for IP over ATM in the long run as the traffic engineering and the resource reservation features of the technology is fully developed and implemented.

4. Transport technology Selection for Core Network

A primary issue in the selection of a proper transport technology for the core network is how to manage quality of service (QoS) for packet-switched traffic. Unlike QoS objectives for circuit-switched traffic in mobile wireless networks, which is typically measured in terms of the blocking rate, QoS for packet-switched traffic is primarily measured in terms of Delay and Throughput. This will affect the choice of transport technology selected for the packet-switched core network, whether dedicated lines, ATM, frame relay, etc..

Dedicated private lines of course offer the best QoS because no congestion is experienced on them. However, they may not always be the best choice because of the costs involved. In particular, private lines can end up to be quite inefficient when the traffic to be carried happens to be bursty resulting in an inefficient usage of the dedicated bandwidth. Another drawback of private lines is reduced reliability. In contrast, both frame relay and ATM not only offer the opportunity for bandwidth on demand and dynamic bandwidth sharing, they also offer enhanced reliability due to the inherent redundancy built into these networks which is shared by many users and hence is cost effective.

Since ATM is a connection-oriented packet-switching technology, which has some built-in QoS mechanisms. It offers the opportunity for integrating multi-service classes within the core network while allowing each one to meet its QoS requirements. The service quality in ATM networks is achieved by controlling delay, jitter, and loss, and dedicating bandwidth to selected traffic types. ATM switches implement proper queuing, scheduling, buffer management mechanisms, traffic shaping, and admission control techniques to provide virtual circuits with appropriate bandwidth and delay characteristics appropriate for each different traffic type. Frame relay on the other hand, offers very little capability for differentiating between diverse services to allow the proper sharing of the network capacity and resources and letting each service to meet its delay and throughput requirements.

As an alternative, VPNs may also be used to transport IP (IP within IP) within the core network. However, IP is a connectionless technology, that has no absolute QoS mechanisms. The only practical QoS mechanisms offered by IP at the moment are the so called differentiated services (DiffServ) which are based on packet prioritization and scheduling mechanisms within the

routers. The DiffServ mechanism helps to differentiate among multiple services and provide service quality on a relative basis (no absolute guarantees of delays and throughputs).

for real-time services, a connection-oriented technology such as ATM is usually better, primarily due to the lowered delay variation or jitter. Delay variation is the time difference between arrivals of subsequent packets, and is especially important for real-time services. For example, voice services cannot suffer more than a certain amount of delay variation. The statistical multiplexing of packets for different users can cause delay variation, or due to the different paths packets may take, causing a variation in the inter-arrival time between packets. The tradeoff between different technologies is in terms of bandwidth efficiency, delay, delay variation, and cost. Determining which technology is better under what circumstances, as well as which is better for long-term evolution is a primary challenge that operators will face.

5. The Design of Core Networks

The design of the core network involves several considerations and decisions that need to be made to ensure a design that is scaleable, reliable, cost effective, and meets the QoS objectives for delay-tolerant data traffic, as well as real-time traffic.

Following are the primary high-level tasks in designing the packet-switched portion of a backbone network:

- Decide on the number and locations of SGSNs, and GGSNs based on expected access traffic and distribution, client site preferences such as co-siting with GSM nodes, equipment traffic handling capacity, required external interfaces, and etc
- Estimate the end to end aggregate traffic flow demand for each service
- Determine the QoS requirements for each service
- Select appropriate transport technologies for interconnecting the nodes
- Decide on a network topology (interconnection between traffic nodes)
- Select the appropriate network infrastructure elements (routers / switches / other traffic nodes) and configure properly to meet QoS requirements
- Allocate IP addresses to various network elements, and users
- Estimate quantities of infrastructure equipment required to support the offered traffic
- Dimension the network elements and the interconnect

All these tasks have to be performed for desired network reliability, requiring considerations of traffic path redundancy and thus traffic node interconnectivity, as well as with cost considerations.

