QoS in Cellular Networks

Dushyanth Balasubramanian, Washington University in Saint Louis, dushyanthb@gmail.com

Abstract:

This paper discusses different schemes for providing Quality of Service (QoS) in cellular networks. Each scheme has its own algorithm to provide QoS and every scheme has its advantages and disadvantages. I have also dealt about an important aspect of QoS which is the Individual QoS (iQoS). iQoS measures the satisfaction rate per user and it is measured individually at each and every user terminal. Voice traffic is very delay sensitive and data traffic is loss sensitive. QoS schemes which try to incorporate both voice and data have to take into consideration this issue. I have also discussed the integration of WLAN/cellular networks in which the high data rate of the WLAN network and the high mobility of cellular networks are utilized effectively.
Back to Raj Jain's Home Page

Table of Contents:

1. Introduction: What is QoS

Quality of Service (QoS) in cellular networks is defined as the capability of the cellular service providers to provide a satisfactory service which includes voice quality, signal strength, low call blocking and dropping probability, high data rates for multimedia and data applications etc. For network based services QoS depends on the following factors [Jain 06]

- **Throughput** The rate at which the packets go through the network. Maximum rate is always preferred.
- **Delay** This is the time which a packet takes to travel from one end to the other. Minimum delay is always preferred.
- **Packet Loss Rate** The rate at which a packet is lost. This should also be as minimum as possible.
- **Packet Error Rate** This is the errors which are present in a packet due to corrupted bits. This should be as minimum as possible
- **Reliability** The availability of a connection. (Links going up/down).

It is for these reasons that providing QoS has been a great challenge in the past and it continues to be a hot topic as there is still a lot of scope to provide better service standards.

1.1 Why do we need QoS

Imagine a situation where you are hardly able to hear what your friend is talking over the phone or the phone gets cut when you are talking something important. These things are highly undesirable and you do not want to get low quality service for paying high monthly bills. Communication plays a major role in today's world and to support it QoS has to be given maximum priority. It is important to differentiate the traffic based on priority level. Some traffic classes should be given higher priority over other classes, Example: voice should be given a higher priority compared to data traffic as voice is still considered as the most important service. It should be noted that more preference has to be given to customers who pay more to get better service, without affecting the remaining customers who pay normal amount. To realize all these things effective QoS schemes are needed. Issues and schemes related to providing better QoS is the main subject of this report.

1.2 Quality of Service Challenges

In wireless mobile networks QoS refers to the measurement of a system with good transmission quality, service availability and minimum delay. In 4G it is expected to have a reliability of at least 99.999

referred to as five nine reliability. The major challenges when considering QoS in cellular networks are varying rate channel characteristics, bandwidth allocation, fault tolerance levels and handoff support among heterogeneous wireless networks. It is fortunate that each layer which includes physical, MAC, IP, TCP and application have got their own mechanisms to provide QoS. It is important to guarantee QoS in each layer so that the network is more flexible and tolerant to QoS issues. Some of the other challenges are efficient usage of the spectrum as its availability is limited. Bandwidth allocation plays a major role with respect to this aspect. It must be made sure that bandwidth is allocated in an efficient manner and also the remaining bandwidth should not be wasted. Some schemes like Renegotiation scheme [Monego 05] takes care of this issue by allocating the remaining bandwidth to lower priority classes. Things get even more complicated when data and voice service has to be supported. Voice services are very delay sensitive and require real - time service. On the other hand data services are less delay sensitive but are very sensitive to loss of data and also they expect error free packets. So both these factors have to be considered for providing QoS for voice and data services.

1.3 What has been achieved so far in QoS

In 1G networks and 2G networks such as GSM and CDMA there was only one aspect of QoS and it is voice. Providing quality speech was the major concern. Now in 3G networks QoS has to be provided for voice as well as data. Still priority is given for voice services as they are considered as the primary service. They are very delay sensitive and require real - time service. Data services are comprised of text and multimedia. These services are less delay sensitive but expect better throughput and less or no loss rate.

The rest of this paper is organized as follows. Section 2 talks about the Universal Mobile Telecommunication System (UMTS) QoS architecture and every layer in the architecture is described with their functionalities. In Section 3 we deal with different classes of QoS and also the various attributes which specify the iQoS. Section 4 deals with the various QoS schemes in cellular networks such as Fault Tolerant Dynamic Channel Allocation, Call Admission Control (CAC) algorithm, QoS provisioning using mobile prediction techniques to improve hand over loss probability, Dynamic channel allocation using Renegotiation techniques and iQoS which is an important aspect of QoS from user's point of view. Section 5 discusses about the coupling of WLAN and cellular networks so that the high data rate of the WLAN and the high mobility of the cellular networks is utilized. Two different architectures loosely coupled approach and tightly coupled approach are considered and explained. Finally Section 6 summarizes the whole report. List of Acronyms and References are given towards the end.

Back to Table of Contents

2. UMTS QoS Architecture [3GPP 23.107]

Overview of Different Levels of QoS

An end - to - end service implies that the communication takes place from one Terminal Equipment (TE) to another. The user of the network service is provided with a QoS and it is the user who decides whether he is content with the QoS or not. A bearer service with clearly specified characteristics and functionalities is to be set up from source to destination to meet the network QoS requirement. The architecture of UMTS bearer service is a layered one as shown in figure 1. Bearer services provide QoS based on services provided by the layers below them.
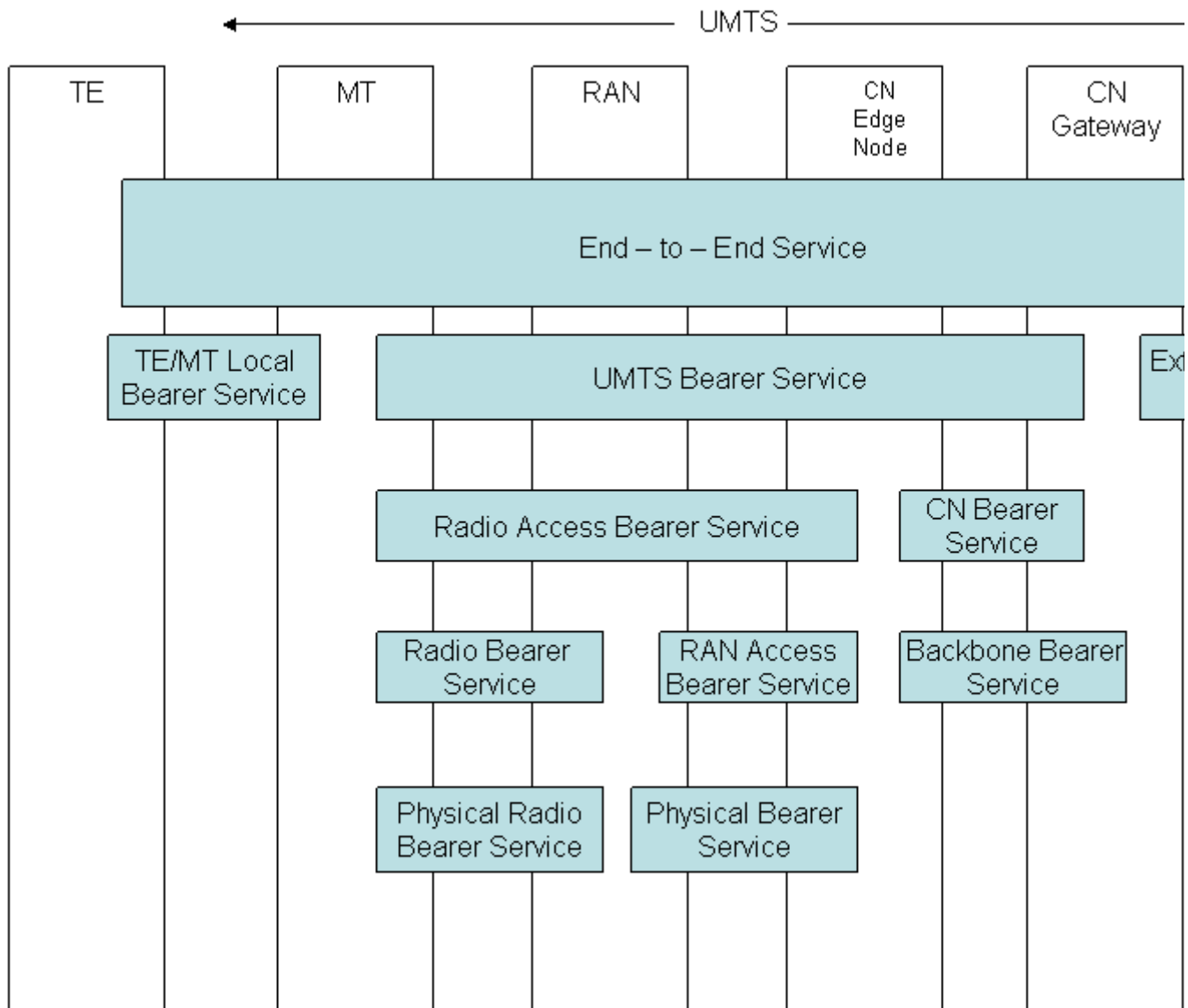
Figure 1: UMTS QoS Architecture [3GPP 23.107]

2.1 End - to - End Service and UMTS Bearer Service [3GPP 23.107]

As shown in the figure 1 the End - to - End service layer is the topmost layer of the QoS architecture. It makes communication possible from one TE to the other. It is shown that a TE is connected to the UMTS network by means of a Mobile Termination (MT). The End - to - End services used by a TE will be realized by the below layers namely TE/MT local bearer service, a UMTS bearer service and external bearer service. The UMTS operator offers services provided by the UMTS bearer service. Thus UMTS bearer service provides the UMTS QoS.

