albentia Systems

ALB-W-000001en

WHITE PAPER

White Paper

QoS Implementation in WiMAX Networks (*IEEE 802.16-2009*)

April 2010
Rev A3

*This document explains in detail an increasingly used term in data Networks: Quality of Service or **QoS**. It will explain its meaning, the reasons that justify its growing importance, and how a technology could implement these mechanisms.*

Afterwards, we will go deeper into the QoS mechanisms implemented in the IEEE 802.16-2009 standard (WiMAX): how every Base Station performs a deterministic resource allocation at Layer2 in accordance with the QoS agreement for each differentiated service for every user. The QoS performance in WiMAX is one of the most interesting strong points of this technology, so this document pretends to introduce this issue in detail.

INTRODUCTION

Origin of QoS

"QoS" is a incredibly extended idea nowadays, and it has become a necessary condition in almost any technology. Nevertheless, it is important to note that historically, it was not always like that.

Internet itself, for instance, was born in the beginning as a data transmission system that followed a "*Best Effort*" (BE) philosophy. This means literally "*do it as best you can*".

BE philosophy perfectly fitted Internet's requirements, so designing advanced QoS mechanisms was not necessary. Every packet was treated in the same way, sharing network's congestion. The target was simple: packets should reach its destination, without ensuring delivery, with no minimum bit-rate required, latency,...

Despite this does not seem *a priori* the best implementation, BE philosophy perfectly fitted many scenarios due to its extreme simplicity, and it took advantage of the IP success, its greatest exponent.

Nevertheless, the system that was born as an experimental network for data exchange, has become during the last years in a technology at its very peak, and there is an increasing interest in services and applications running over IP technology.

Besides, social needs are changing: while some years ago a customer only asked to its operator for an acceptable Internet connection, nowadays customers demand a high quality broadband connection, including *crystal-clear* voice services and IPTV with multimedia contents.

The problem is that these "*new*" services over IP might be latency-sensitive, or might not accept packet discarding, for example. Because of this, it is mandatory to include additional mechanisms that allow guaranteed packet transmission. This is how the need of QoS mechanisms appears.

Definition of QoS

QoS *(Quality of Service)* is a kind of ambiguous idea that may have different meanings, depending on the area of application. QoS, if literally translated, seems to mean "*a good-quality service*", but this definition is not precise enough.

In our concerning field, communication networks and more specifically wireless *IEEE 802.16-2009* (WiMAX) technology, a good definition might be:

> *QoS guarantee is the guarantee of a Service Level Agreement between a provider and a user of a service.*

This paragraph may be shorted in the sentence: "**QoS is the guarantee of a SLA**". A SLA defines the relationship between a service Provider and a service Customer in a formal way. It is a contract where the required minimum service parameters regarding availability, performance, charging, etc... are established. It is the commitment of meeting the minimum required service levels, considering penalizations if the contract is not met. In conclusion, a SLA is a table specifying the parameters the operator will guarantee to the customer, and the agreed levels (1Mbps, 30ms…).

In WiMAX, communication is performed using differentiated service flows, and the operator is able to establish the minimum levels in some parameters: download bit-rate, maximum delay, jitter,… This is the **most important idea** in QoS: there are some measurable parameters the customer is going to pay for, and the operator may and should be able to provide and guarantee them.

In conclusion, when talking about systems, standards or protocols that guarantee QoS, it means that they have the technological resources that allow, in a deterministic way and without any doubt, to guarantee the service levels that are established in the "contract" between operator and final customer.