Two major decisions that need to be made in the design process, besides the selection of an appropriate transport technology, is a proper model for estimating the traffic demand for various services, and the selection of a topology choice for interconnecting the nodes. These are discussed in the following two sections.

5.1 Traffic Model

One of the most difficult data to obtain for designing a network initially is usually the information describing the traffic that will be using the network. Traffic data can be used to determine the placement of links, the sizing of links and routing configurations that obtain desired utilizations and service levels. Traffic data may be generated from actual measured traffic on a similar network or manufactured based on marketing estimates of the applications that will be using the network. In either case, a methodology for choosing the data set or sets against which to design a network must be determined. Generally, a conservative estimate such as peak-day, peak-hour traffic is assumed. When actual traffic information is unknown, it is recommended that several traffic scenarios be generated based upon whatever information is available in order to test the sensitivity of designs to changes in network traffic distributions. Once the network is designed and operated, node configuration data can be drawn from MIBs, configuration files, or network management systems for future capacity planning purposes. Such detailed configuration data collected from a running network can be used to accurately model routing in complex operational networks that use route-maps, route-filters, route-redistribution, etc.

The steps involved in choosing a traffic model, and using that to estimate the network link capacities are as follows:

- Estimate the end-to-end peak hour traffic matrix based on either a similar representative network that exists, or marketing estimate of the traffic the applications are expected to generate
- Select a network topology (discussed in next section)
- Select a traffic routing scheme
- Size up the links connecting the nodes

Subsequent steps for validating the initial design are:

- Fail single links, and nodes and using a network performance analysis tool, re-run the traffic through the network and make sure there are no over-utilized links. If so, resize any over-utilized links or add new links, and rerun the traffic to make sure the bottlenecks are eliminated.
- Vary the assumed traffic flow demand within expected ranges and patterns, and see if the design can still meet it. Otherwise modify the design by resizing links or adding new ones until the sensitivity to the traffic variation is resolved.
- Fail single links and nodes, and re-run the model with the new traffic flow demand and re-iterate the design as before if needed

5.2 The No Traffic Information Scenario

In this scenario, there is not information on the traffic flows in the network. An algorithm may be used to support analysis of the underlying topological design without any traffic information. Design algorithms that focus on addressing fundamental topological characteristics such as connectivity, min-hop distances, number of links, costs of links, and network hierarchy may be used to generate initial network designs. Technologies may be deployed on the model and tested for validity. Without traffic information, utilizations can not be computed so link sizing must be based upon user specified design criteria. This type of approach is appropriate for preliminary

investigations such as those conducted by pre-sales organizations or network planning groups that are exploring completely new networks designed from the ground-up. Designs based on "traffic insensitive" approaches are appropriate. These types of design algorithms provide a method for creating low-cost solutions that are relatively insensitive to changes in the distribution or types of traffic on the network. After examining some of the basic architectural issues at this level, the best designs will generally be used as the starting point for more detailed evaluations integrating some type of traffic information.

5.3 Locating core network nodes

In order to reduce interconnection costs, the network operator can deploy the GPRS support node (GSN) at a location most suitable for traffic concentration. The SGSN should be placed close to the MSCs, in order to concentrate traffic from an entire MSC area.

The GGSNs should be as close to the network as possible, perhaps resulting in dedicated GGSNs that incorporate specific protocols according to traffic types. Depending on traffic volumes, a single GSN could accommodate more than one protocol. At any rate, SGSN and GGSN interwork via a frame relay or ATM backbone, which is sufficiently powerful to handle the associated data streams.

5.4 Core Domain Dimensioning

The system constraints and operator planning assumptions impact the dimensioning of each GPRS Core network element differently. An overview of how these drivers apply to each core GPRS Network element is shown in Table 1.