2.2 The Radio Access Bearer Service and the Core Network Bearer Service [3GPP 23.107]

The UMTS Bearer service is comprised of two parts which are the Radio Access Bearer Service and the Core Network Bearer Service. Both these services take care of the Bearer service over the network topology taking into consideration attributes such as mobility and mobile subscriber profiles. The Radio

Access Bearer Service makes provision for the transport of signaling and user date between MT and CN Edge Node with QoS adequate to the negotiated UMTS bearer service or with default QoS for signaling. Radio Bearer Services provides the unequal error protection if that feature has to be supported.

The Core Network Bearer Service of the UMTS core network connects the UMTS CN Edge Node with the CN Gateway to the external network. This service controls and utilizes the backbone network efficiently in order to provide the contracted UMTS bearer service. [3GPP 23.107]

2.3 The Radio Bearer Service and the RAN Access Bearer Service [3GPP 23.107]

The Radio Access Bearer Service is realized by a Radio Bearer Service and an RAN Access - Bearer Service. The Radio Bearer Service covers all the aspects of the radio interface transport. [3GPP 23.107] To support unequal error protection, RAN and MT has the ability to segment/reassemble the user flows into the different subflows requested by the Radio Access Bearer Service. The segmentation/reassemble is given by the Service Data Unit (SDU) payload format signaled at Radio Access Bearer establishment. [3GPP 23.107]

2.4 The Backbone Network Service [3GPP 23.107]

The core Network Bearer Service uses a generic Backbone Network Service. The Backbone Network Service covers the layer 1/layer 2 functionality and is selected according to the operator's choice in order to fulfill the QoS requirements of the CN Bearer Service. [3GPP 23.107]

Back to Table of Contents

---

3. UMTS QoS Classes and Attributes [3GPP 23.107]

There are four different QoS classes:

- Conversational class;
- Streaming class;
- Interactive class;
- Background class

The main differences between these QoS classes are how delay sensitive the traffic is. Conversational class is meant for traffic which is very delay sensitive while the Background class is the most delay insensitive class.

Conversational and Streaming classes are used to serve real - time traffic flows which are very sensitive to the delay. Examples of Conversational classes include real - time services like video telephony and data streams.

Interactive class and the Background are mainly meant for applications like WWW, Email, FTP, News and Telnet. Since these classes are less delay sensitive compared to the conversational and streaming classes, both the classes provide better error rate by means of channel coding techniques and retransmissions. This means that packet retransmission is done whenever packet error/packet loss/packet order mismatch takes place. The reason is that these classes are delay insensitive but expect high throughput and less error rates. The main difference between Interactive and Background class is that

Interactive class is mainly used for interactive applications like interactive Email and interactive Web browsing, while Background class is meant for background traffic, e.g. background downloads or Emails and background file downloading. [3GPP 23.107] The scheduling algorithm gives more priority to the interactive class than the background class. Therefore the background applications use the transmission resources only when the interactive applications do not need them.

Table 1: Traffic Classes [3GPP 23.107]

Traffic class	Conversational class conversational RT	Streaming class streaming RT	Interactive class Interactive best eff
Fundamental characteristics	- Preserve time relation (variation) between information entities of the stream - Conversational pattern (stringent and low delay)	- Preserve time relation (variation) between information entities of the stream	- Request respons: pattern - Preserve payloa content
Example of the application	- voice	- streaming video	- Web browsing

3.1. Conversational class

Applications which use this class include telephony speech, voice over IP and video conferencing. Real time conversation is always performed between groups of humans and so this is the only scheme where the required characteristics are strictly given by human perception. [3GPP 23.107]

Real time conversation scheme is characterized by that the transfer time should be low because of the conversational nature of the scheme and at the same time the time variation between information entities in the stream should be preserved in the same way as for real time streams. [3GPP 23.107] The maximum transfer delay is dictated by how much delay the humans can tolerate for audio and video. Therefore the bounds for acceptable transfer delay are very stringent, and if transfer delay is not low enough then it affects the quality. The transfer delay should be lower and stringent than the round trip delay for this class.

3.2. Streaming class

Applications for this class includes listening to or looking at real time video (audio). This scheme is

characterized by that the time relations (variations) between information entities (i.e. sample packets) within a flow shall be preserved, although it does not have any requirements on low transfer delay. [3GPP 23.107] The delay variation of the end - to - end flow should be limited to preserve the time variation between the information entities of the stream.

3.3.Interactive class

Applications for this class includes browsing the web, database retrieval, access of server etc. Interactive traffic is a communication scheme which is characterized by the request response pattern of the end user. Round trip delay is the most important attribute for this class. Another important attribute is that the error rate should be very low in the data transfer. [3GPP 23.107]

3.4. Background class

This is a service class in which the applications run in the background. Example includes an email program which is running in the background. It sleeps for most of the time and wakes up when an email arrives. Other examples include SMS, download of databases. Background traffic is characterized by that the destination is not expecting the data within a certain time. Thus this class is less delay sensitive and another characteristic is that the contents should be delivered with low error rate. [3GPP 23.107]

3.5. QoS Attributes

UMTS bearer service attributes describe the service provided by the UMTS network to the user of the UMTS bearer service. A set of QoS attributes (QoS profile) specifies this service. [3GPP 23.107] Below is a list of attributes.

Maximum bit rate (kbps) is defined as the maximum number of bits delivered by the UMTS and to the UMTS at a Service Access Point (SAP) within a period of time, divided by the duration of the period. The maximum bit rate is the upper limit a user or application can accept or provide. [3GPP 23.107] The purpose of this attribute is (1) to limit the delivered bit rate to applications or external networks with such limitation and (2) to allow maximum waned user bit rate to be defined for applications able to operate with different rates.

Guaranteed bit rate (kbps) is defined as the guaranteed number of bits delivered by UMTS at a SAP within a period of time (provided that there is data to deliver), divided by the duration of the period. The purpose of this attribute is that it describes the bit rate the UMTS bearer service shall guarantee to the user or application. [3GPP 23.107]

Delivery order (y/n) indicates whether the UMTS bearer shall provide in - sequence SDU delivery or not. The purpose of this attribute is to know whether out- of - sequence SDU is allowed or not. [3GPP 23.107]

Maximum SDU size (octets): is defined as the maximum SDU size for which the network shall satisfy the negotiated QoS. The purpose of this attribute is that it is used for admission control and policing and optimizing transport. [3GPP 23.107]

SDU format information (bits) defines the list of possible exact sizes of SDUs. [3GPP 23.107]

SDU error ratio: indicates the fraction of SDUs lost or detected as error packets. SDU error ratio is defined only for conforming traffic. The purpose of this attribute is that it is used to configure the

protocols, algorithms and error detection scheme. [3GPP 23.107]

Residual bit error ratio: indicates the undetected bit error ratio in the delivered SDUs. If no error detection is requested, Residual bit error ratio indicates the bit error ratio in the delivered SDUs. The purpose is that it is used to configure radio interface protocols, algorithms and error detection coding. [3GPP 23.107]

Delivery of erroneous SDUs (y/n) indicates whether SDUs detected as erroneous should be delivered or discarded. This is used to decide whether error detection is required or not and also whether frames detected as errors should be thrown or not. [3GPP 23.107]

Transfer delay (ms) is defined as the maximum delay for 95th percentile of the distribution of delay for all delivered SDUs during the lifetime of a bearer service, where delay for an SDU is defined as the time from request to transfer and SDU at one SAP to its delivery at the other SAP [3GPP 23.107]

Traffic handing priority specifies the relative importance of handling of all SDUs belonging to the UMTS bearer compared to the SDUs of other bearers. [3GPP 23.107]