Importance of QoS

The use of technologies with advanced QoS mechanisms is a very advantageous choice for anybody who wants to provide (generally, charging for it) a telecommunication service, for several reasons:

- **SLA performance guarantee**. QoS allows the operators to guarantee the minimum levels (throughput, latency, ...) for its customers in the service they have paid for, being sure that the technology will be able to cope. Failure to comply with the SLA may entail important consequences to the operator, what may be avoided with the right technology.
- **Satisfactory user experience**. Obviously, if a customer contracts a service and the operator is able to provide it properly, this supposes a satisfied customer, which is the goal of any services' provider.
- **User differentiation, Services and Pricing Policies.** Implementation of QoS allows providing different priorities to the users and applications. Customers that have critical traffic (i.e. companies) may be willing to pay more for a higher priority service or benefits, while there are other profiles that do not require such demanding conditions (i.e. particular users). QoS mechanisms allow operators to have a distinction between users and/or services, and thus it is possible to establish different charging depending on the provided features.
- **Transmission of demanding applications**. In any telecommunications network, there are many causes that may affect the data packets transmission: delay, jitter, errors,… This is not a critical problem to some applications, or it may be solved by higher protocols. Nevertheless, in some other cases it is not enough to access the information only "if possible" (BE service), and some minimum quality levels should be guaranteed. Is in these scenarios where QoS plays a fundamental role. Some of these applications are:
  - **Video over IP,** or any type of real-time multimedia streaming (**IPTV**, multicast video,...). It requires mainly a minimum sustained throughput and a minimum packet lost rate, to avoid errors in the video (pixels, interruptions,...).
  - **Voice over IP** and videoconferences (**VTC**, ...). These applications does not usually require too much throughput but low levels of jitter, because voice must be correctly reproduced on its destination.
  - **Online games.** In this case playing fluid is very important, so low latencies to the gaming server are required.

Field of Application

¿Where is QoS applied? Traditionally, QoS in networks refers to the quality that networks are able to provide, as a whole. In a typical packet network like ATM, when talking about QoS, it is referred to the parameters to be ensured between final users (end-to-end QoS). However, this document is about QoS applied to a specific scenario: **wireless communications** (air interface). It is not an end-to-end QoS, is a QoS in a particular interface.

QoS makes sense if any of next circumstances is given. The more of them are appliable, the more important becomes ensuring QoS:

⇒ Maximum net capacity in the interface is lower than the network connection capacity.
⇒ Multiple users accessing the interface (Point to Multipoint—PtmP interface).
⇒ Multiple services with different requirements over the interface.

QoS wireless environments

Generally speaking, radio interface meets any of the previously described circumstances (and usually all of them). In a broadband wireless access, multiple users may be served, multiple services may be required, and normally, radio interface capacity (some Mbps) is much lower than network connection capacity (tens of Mbps). It may be concluded that QoS mechanisms are specially important in wireless networks due to:

- **Unstable radio interface**: instantaneous capacity offered by wireless networks is variable as a function of medium conditions. The transmission time required for a byte depends on the used modulation and codification. These are dynamically adjusted to the medium conditions, which supposes an additional difficulty for QoS mechanisms (they must know the medium capacity constantly). This does not happen in wired networks, that have generally a high fixed capacity and a low error probability.
- **Available spectrum reduced**: radio spectrum is limited and over-provisioning resources is not viable.
- **Control over transmission direction**: in an ATDD (*Adaptive Time Division Duplexing*) wireless system with centralized QoS classification (like a BS controlling QoS in both directions), *Scheduler* must also decide frame by frame how UL and DL channels are going to be distributed.
- **Receiver's capacity to handle high error rates**: if due to a fading phenomenon, a transmission is not able to reach the receiver unit, requesting an immediate data retransmission is not convenient, because there is a high probability that it will fail again. Besides, a high re-transmission rate to a particular receiver not only increases the average latency to that unit, but it affects other units' QoS in the network. QoS classifying mechanisms will have the hard challenge of balancing the throughput and the use of the radio electric medium for keeping the system in balance.

HOW TO ENSURE QoS

Introduction

A priori, the simplest way to provide QoS in a network is to over-provision the available resources, ensuring that all needs will be always attended, even in the worst case (with the maximum achievable traffic). For instance, if a few computers with low throughput needs are interconnected using a wired *Gigabit-Ethernet* network, it is quite probable that it will not be necessary to implement any QoS mechanism, as long as the network itself will be enough (the available resources will be higher than the current needs).