GPRS Core Element	Key Dimensioning Factors
Gb interface	<ul style="list-style-type: none"> • Average throughput per user • Number of cells per BSC/PCU
SGSN	<ul style="list-style-type: none"> • BH number of Attached Subscribers • BH peak throughput • Number of cells per SGSN • Number of Routing Areas per SGSN • Number of Gb E1 ports
Gn interface	<ul style="list-style-type: none"> • BH number of Active PDP contexts • BH peak throughput • Total number of Routing Areas • Number of switch sites
GGSN	<ul style="list-style-type: none"> • BH number of Active PDP contexts • BH peak throughput • Number of Gi ports required

Gi interface	<ul style="list-style-type: none"> • APN throughput requirements • Customer security requirements • Customer cost consideration
--------------	--

Table1: Dimensioning Factors for GPRS Network Elements

5.5 IP Address Allocation

IP address allocation involves proper strategies for allocating IP addresses to internal network elements, and to subscribers.

The IP address assignment to internal network elements such as the SGSN, GGSNs, and application servers should be made in such a way to allow route aggregation, and proper subnetting for efficient routing and minimization of traffic overheads due to routing updates. User IP addresses can basically be assigned in three ways:

- Fixed: IP address stored in the HLR
- Dynamic: A set of IP addresses are allocated to the GGSN domain. A mobile station is assigned an address from this pool upon PDP context activation.
- Dynamic: The IP address can be allocated by an external ISP (Radius) server

6. Backbone Topology Design

The basic problem of backbone topology design for the core network is the following: Given a set of edge nodes (i.e, SGSNs, and GGSNs), and a set of economic costs for candidate links that could connect them directly or through intermediate transit nodes, select a set of N disjoint links, and transit nodes (if any) that result in a connected network that meets the traffic demand, redundancy and delay performance in a cost effective manner. There are a variety of ways in which the connectivity between the network edge nodes can be architected. These are discussed below.

6.1 Fully Meshed Architecture

In the fully meshed architecture, each edge node is connected to every other edge node in the network either physically, for instance through leased lines, or logically, for instance through frame relay or ATM virtual circuits. The fully meshed topology provides a highly redundant, and simple connectivity, but can end up being expensive particularly when the number of nodes is considerable. In this architecture, the sizing of each link in the network will be based on the traffic flow predicted between the two nodes. Extra capacity can be provided on each link to handle at least a portion of re-routed traffic between other neighboring node pairs when the link between them fails. The pros and cons of the fully meshed topology are:

- Enhanced reliability due to the reach connectivity between nodes
- Reduced transport delays due to no transit routing or switching

- Costly for large number of nodes
- Not well scaleable (addition of a new node would involve many new connections and configuration in the existing nodes)
- Increased number of peering routers, which causes too many routing update messages between nodes, and a slower convergence of routing updates
- For the case of using private lines (leased lines), requires too many physical ports on each router or switch for large networks

6.2 Partially Meshed Architecture

In the partially meshed architecture, direct link are provided only between certain nodes, usually the ones with heavy traffic demand between them. Then, the routing of traffic between nodes with no direct connection is provided through transit routing (or switching) through other nodes. The advantages of this over the fully meshed architecture are:

- Is more cost effective
- Is more scaleable
- Reduces the number of peering routers

A basic problem in designing the partially meshed topology is which node pairs to interconnect directly. Generally given N nodes, there is not just one optimum or even one workable scheme of interconnecting them through a partially meshed topology to support a given traffic flow pattern. However some topologies will provide better cost efficiencies, failure resilience, and improved transit time than others, though many will work. Finding a good topology is a combination of heuristics, science and experience. Most often, a topology which is designed to work with an initial traffic pattern projection is periodically enhanced and updated as the network is installed and measurements are collected on the actual traffic flow pattern in the network.

6.3 The Layered Architecture

In the layered architecture, intermediate nodes (transit nodes) are introduced in the network which are used to provide transit routing of traffic between the network edge nodes. In this scheme, the traffic routing between the edge nodes takes place either directly (if they are connected), or indirectly through either the intermediate nodes or possibly through other edge nodes. The intermediate nodes in the core network can be routers, or ATM like switches if the MPLS technology is used to implement the core network. In a way, the partially meshed topology which was discussed in the previous section is one form of the layered architecture in which the edge nodes themselves act as intermediate nodes for handling transit traffic. The design of the layered architecture follows the same rules and guidelines as in the case of the partially meshed topology (a special case of the layered architecture).