Back to Table of Contents

4. Different QoS Schemes In Cellular Networks

There are many QoS schemes which have been deployed for cellular networks and each scheme has its own advantages and disadvantages. In this section we are going to look into some of the fundamental and effective QoS schemes which are used for providing voice and data services. Fault Tolerant Dynamic Allocation scheme looks into methods of reusing the channels effectively between two cells, which are separated by a minimum distance so that they do not interfere with each other. The channels are allocated dynamically as opposed to static allocation where the channels are allocated and reserved beforehand. [Yang 03] The next scheme is the Call Admission Control (CAC) which employs pre - blocking of calls based on the available bandwidth for handling calls. This algorithm is based on two schemes which were used earlier namely Pre request scheme and the guard channel scheme. CAC algorithm utilizes both the schemes and gives better performance in terms of successful call completion rates (SCCR) and provides guaranteed QoS for profiled users. [Kovvuri 2003] In the Mobility prediction techniques hand off losses are reduced and due to which the blocking and the dropping probabilities are significantly reduced. In this mobility prediction scheme road topology information is gathered and stored in a database and the path or the trajectory of the mobile host is calculated. No assumption about the shape of the cell is assumed. [Soh 03] The renegotiation scheme is a scheme in which the bandwidth allocation is changed dynamically based on the availability. If a low priority service has been admitted with a bandwidth less than what it had asked and after sometime extra bandwidth is available due to completion of a high priority service then the remaining bandwidth is given to the low priority service and thus increases the QoS of the lower priority service. This scheme also ensures that the higher priority services get their requested bandwidth and they are not affected in any way. [Monego 05]

4.1 Fault Tolerant Dynamic Channel Allocation Scheme [Yang 2003]

In this scheme the channels are allocated dynamically based on demand and hence increasing the channel utilization and also the QoS. The channel allocation schemes are centralized and distributed. In a centralized approach there is a central controller which is responsible for channel allocation and requests are sent to the controller, whereas in a distributed approach there exists a Mobile Service

Station (MSS) in each cell which takes care of the channel allocation for that particular cell. Distributed approach is more scalable and reliable and thus this approach is used in this algorithm.
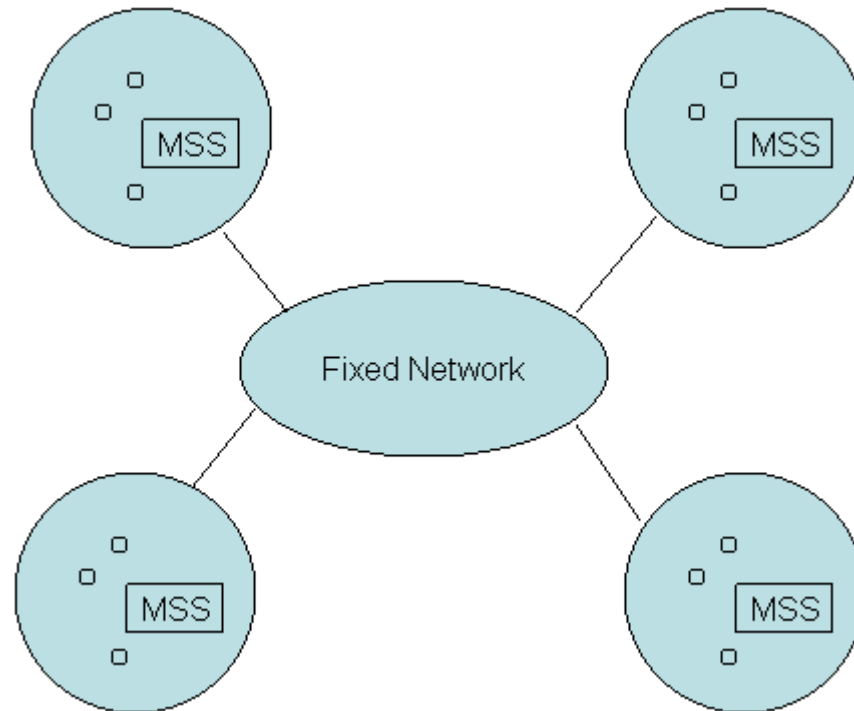

Figure 2: A Model of Wireless Communication Networks [Yang 03]

Figure 2 shows the model of wireless communication. There are many cells which are comprised of many Mobile Hosts (MH) and a Mobile Service Station (MSS). The cells are connected to a fixed network. Each MH when it wants to communicate with another mobile host has to request a channel from MSS for communication. There are two types of channels namely communication channel and control channel. A communication channel is used to support communication between a MH and a MSS whereas control channels are used for sending control messages like MH requesting the MSS to set a communication channel by means of a control channel. The MSS allocates channels to the MH based on whether co - channel interference occurs. If there is no such interference then it allocates the channel to the MH for communication.

There two different approaches for requesting channels. In Centralized approach each MH requests for channels from a central controller called Mobile Switching Center (MSC). So the channel allocation is done by MSC in such a way that no co - channel interference takes place. The disadvantage of this approach is that the MSC can become a bottleneck in the network and the failure of the MSC means the whole network is affected. Thus this approach is not at all scalable and robust.

In the Distributed approach there is no central controller and each cell has its own MSS. It is the job of the MSS to allocate channels and it makes sure that no channel interference occurs. In this approach a cell updates its neighbors about its usage of channels and lets them know when it relinquishes any channel so that others can use it. When a cell wants a channel it sends a request message to all its neighbors and when it gets reply from all of its neighbors that a channel can be used, then it uses that channel.

4.1.1 System Model [Yang 2003]

In this section we talk about the system model of this distributed fault tolerant approach. Channels are allocated dynamically and not allocated in advance to any cells. A three cell cluster model is used and each cell has got six neighbors. If a channel is used by one particular cell then none of its neighbors can use the same channel and thus prevents co - channel interference. A cell can borrow a channel to use if it does not have any channel for communication. The request is to be satisfied by any of the cell's neighbors who can lend a channel and also making sure that co - channel interference does nor occur. There are two modes of operation which are update mode and search mode. In an update mode a cell informs its channel usage information to all its neighbors whenever it borrows or releases a channel. Thus it updates all the neighbors about the status. In a search mode, when a cell needs to borrow a channel it sends request message to all of its neighbors. Based on the reply it gets from the neighbors it borrows a channel.

4.1.2 Distributed Channel Allocation Algorithm [Yang 2003]

In this model a three cell cluster model is there and each cell has six neighbors and each of them have a unique neighbor id ordered from one to six. Channels are not pre-allocated to any cell. The channel with lowest frequency has the minimum order and the channel with highest frequency has the maximum order. A cell C_i selects a channel with the highest order to support a call if it is available.
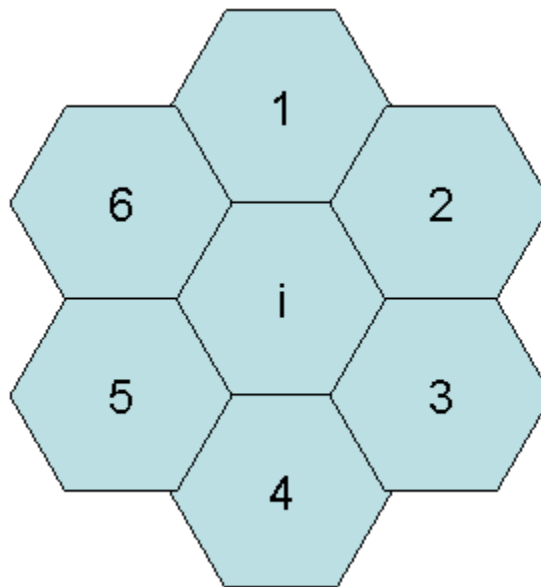

Figure 3: Cell Cluster Model to Illustrate Distributed Channel Allocation [Yang 03]

In figure 3 each cell is numbered from 1 to 6 and we are going to use cell C_i to represent a particular cell. When C_i requests for a channel to use it is in search mode and it's called a borrower. It sends a broadcast message to all its neighbors for the channel and sets a timer and waits till the timer times out. When a cell is in borrower mode it does not respond to any query from any other cell requesting channels. This is because the borrower itself is in search for channels and so it's not going to be of any help to other channels which are requesting. This scheme is nice in the way that it prevents unnecessary congestion in the network. Now after the timer times out, the cell C_i will borrow a channel based on the replies which it got from its neighbors.

To prove that this algorithm works lets assume that the cell C_i has got replies only from 2 neighbors, 1

and 4 and there is a channel **r** which is allocated to both the cells 1 and 4. In this case the cell C_i can borrow the channel **r** even though other cells (2,3,5,6) have not replied other than cells 1 and 4. The reason is that since channel **r** is allocated to cells 1 and 4, it cannot be allocated to the neighbors of cells 1 and 4 (cells 2, 6, 3 and 5) due to co - channel interference. Thus channel C_i can borrow the channel with the replies it got from just two channels. This proves that this algorithm is fault tolerant.

Thus the above algorithm which we saw was fault tolerant as it need not wait for replies from all the channels. This also means that the algorithm is more scalable. It uses distributed channel allocation scheme in which channels are allocated dynamically based on request. This algorithm also allows reuse of channels. This means that a lender can lend a channel to two or more cells simultaneously provided that no co - channel interference occurs.