However, as long as the host number or the traffic needs increase, over-dimensioning becomes completely unviable, as it supposes really elevate costs that operators and network administrators are not ready to cope with. Besides, in wireless technologies the available radio spectrum is really short, so it is not possible to over-allocate it infinitely, so some other solutions must be found for being able to guarantee QoS.

A more realistic approach to the QoS is based on a static resource allocation. This is the way used in *circuit-switching* systems to guarantee QoS: every *peer-to-peer* connection is established using a dedicated communication channel called "circuit", allocating specific resources (for example in a traditional phone call). However, this practice is not possible in *packet-switching* systems. The challenge in this kind of networks is to implement adaptable QoS techniques capable of using the available resources in the most efficient way.

Today, the most used model for guaranteeing QoS in a network or interface is the **Differentiated Services** model (*DiffServ*). In this model data packets include a *tag* which indicates the transported high-level application, and with this information operators are able to configure the intermediate network nodes with different routing policies for each service type. Thus, a bigger priority may be given for packets that require the minimum possible latency, for instance.

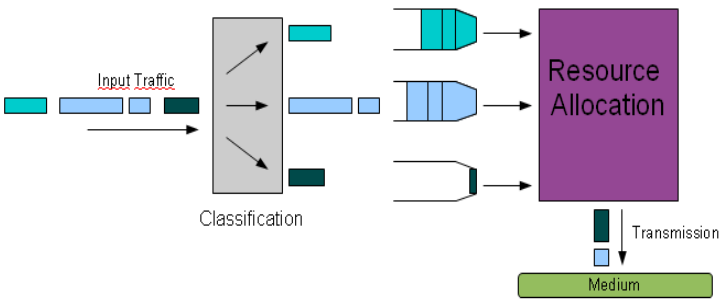

Figure 1 - Simplified QoS model

Basic implementation

Figure 1 shows a simplified but useful scheme to understand better how the QoS mechanisms work. Any QoS mechanism should implement these two essential procedures:

a) CLASSIFICATION: when a node that implements QoS control receives a higher traffic rate than the amount it is able to transmit, it is obvious that it should decide which data packets should be first attended. A simple decision could be to transmit first the most important packets, queuing or even discarding the less important ones. But... ¿how may that host know the *importance* of each packet?

Before transmitting, this hosts should classify all the incoming traffic, so later different priority criteria may be applied to each traffic type (which should be previously defined by the network administrator).

There are many traffic classification criteria. They may be up to Layer 2 (i.e. MAC address), Layer 3 (i.e. IP address, TOS/DSCP field,...) or Layer 4 (source/ destination port, transport protocol,...). The main goal of this filtering rule is to be able to divide a unique inbound traffic into several differentiated and smaller flows, which may be associated to specific QoS mechanisms. The node performing the classification process will also have information about the QoS patterns and contracts, being able to give more or less priority to the different classified data flows and services.

b) RESOURCE ALLOCATION: once the inbound traffic has been classified, and considering that some QoS parameters have been defined by the network operator, next step would be to make the resource allocation in the current interface. The node should transmit the packets to the medium (wired or wireless) in the correct order. The part responsible of performing the allocation is usually known as the *Scheduler*, an essential component of any QoS mechanism, which makes the bandwidth allocation for every service flow.

In a system with no QoS support, all the inbound traffic from all the applications is treated as a FIFO queue (*First-In, First-Out*). On the other hand, in a system with QoS the Scheduler must ensure that traffic is served following the QoS rules and not the entrance order, allowing that a maximum-priority service is not affected when the overall network traffic load is increased.

The Classification phase is common to all technologies that implement any kind of QoS, while the main difference comes with Resource Allocation phase. There are two mechanisms so general that could receive a name: "Layer 3 QoS (L3QoS or IPQoS)" and "Layer 2 QoS (L2QoS or MACQoS)". The differences between them will be explained in the following points.