The benefits of the layered architecture are:

- Is cost effective (reduces the number of lines)
- Is scaleable (easy to add new nodes and sites)
- Makes small number of peering routers

- Requires small number of physical ports on routers (when private line are used)
- Can help to place the edge nodes closer to users, and thus reduce the cost of access lines by sharing the larger backbone lines among more users
- Provides the possibility to use a smaller number of high capacity backbone links to share the traffic routing between more edge nodes in the network

6.4 The Ring Architecture

By the ring architecture, it is meant that all the edge nodes are placed on a ring in which the routing between a node pair is provided either directly when adjacent, or through other nodes when they are not adjacent to each other. Examples of this type of ring operation are the token ring, or a slotted ring. In a ring architecture of this type, the maximum transport delay occurs normally between nodes that are opposite to each other on the circle where the routing between them takes place through half the number of nodes on the ring. In ring architectures, dual counter rotating rings are implemented to provide continued connectivity between the nodes when one of the nodes fail. However, the ring architecture though is simple to design, it is not much advocated here for the core network because of the fact that it can significantly limit the bandwidth available to each node and increase the transport delays in a heavily loaded network. Besides, rings of this type have been mostly used in the context of local area networking.

7. Design Validation Process

The design process for a network of considerable size consisting of more than a few nodes will require the use of a computer based network performance and analysis tool, in most cases. The tool is used to assess and compare the performance and reliability of alternative design topologies and transport options in the design process. One such tool which is found to provide competitive performance and functionality is the Opnet Modeler, or the Decision Guru version of Opnet. A network performance and analysis tool such as the Opnet can be used to perform the following specific functions in the design process:

- To evaluate the backbone edge to edge network delay statistics, as well as link and node (CPU) utilizations under a given traffic load
- To experiment with different traffic load statistics and models and verify the network performance under all possible traffic flow patterns which may occur in practice
- To perform what if studies to make sure that a selected design will meet the required performance and reliability under specified link and node failure scenarios
- To model the QOS mechanisms implemented within actual network routing and switching hardware and verifying the efficiency of a selected design in meeting service differentiation performance specifications
- To model the effect of the protocol overheads on the network traffic
- To experiment with alternative routing protocols and mechanisms for selection of the optimum routing scheme for a network of considerable size
- To experiment with alternative transport/transmission technologies, and performing cost/performance tradeoffs

8. Conclusion

The emerging 3G networks have been designed with wideband capabilities, allowing significantly higher data rates for mobile multimedia and Internet-based services of the future. The effects of the introduction of packet-switched services as well as higher data rates will pose significant challenges to the design of the core networks. The design of the core networks involves issues related to node location and dimensioning, the choice of appropriate transport technologies, interconnect topologies, traffic considerations, interconnect sizing and quality of service. This paper has reviewed the alternative choices and their merits as well as some of the design parameters that the network operator will face in providing a cost effective network to provide the necessary performance. Building the core network requires an operator to make several decisions, which will affect the scalability, reliability, quality, and cost of the network.

References

- [1] TS 23.060, General packet Radio Service Description
- [2] TS 23.925, UMTS Core network based ATM transport
- [3] TS 26.915, Transmission planning aspects of the services in 3G PLMN
- [4] M. Rahnema, "Overview of GSM Systems and Protocol Architecture", IEEE Communication Magazine, April 1993.
- [5] M. Rahnema, "Frame Relaying and the Fast Packet Switching Concepts and Issues", IEEE Network Magazine, July 1991.
- [6] J. Cai and D. J. Goodman, "General Packet Radio Services in GSM", IEEE Communication Magazine, October 1997
- [7] R. Kalden, I. Meirick and M. Mever, "Wireless Internet Access based on GPRS", IEEE Personal Communication Magazine, April 2000