4.2 Call Admission Control (CAC) Scheme [Kovvuri 2003]

In the CAC algorithm new call arrival rates are estimated continuously and if they are higher than a predetermined level some calls are blocked irrespective of whether a channel is available or not. The objective of this scheme is to maintain the new call arrival rate lesser than a predetermined level. In this scheme a comparison is made with the existing two schemes namely prerequest scheme and the guard channel scheme and various advantages and disadvantages are given for the two schemes and then a CAC algorithm is developed which provides a better QoS than the existing two schemes. The two metrics used for QoS in this algorithm are Forced Termination Probability (FTP) which is defined as the ratio of the number of calls which are forced to terminate because of failed handoff to the number of calls that successfully entered the network. Another metric is the Successful Call Completion Rate (SCCR) which is defined as the number of calls which are completed successfully in a unit time by each cell. [Kovvuri 03] So lower FTP and higher SCCR is what ideal algorithms will try to achieve and this algorithm achieves that.

4.2.1 Channel Prioritization Schemes

In the Guard Channel Scheme some channels are exclusively reserved for the handoff calls and these channels are called guard channels. A certain number of channels say G out of C channels are exclusively reserved for handoff calls of the profiled users. The remaining channels are shared for all types of calls. These calls include new calls, handoff calls of profiled users and handoff calls of non profiled users. This is an example of static allocation of channels.

In the channel prerequest scheme the channels are requested beforehand before the handoff occurs. The information about the mobility patterns of the profiled users is in the Home Location Register (HLR) and using this information the handoff behavior of the profiled users can be predicted. The prerequest channels are requested from the neighboring cell for a certain amount of time called as the reservation period. By increasing the reservation period the probability of forced termination can be significantly reduced.

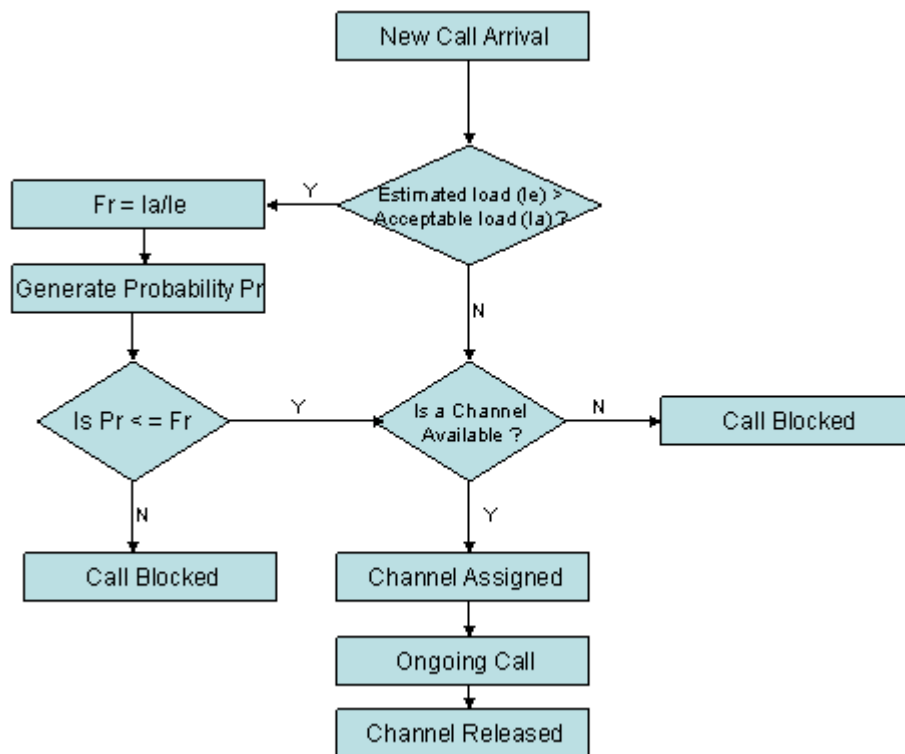## 4.2.2 Call Admission Algorithm (CAC)

Figure 4: Flow Chart for CAC Algorithm [Kovvuri 03]

In the CAC algorithm the acceptable load is calculated based on simulation results and this value is used for comparison purpose. The estimated load is also calculated and it is checked with the acceptable load. If the estimated load is lesser than or equal to the acceptable load, then attempts are made to allocate channels for all the incoming calls. If the estimated load is greater than the acceptable load then only a fraction of the incoming calls will be allocated channels and the remaining fraction of the calls will be discarded even if there are available channels. This is called pre - blocking of channels and this scheme improves the FTP and SCCR of the profiled users.

4.3 QoS based on Mobility Prediction Techniques [Soh 2003]

Mobile prediction techniques are employed to find the path or the trajectory of a mobile node and it is stored in a database from time to time. This technique helps in reserving resources for MH before a hand off occurs so prioritizing resources takes place for a node before its hand off and thus this decreases the call blocking rate at hand offs. However the new call admission rate is reduced as more resources are reserved for hand off calls. Forced termination can be reduced by increasing the number of new call blocking probability. But this is not very efficient utilization of the radio resources. By using mobile prediction techniques, we can know in advance when a handoff will take place and so dynamic allocation of resources for the handoff can be done. This ensures that the new call blocking probability is not increased so much to preserve resources for the hand off calls. There are many mobile prediction techniques which have been employed in the past which includes GPS positioning methods in which each MH is integrated with a GPS receiver so that the path of the MH can be traced.

4.3.1 Road Topology based Mobile Prediction Techniques [Soh 2003]

This technique is based on the fact that MH's in vehicles will encounter the most frequent hand offs and so studying their characteristics would prove to be more beneficial. Since the vehicles travel on the road, the road topology is studied and is used in the prediction algorithm. In this scheme the base station will

do the mobile prediction based on the road topology information it has. Since the base station has more storage capability and resources than the MH, it is used for better accuracy.
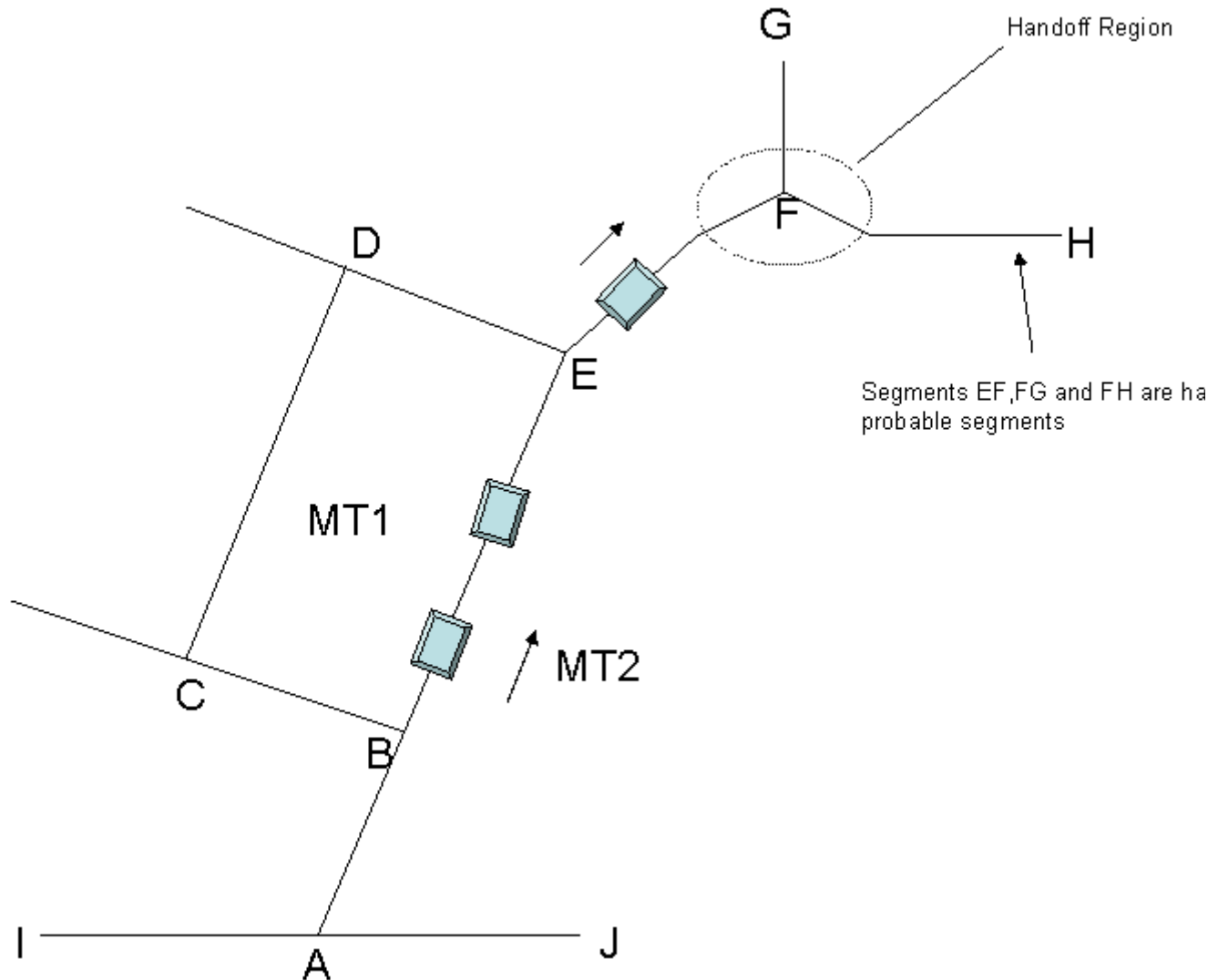