QoS AT MAC/IP LAYER

L3QoS: IP Level QoS

The techniques used in this kind of QoS mechanisms are the ones used by the traditional *Traffic Shapers* (TS). A TS classifies first all the inbound traffic following some pre-defined rules and QoS contracts. Thus, the **Classification** phase is performed exactly in the same way as explained in the previous section, dividing a common inbound traffic flow in several thinner data flows. The main difference comes with the **Resource Allocation** phase.

Once the traffic has been classified, the TS statistically assigns transmission resources in the medium. For example, if the queue of a low-latency service is getting full, it will try to empty it as soon as possible, or if the queue of a service with a minimum sustained traffic starts storing packets, it will try to maintain a transmission bit rate equal or higher to that minimum guaranteed rate.

These Layer-3 QoS techniques (sometimes also called "*IP Level QoS*"), are based in priority queues which are associated to DSCP or TOS fields in the IP header, for instance.

The main problem of L3QoS is that it is impossible to exactly know the capacity and availability of the transmission medium. In radio scenarios, the maximum gross throughput towards every CPE is very variable as it depends on the medium conditions (interferences, fading, non-line-of-sight, obstacles, reflections,…). Using L3QoS techniques in these scenarios, where the available throughput towards every user cannot be known every time, becomes really inefficient: it is not possible to guarantee QoS in absolute terms, but in relative ones.

This means that if for instance, two services of 1Mbps and 2Mbps have been defined, a L3QoS system can only guarantee that the traffic rate in the second one will double the traffic on the first one, but cannot guarantee the exact value of minimum guaranteed traffic, as it has not information about the medium state and availability.

This problem gets worse when the Medium Access is controlled using Contention mechanisms (such as in Wi-Fi, Ethernet,…). In these technologies the medium use itself is statistic, and even the Layer-2 is not capable of knowing exactly the medium availability. In the worst case of a nearly saturated network, it is not possible to determine when a new packet is going to be transmitted, as the collision possibility will be really high.

L2QoS: MAC Level QoS

When the Resource Allocation is performed at Layer 2, the system that assigns *time slots* and that decides the packet transmission rates knows every moment the gross capacity in the medium, so it also knows the net throughput available for every user. This deep knowledge allows to implement algorithms that may guarantee the traffic assignation in an absolute way.

WiMAX technology, for instance, is a L2QoS system. The Base Station (BS) is the network Master node, and decides the data transmission both towards the users (*Downlink*) and from the users (*Uplink*). The presence of a node that acts as an arbitrator allows to avoid the Medium-Access contention, guaranteeing that the BS knows every time the radio medium availability. Besides, a WiMAX BS has knowledge about all the connected CPEs, about their radio signal quality and modulations, and in conclusion, about the state of the Physical Layer, so it can assign in a completely deterministic way the traffic distribution both in Downlink and Uplink.

Obviously, the Layer-2 QoS mechanisms are not unique in WiMAX. DVB-RCS, for example, is a standard protocol for satellite multiple-access communications with a similar scheme: a central node that assigns traffic, an exhaustive knowledge of the radio medium, and in conclusion, a QoS that may be completely guaranteed.

L3QoS

- IP Level (Layer 3) does not have the knowledge about the Physical medium (Layer 1) every time
- The packet transmission is based on relative (not absolute) priorities
- Statistic resource allocation

L2QoS

- Mac Level (Layer 3) has a complete knowledge about the Physical medium (Layer 1) every time
- Deterministic resource allocation

QoS IN WIMAX

Medium Access

As explained before, the *IEEE 802.16-2009* standard supports up to Layer-2 QoS levels. This is possible thanks to its extremely efficient Medium Access Control (MAC) Layer, the really strong point of WiMAX. In other technologies, the Medium Access is random and it is performed using contention mechanisms: the hosts directly compete for using the medium. With this policy, in situations with high traffic rates or with a big number of hosts, the collisions will be very frequent, so the channel usage will not be efficient and the total achievable net throughput will become dramatically decreased.