Figure 5: Road Topology Information for Mobility Prediction [Soh 03]

In this technique there are several base stations, which are updated with the MH's latest position at regular interval of time. Each base station will maintain a database which contains information about the road topology. As shown in figure 5 the road which consists of bends are broken into many piecewise linear line segments and the coordinates for these line segments are stored separately. The database has information such as average time to transit a segment, neighboring segments at each junction and the probability of the MH to do an hand off and go to a next neighboring segment. The database is updated periodically and at every instant the information about the location of the MH is obtained. Segments which have already had hand offs are referred to as handoff probable segments. Using this model, accurate prediction is done, which is useful in prioritizing resources for the handoff of an MH.

4.4 Dynamic Allocation Scheme using Renegotiation [Monego 05]

In a Dynamic Allocation scheme using Renegotiation, the unused resources of the network are explored and they are allocated to services which got a lesser bandwidth at the time of admission when they

actually wanted more. In other words, renegotiation of the bandwidth is done for a lower priority service when the medium is free, thus increasing the overall bandwidth of the lower priority services. The scheme on the other hand also maintains the bandwidth of the higher priority sources.

4.4.1 Different Service Classes

Conversational and streaming classes are associated with real time traffic and are extremely delay sensitive. Examples of these kinds of services are video telephony, telnet, voice and video. On the other hand Interactive and background processes are less delay sensitive and they include applications like FTP, WWW, email and news. Since they are less sensitive to delay both these classes provide better error rates by using channel coding techniques. The difference between the interactive and the background services are that interactive services are mainly used by interactive applications like for example, interactive email. Background classes are used for background applications which include background browsing and background emails.

4.4.2 Renegotiation Scheme [Monego 05]

In this scheme the conversational classes are given the maximum priority and they are priority 1 class. The streaming class is also given high priority and it belongs to priority 2 classes. Now these priority 1 and priority 2 classes will be admitted only if there is enough resources (bandwidth) to satisfy their requirements. If not their requests won't be admitted. The interactive service is given the lowest priority and it is 3. So the request from these classes are admitted even if the network have a lesser bandwidth than what they had requested. The advantage of the Renegotiation scheme when compared to the CAC scheme is that in the CAC scheme when a bandwidth is allocated to priority 1 and 2 they cannot be transferred to a lower priority class even after the higher priority class leaves the network. In renegotiation scheme a priority 3 application can use more bandwidth that what was allocated for it. This is made possible because any unused resources by the higher priority classes can be transferred to this class. On the other hand when a high priority class arrives again it will not be blocked if the system has low resources. At that time the bandwidth given to a lower priority class will be taken back and will be allocated to the high priority class. Thus the higher priority class is not harmed in any way.

4.4.3 Renegotiation by Flow Termination [Monego 05]

The renegotiation by flow termination is a means by which more bandwidth is allocated for a low priority flow when a high priority flow ends or when there are more resources available in the network.
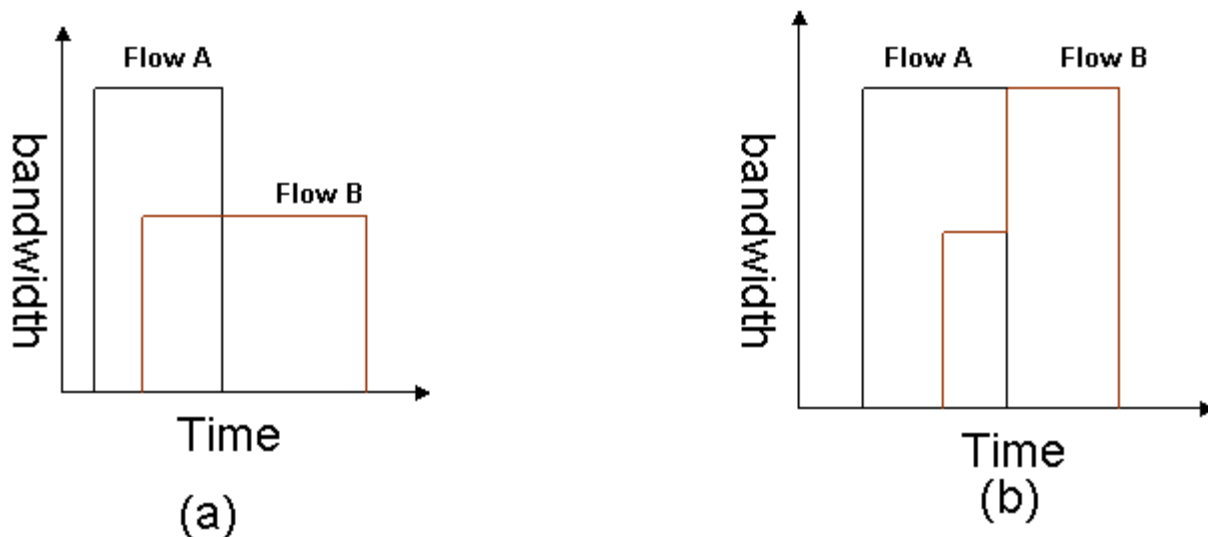
Figure 6: Behavior of Two Flows Without (a) and With (b) Renegotiation [Monego 05]

In figure 6 behavior of two different flows are shown without and with renegotiation. In case (a) without renegotiation the bandwidths of A and B are shown. Even after the termination of flow A the bandwidth of B remains the same. In case (b) after the termination of flow A the bandwidth of flow B is increased due to the availability of resources.

4.5 Individual QoS for Voice Services [Medeisis 04]

The Individual QoS is a very important measurement of QoS and its value is often overlooked. So far QoS based on cell or network level is what is done in all the QoS studies. Individual QoS (iQoS) measures the satisfaction rate per user and that is one of the most important parameters which has to be considered. It is done individually at each and every user terminal. The estimation of overall QoS in the network does not give a very accurate picture of what each and every individual user is encountering. For Example the average rate of dropped calls in the whole network may be as low as 1 - 2 percent. However, some users may experience much more frequent dropped calls, probably 30% of their calls may be dropped. This means that the overall network level QoS is high (more than 95 percent) but some users have become victims of this average process and they are charged by their network operators for the QoS which they actually did not receive. It turns out that users in area with very less radio coverage often experience this kind of loss then the average cannot be taken into account for those users and their dissatisfaction is hidden in the average. Thus it is very important to measure each users satisfaction in a network.

4.5.1 Purpose of Measuring iQoS [Medeisis 04]

It may be observed that whatever global QoS management concept is realized in a network by definition it can never produce the same amount of user level satisfaction. The radio strength which a user gets depends on his physical position with respect to the base station. For users who are very much near the base station, they might get very good signal strength and voice quality but for users who are very far away from the base station may experience a very poor network connection and also lots of dropped calls. The median signal level characterization of radio link budget for particular point may be easily vary by some 10 - 20 dB just within a short distance of few tens of meters. Thus it will be much better if iQoS is included in the billing so that every individual user is aware of what QoS he is getting and he knows whether he gets the quality for what he is paying. [Medeisis 04]

4.5.2 Concept of iQoS Rating [Medeisis 04]

It is evident that in the users perspective setup failure and drop call will be the most annoying thing as he will have to pay for the services not received if error occurs after the billing timing is triggered. So for measuring iQoS first of all differentiation of users who have not received a dropped call versus users who have received dropped call should be made. The Quality of service should be measured for only those people who experience bad service like dropped calls and call blocking.

Back to Table of Contents

---

5 QoS in Cellular/WLAN Networks [Begain 04] [Song 05]

Wireless LAN (WLAN) systems provide very high data rate but less mobility and on the other hand cellular networks provide a very high mobility but not high data rate. So a combination of both these networks will result in a network with a very high data rate as well as mobility. It is using this idea that people tried to combine both the WLAN and Cellular networks and it has proved to provide good data rate as well as mobility. The issue to be considered here is that the QoS of cellular networks is well defined and it is much easier to provide compared to the QoS in the IP network. There is no assured QoS in the IP network and hence when combining both these networks we will have QoS issues which needs to be solved and its much difficult to manage QoS in Integrated Networks.
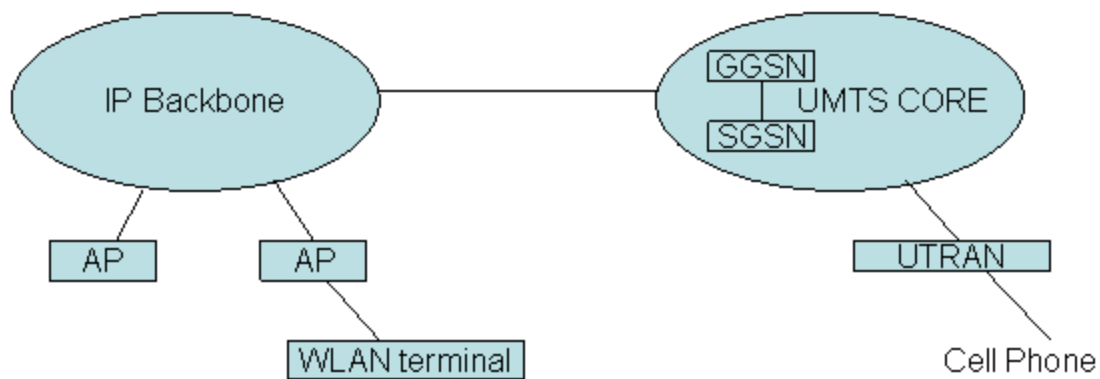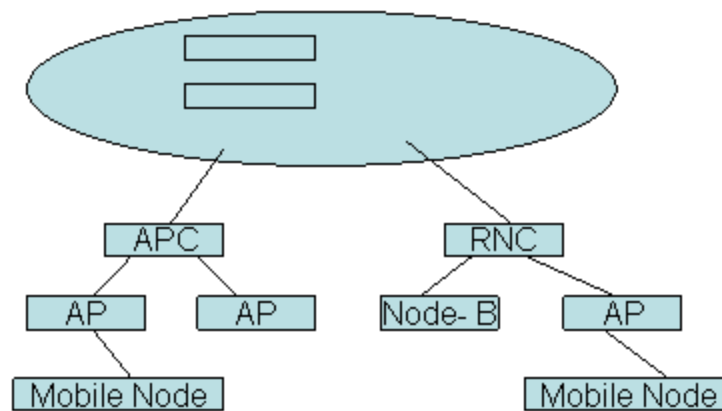
Figure 7: Loose Coupling (a) and Tight Coupling (b) [Begain 04]

In Tight coupling the WLAN technology is employed as a new radio access technology within the cellular system. Regardless of the access technology there will be one common cellular core networks. This can be done by connecting WLAN AP to a radio network controller (RNC). Loose coupling on the other hand is less connection between the two networks. In this scenario WLAN and the cellular networks are two separate access networks. The WLAN access network is attached to the Internet backbone and the cellular network into the cellular core network. [Begain 04]

5.1 Tight Coupled QoS Model Structure [Begain 04]

In this Tight coupled architecture the cellular network is based on an IP based packet switching network. This architecture assumes a packet switching core network based on UMTS network architecture. [Begain 04] Figure 8 shows the components of this tightly coupled architecture.
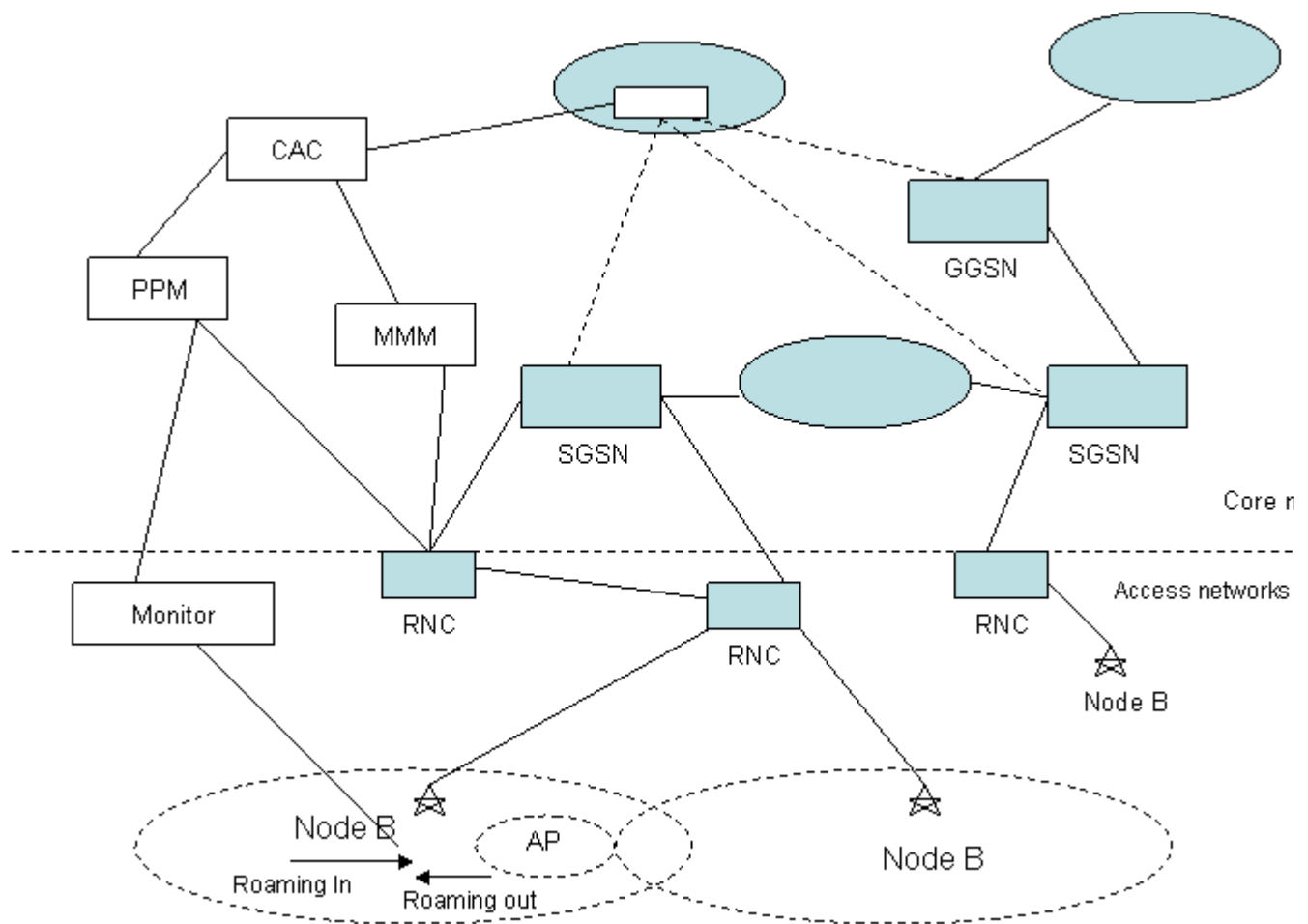
Figure 8: Tightly Coupled QoS Model Architecture [Begain 04]

A Policy Provisioning Module (PPM) The PPM takes the users QoS and decides which class the traffic belongs to (priority level) and then it handles the users request to the Connection Admission Control Module (CAC) for further processing of the request as to allow the traffic or not.

A Connection Admission Control Module (CAC) The CAC modules accepts the traffic flows and allots bandwidth to all the flows. It can also reject flows if there is no available bandwidth for a particular flow. It also maintains the QoS of the already existing connections. The CAC module first receives a connection request from the PPM and then it consults the Mobility Management Module (MMM) to know about the mobility status and the information about nodes. It then uses reservation protocols such as RSVP to reserve connections.

A QoS Mobility Management Module (MMM) MMM monitors the terminals which are no longer connected to the network, idle terminals and the terminals which are connected to the network. It also monitors the connected nodes which are mobile at high speeds. Issues such as handover between a cellular to WLAN and WLAN to cellular networks are taken care by this module. Another important aspect of the MMM is that it monitors the QoS of the users and if the QoS profile is not satisfied then it has a feedback mechanism in which the it informs the CAC about the current QoS of the user and triggers more action for improvement of the QoS.