As opposed to this random technologies, in WiMAX the Base Stations act like an "*arbitrator*", controlling the medium access of all the connected subscriber stations (CPEs). It is a framed technology, and the BS will assign to every registered user some transmission opportunities in the frame (*time slots*) using TDMA multiplexing, so they are able to transmit without competing between them. It uses a completely deterministic Medium Access where the BS always knows the radio modulation of every single connected user. Having a complete knowledge about the radio medium allows to allocate time slots following the configured QoS criteria. That's why it is considered a L2QoS, as the BS allocates resources at layer 2.

IEEE 802.16-2009 System

- **Framed**: TDD and TDMA (*time slots*)
- **Centralized**: BS acting as "arbitrator", controlling all transmissions
- **Deterministic MAC Layer**: no Collisions, no Contention
- **Total knowledge about the radio medium:** the BS always knows the state of the radio medium and the traffic demands.

Service Flows

WiMAX transmissions are based on the "Service Flow" concept. These flows are individual and unidirectional data connections that may be established between BS and CPE. They may be viewed as "*logical pipes*" where data packets are inserted, flowing peer-to-peer. Figure 2 in the right side of the page shows a sample diagram of a BS with two connected CPEs and with some different Service Flows.

Data packets cannot be transmitted to the air if they are not inserted in a valid Service Flow. The idea is that every CPE will have different Service Flows with different sizes and QoS policies, created for transmitting traffic from different applications such as video, data, voice, etc.

When a CPE is connected to a WiMAX cell, the BS will create the Service Flows specified by the Operator for this CPE in the provisioning Database. Every Service Flow will include the following independent properties:

- **QoS contract:** it defines the specific features of that flow, and is stored in the provisioning database of the BS. It specifies parameters such as the maximum achievable binary rate, minimum guaranteed rate, type of service (BE, UGS,…), maximum latency,… In conclusion, a sort of parameters that define the features of every flow (its size, priority,…).
- **Filtering policy**: it includes a sort of rules that allow to determine which kind of data packets will be inserted to each Service Flow. They are responsible of the "packet classification phase" that must be performed in any kind of QoS procedure, as explained before.

In conclusion: WiMAX supports differentiated QoS for multiple users and multiple Service Flows, using these flows to perform the two essential tasks in any QoS mechanism: **Traffic Classification**, and **Resource Allocation**. Each flow will be used to separate traffic from different applications (VoIP, video, web browsing,…) depending on the type of service specified in the contract that has been defined by the operator for every user and service. Using Service Flows is a distinguishing element of this technology and is absolutely necessary for offering QoS guarantees at operator level.

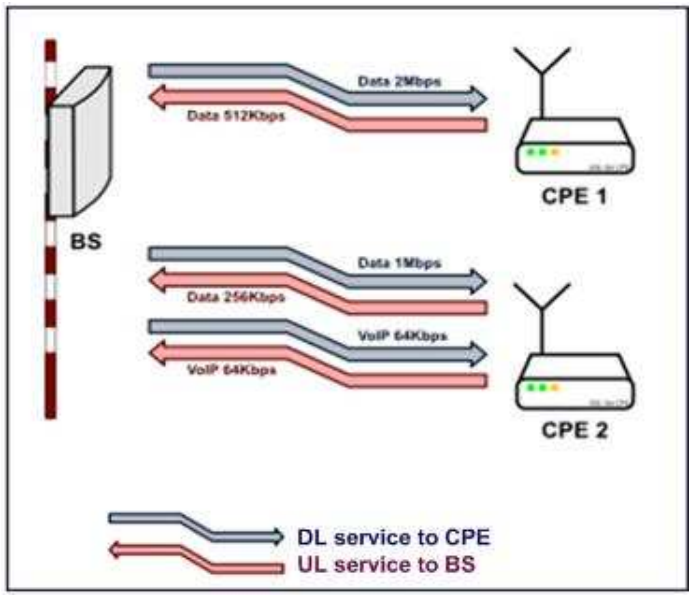

Figure 2 - Service flows over *802.16-2009*

Types of Service

IEEE 802.16-2009 standard defines up to 5 different types of services depending on the type of QoS they provide. Even if the standard defines them only for the Uplink sense, they are all usable both in UL and DL. These 5 types may be sorted into three main groups:

- **BE** (*Best Effort*): this type offers a good performance for "Best Effort" traffic, and is usually used for those data services that do not require special guarantees of QoS. These services are served depending on the cell availability, and always after services with a higher priority. They are designed to support this model of binary rates:

  *Maximum Sustained Traffic = **X***
  Minimum Guaranteed Traffic = 0

- x**RTPS**: this class includes 3 different types of service with similar features: **RTPS** (*Real Time Polling Service*), **NRTPS** (*Non Real Time Polling Service) and* **eRTPS** (*Extended Real Time Polling Service*). They are mainly differentiated by the latencies to the final user, and are oriented to those applications that require a minimum guaranteed throughput. They are a perfect choice for applications such as VoIP. When a CPE needs to transmit a packet belonging to a xRTPS service, the BS will give him preference against BE flows. They are designed to support this model of binary rates:

  *Maximum Sustained Traffic = **Y***
  *Minimum Guaranteed Traffic = **X** (X<Y)*

- **UGS** (*Unsolicited Grant Service*): a perfect type for constant bit rate applications which must be guaranteed, such as constant video transmission or E1/T1 frame transmission. When this kind of service flow is assigned to a CPE, the BS will **always** allocate a certain time in the WiMAX frame for it, even if the CPE has nothing to transmit. When it does, it will put the information on the pre-reserved *time slots*, no needing to send a bandwidth request to the BS. UGS is the type that offers the best QoS by means of continuously reserving time slots in the frame, so it is convenient to correctly calculate the number of services of this type so not to reduce too much the aggregated throughput for the rest of the users. They are designed to support this model of binary rates:

  *Maximum Sustained Traffic = **Y***
  *Minimum Guaranteed Traffic = **Y***

Except of in the UGS case, a CPE must always send a *Bandwidth Request* to the BS for transmitting, located in an specific place of the WiMAX frame. When the BS receives this message, it will assign to the CPE the maximum available throughput, never exceeding the "*Maximum Sustained Traffic*" specified for the Service Flow, and serving before the Service Flow with a higher QoS type.

Benefits of WiMAX Services

A Service-Flow-based technology is very efficient assigning resources in a shared radio medium, and in addition it gives absolute control over the bandwidth usage, avoiding the misuse or simply controlling that the traffic data of a CPE is not affecting its own VoIP conversation, for instance. This means a better browsing experience for users, the possibility of using differentiated services, and an enhanced ability for the network operator to provide multiple service levels and value added applications.

For establishing different service flows associated to different commercial products, operator should:

- Clearly define the offered products: data or voice, low, medium or high capacity, with or without QoS guarantees, ...
- Define the WiMAX services that will be matched to the previously defined commercial products: QoS types, priority, UL and DL rates,...
- Create a provisioning system which includes different Service Flows to every user according to the contracted commercial products.

Some of the benefits obtained by both operator an customer are the following:

- The operator increases the spectrum usage efficiency, reducing the bad use and the over-provisioning, and in conclusion, being able of increasing the potential users in every cell.
- The customer gets a high-quality service and fully compliant with the contracted service levels.
- The customer gets a better experience than in any other wireless technology, and even better than in some wired systems.
- It allows to offer a new range of services: *carrier-class* quality VoIP is now possible, as a service flow differentiation makes possible to give the maximum priority to IP phone calls, avoiding that data traffic in the cell affects the voice quality.