5.2 Loosely Coupled Cellular/WLAN interworking Model [Song 05]

In this loosely coupled architecture there is no direct link between the WLAN and the cellular networks. The gateway connects the WLANs to the Internet backbone. The advantage of this approach is that there is not dependency between the two networks. Each one can be deployed independently and one network wont be a bottleneck to the other. An all IP DiffServ platform is the most popular architecture for this loosely coupled scheme and we are going to look into this architecture.
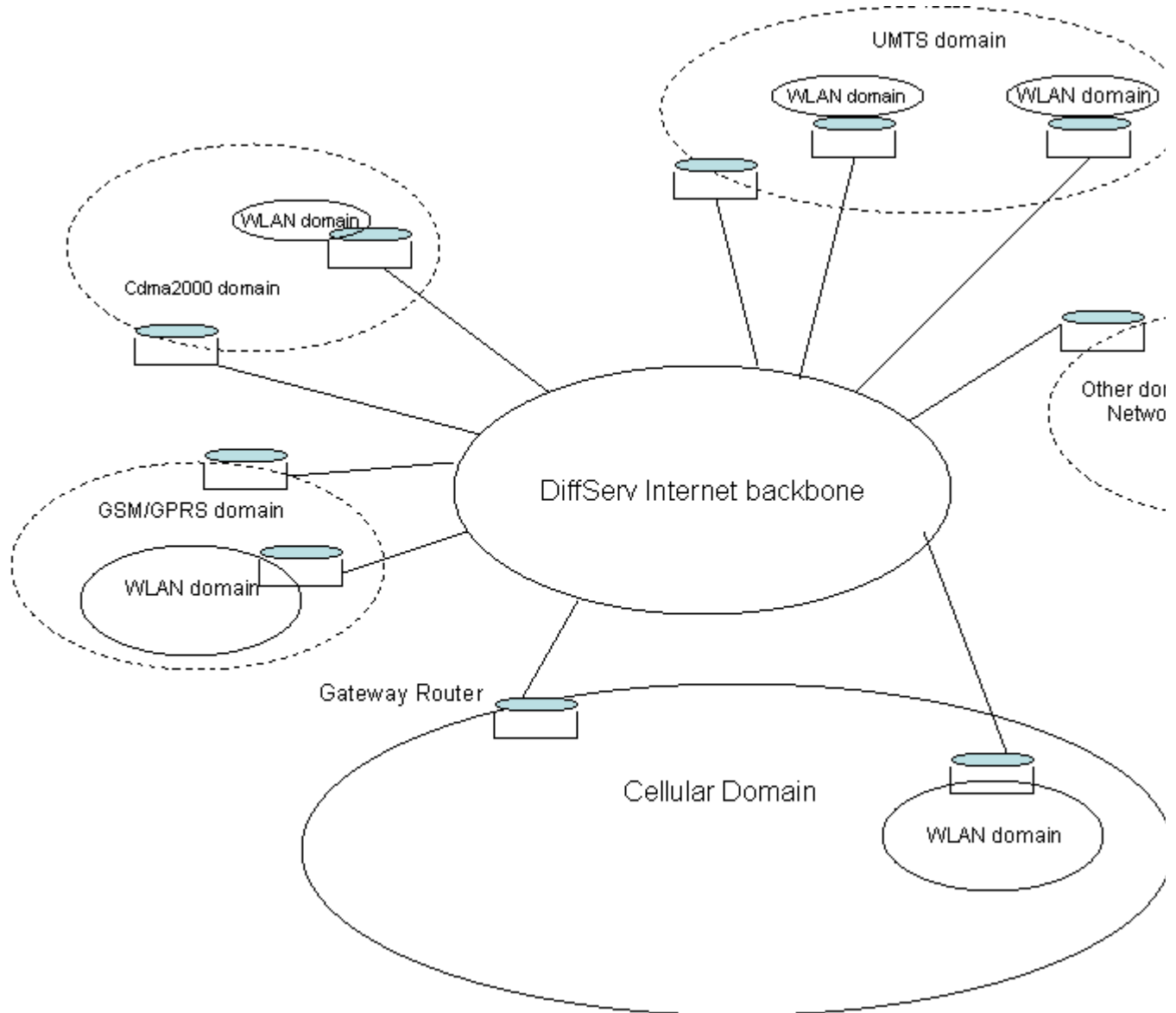

Figure 9: Loosely Coupled cellular/WLAN Architecture in a DiffServ Platform [Song 05]

Figure 9 shows the loosely coupled cellular/WLAN architecture in a DiffServ platform. This illustrates the integration of WLAN with different cellular architectures such as UMTS, code division multiple access (cdma2000) and GSM/GPRS. The DiffServ platform is a domain based architecture in which different domains can freely choose their own system mechanisms and it should ensure that the service level agreements (SLAs) with neighboring domains are satisfied.

5.2.1 Call Admission for Voice and Data Traffic

The two most important services voice and data are considered in this section. Voice traffic is delay

sensitive and requires real - time transmission. It can tolerate some data losses. On the other hand data traffic is not so delay sensitive but they are very sensitive to packet loss and error rates. Voice calls are always given priority over data traffic in cellular networks, as voice calls are more delay sensitive and also cellular networks provide better QoS which is needed for voice services. On the other hand data traffic will most likely end in WLAN coverage as it gets better bandwidth in that region.
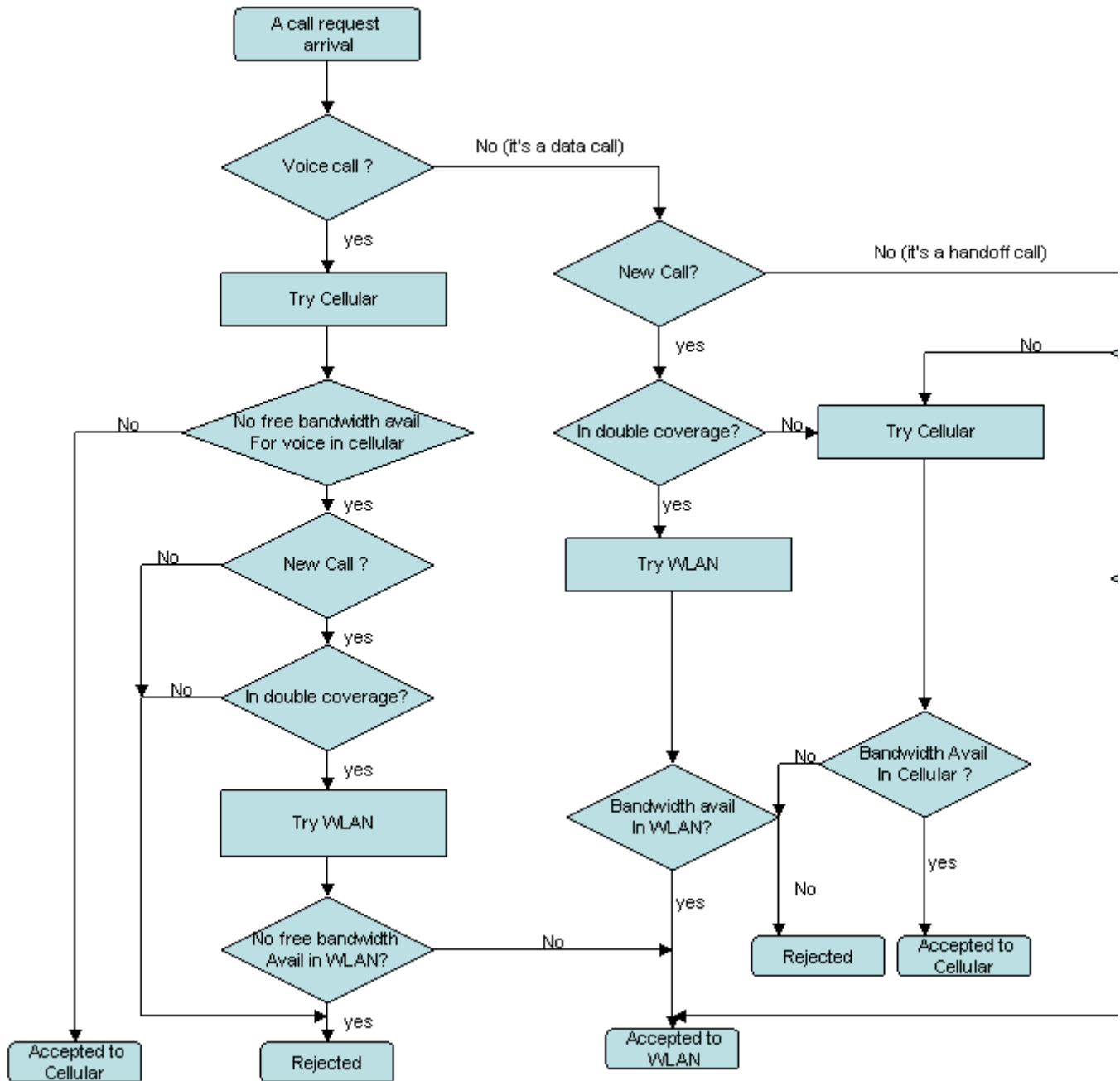

Figure 10: Call admission procedure for cellular/WLAN interworking [Song 05]

Figure 10 explains in detail the Call admission procedure for cellular/WLAN networks. An area with only cellular coverage is called as cellular - only area and an area with WLAN and cellular coverage is called double - coverage area. For voice calls, it has got two choices. It can either try cellular area or WLAN coverage area. If a voice call is rejected in a cellular coverage area the it tries in WLAN area. The request is rejected if there is not enough bandwidth to accommodate the voice call in this area. A

data call has only one option which is the WLAN area. If it is rejected there it wont try the cellular area as the data traffic overflow in cellular area will degrade the QoS of voice traffic. To provide a priority to voice calls in cellular - only area limited fractional guard channel (LFG) policy is used. [Song 05]

Back to Table of Contents

---

6 Summary

Quality of Service plays a major role in cellular networks. In this report we have seen the various schemes for maintaining the QoS in cellular networks and each scheme has its own advantages and disadvantages. The Dynamic channel allocation scheme is fault tolerant and it provided a way to reuse the channels and also there was no need to get response from every neighbor for borrowing a channel. Therefore this scheme reduces the overall congestion in the network.