In conclusion: WiMAX services allow the operator to offer a wide range of last-mile commercial products (QoS, data and voice differentiation,…), but with the advantages of a wireless technology (fast deployment, cost-effective, scalable,…).

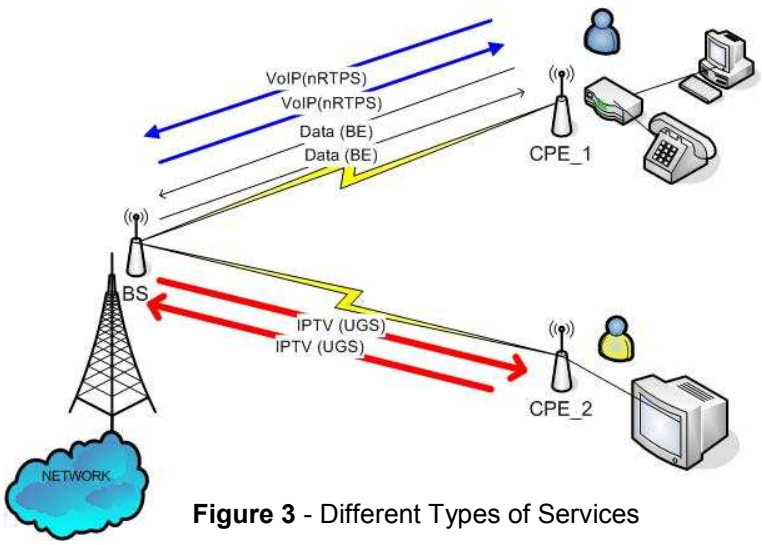

Figure 3 - Different Types of Services

CONCLUSIONS

In the following, the most important points that have been discussed in this document will be listed:

1) QoS Definition

*QoS term may be used for defining those mechanisms that are able to guarantee minimum performance levels in a communication, establishing a **contract** and a "Service Level Agreement" between Provider and Customer that the Provider must always meet, using for that propose a technology capable of guaranteeing it.*

2) Need of QoS

Being capable of providing QoS is essential for Service Providers, as it allows them to guarantee minimum quality levels and to create different service levels with different pricing policies. Besides, nowadays many common IP-applications require, by nature, an strict control of QoS: VoIP, video surveillance, IPTV,...

3) QoS in wireless networks

Wireless communications are becoming very popular and increasingly used, so they must be ready to offer a wide range of services and applications, including those with QoS needs. It is quite difficult for a wireless technology to succeed if it is not prepared to include strict and efficient QoS mechanisms and similar features than the ones in more traditional systems.

However, these technologies have by nature more complications to perform QoS: an unstable propagation medium with a higher error rate and with a limited spectrum means an additional challenge for QoS in wireless technologies, specially in NLOS scenarios.

4) QoS in WiMAX

QoS in WiMAX is based on the following points:

- ***MAC Layer**: the Medium Access is the key element of this technology. With a framed structure, high spectral efficiency, and a BS that acts as "arbitrator" and that manages the radio medium, WiMAX allows to implement **on a deterministic way** any QoS mechanism.*
- ***Layer2QoS**: one of the strong points of WiMAX is that it implements up-to-Layer2 QoS mechanisms, which allow to offer differentiated services as the BS always knows the state of the radio medium (it controls the modulations for all users), being able to allocate the necessary resources every moment.*
- ***Service Flows:** data packets flowing in the air must be transported by Service Flows. These flows are unidirectional and individual for every CPE. WiMAX technology is based on using differentiated flows for differentiating applications.*
- ***Types of Service**: there are up to 5 different service types in WiMAX depending on the desired priority level.*

In conclusion, nowadays QoS mechanisms are essential for telecommunications, and wireless broadband systems based on the 802.16-2009 standard (WiMAX) are the most suitable ones for offering all the necessary QoS mechanisms, allowing operators and WISPs to directly map their commercial services into over-the-air transmission services, with all the service compliance guarantees.