The CAC algorithm combined the advantages of two schemes namely pre request and the guard channel scheme and it proved to be much better than both these schemes for providing a better QoS for profiled users. In this techniques there was a pre - blocking feature where if the estimated load is greater than the accepted load then some fraction of the calls will be blocked even if there is channel for them.

In the Renegotiation scheme it was proved that when a lower priority class is admitted with a lesser bandwidth than what they had requested, they are given a much higher bandwidth when there was unused resources by the higher class services. On the other hand this scheme also guarantees the higher class services the bandwidth they are supposed to get. Thus it improves the overall QoS of the lower class services without affecting the higher class services.

It was also seen that iQoS was a very important parameter which has to be considered. It measures the overall user satisfaction for the full duration of the call and that is more important as some users QoS is buried in the averaging policy and so the real truth about users QoS is found by this technique. It was also seen that users who were not close to the BS were not getting good signal and the location of the users played a major role.

Finally we saw the integration of both WLAN and cellular networks and this is a scheme where in the mobility of the cellular network and the high data rate of the WLAN network are combined to provide an integrated network with high data rate for the users as well as good mobility. Issues such as QoS in IP is not assured and that is a major issue in this integrated network. Two different architectures namely Tight coupled and loosely coupled architecture were discussed. In a tight coupled scheme the WLAN technology is employed as a new radio access technology within the cellular system. The disadvantage of this scheme is that it is not scalable. In a loosely coupled architecture there is no direct link between the WLAN and the cellular networks. The advantage being that each network can be deployed independently and there is no bottleneck. The disadvantage is that it is difficult to guarantee end - to - end QoS.

Back to Table of Contents

References

[3GPP 23.107] 3GPP 23.107 V7.4.0,"Quality of Service (QoS) Concept and Architecture," June 2006

http://www.3gpp.org/ftp/Specs/html-info/23107.htm

[3GPP 23.207] 3GPP 23.207 V6.6.0,"End - to -End Quality of Service (QoS) Concept and Architecture," June 2006 http://www.3gpp.org/ftp/Specs/html-info/23207.htm

[Jain 06] Lecture Notes of Prof. Raj Jain in QoS, Apr 2006 http://www.cse.wustl.edu/~jain/cse574-06/j_9qos.htm

[Yang 03] Jianchang Yang, D.Manivannan and Mukesh Singhal, "A Fault Tolerant Dynamic Channel Allocation Scheme for Enhancing QoS in Cellular Networks," http://ieeexplore.ieee.org/iel5/8360/26341/01174857.pdf?tp=&arnumber=1174857&isnumber=26341

[Kovvuri1 03] Satya Kovvuri1, Vijoy Pandey2, Dipak Ghosal2, Biswanath Mukherjee2 and Dilip Sarkar1, "A Call-Admission Control (CAC) Algorithm for Providing Guaranteed QoS in Cellular Networks," http://www.springerlink.com/(rpophtf1vxmipbab1vtfyv45)/app/home/main.asp?referrer=default

[Soh 03] Wee-Seng Soh and Hyong S. Kim, "QoS Provisioning in Cellular Networks Based on Mobility Prediction Techniques," http://ieeexplore.ieee.org/iel5/35/26307/01166661.pdf?tp=&arnumber=1166661&isnumber=26307

[Monego 05] Hermes Irineu Del Monego1, Eliane Lucia Bodanese2, Luiz Nacamura Jr1, and Richard Demo Souza1, "A Dynamic Resource Allocation Scheme for Providing QoS in Packet-Switched Cellular Networks," http://springerlink.metapress.com/(xxolql45nijqxu55b4eq5345)/app/home/main.asp?referrer=default

[Medeisis 04] Kajackas A., Batkauskas V., Medeisis A , "Individual QoS Rating for Voice Services in Cellular Networks," http://ieeexplore.ieee.org/iel5/35/28971/01304238.pdf?tp=&arnumber=1304238&isnumber=28971

[Begain 04] Xin Gang Wang, John Mellor and Khalid Al-Begain, "Towards Providing QoS for Integrated Cellular and WLAN Networks," http://www.cms.livjm.ac.uk/pgnet2003/submissions/Paper-23.pdf

[Song 05] Wei Song, Hai Jiang, Weihua Zhuang, and Xuemin (Sherman) Shen, "Resource Management for QoS Support in Cellular/WLAN Inter working," http://ieeexplore.ieee.org/iel5/65/32333/01509947.pdf?tp=&arnumber=1509947&isnumber=32333

[Cotanis 03] Nicolae Cotanis, "QoS Estimation for Cellular Packet Data Networks," http://www.lcc.com/ABOUT/MEDIA/QoS%20Estimation%20for%20Cellular%20Packet%20Data%20Networks.doc.pdf

[Jain 01] "Issues in Emerging 4G Wireless Networks," http://www.ee.oulu.fi/~skidi/teaching/mobile_and_ubiquitous_multimedia_2002/issues_in_emerging_4G

[Sadeghi 04] Bahareh Sadeghi and Edward Knightly, "Architecture and Algorithms for Scalable Mobile Qos," http://portal.acm.org/citation.cfm?id=602414&coll=portal&dl=ACM&CFID=70683969&CFTOKEN=41794812&ret=1#Fulltext

[Nokia] "Multiplayer Game Performance over Cellular Networks v1.0"

http://www.forum.nokia.com/info/sw.nokia.com/id/b2b3781e-3dce-42dc-9e75-110e76b3e784/Multiplayer_Game_Performance_over_Cellular_Networks_v1_0_en.pdf.html

[Wu 03] "2.5G/3G Quality of Service - What has been Achieved so Far," http://www.touchbriefings.com/pdf/20/wire02_t_Ascom.pdf

[Yu 04] Fei Yu, Vincent W.S Wong and Victor C.M Leung, "A New QoS Provisioning Method for Adaptive Multimedia in Cellular Wireless Networks," http://www.ieee-infocom.org/2004/Papers/44_1.PDF

[Rachidi 05] T. Rachidi, M. Sebbane, A.Y. Elbatji and H. Bouzekri , "An Integrated System for QoS Provisioning in 3G WCDMA Cellular Networks," http://www.ctr.kcl.ac.uk/mwcn2005/Paper/C200516.pdf

"QoS in Integrated 3G Networks by Robert Lloyd-Evans, Artech House,2002"

Service Performance and Optimization in 2G/3G Networks: Providing End to End Quality of Service edited by Gerardo Gomez, Rafael Sanchez , John Wiley and Sons ,2005

End-to-End Quality of Service over Cellular Networks : Data Services Performance Optimization in 2G/3G by Gerardo Gomez and Rafael Sanchez (Hardcover - May 31, 2005),John Wiley & Sons

QoS in Integrated 3G Networks (Hardcover) by Robert Lloyd-Evans, 2002

Back to Table of Contents

List of Acronyms

3G- 3rd Generation
3GPP- 3rd Generation Partnership Project
CAC- Call Admission Control
DiffServ- Differentiated Service
iQoS- Individual QoS
IntServ- Integrated Service
FTP- Forced Termination Probability
GPS - Global Positioning System
HLR- Home Location Register
km- Kilometer
LAN- Local Area Networks
LFG- Limited Fractional Guard Channel
MAC- Medium Access Control
MH- Mobile Host
MMM- Mobility Management Module
MPLS-Multi-Protocol Label Switching
MSC- Mobile Switching Center
MSS- Mobile Service Station
MT - Mobile Terminal

MAN- Metropolitan Area Networks
Mb/s - Mega bits per second
PPM- Policy Provisioning Module
QoS- Quality of Service
RNC- Radio Network Controller
SAP- Service Access Point
SCCR- Successful Call Completion Rate
SDU- Service Data Unit
SLA- Service Level Agreements
TE- Terminal Equipment
UMTS- Universal Mobile Telecommunication System
WAN- Wide Area Networks
WLAN- Wireless Local Area Networks

Last Modified on Apr 20,2006 by Dushyanth Balasubramanian

Note: This paper is available on-line at http://www.cse.wustl.edu/~jain