Introduction to QoS

There are some definitions for Quality of Service: the one consider is that QoS is a term to illustrate performaces from the user's point of view. In order to do this some indices are introduced:

. speed, throughput, bit rate, bandwidth (for all measurements is important specify precisely how long is time used in that measures);

. delay: average, percentile, maximum (worst case), variance, jitter (delay variability);

. loss probability;

. error probability.

These indices concern packets, as it was said before, in particular from the user point of view. Other indices are:

. blocking probability (if network has not enough resources communication is refused);

. fault probability (also called avaiability; this index is correlate with blocking probability because if a call is refused network is not avaiable);

. recovery time after a fault (to guaratee trasparent performances must be the smaller possible);

. others such as time needed to open a connection, costs and tarifs ...)

These indices describe how much efficient is the network because if there are enough resources blocking or fault probability are negligible or null, but if the condition is not verified users detect faults or are blocked.

Different services requires different indices of quality: in a scenario where only one service is provided, QoS parameters are very well satisfied, but in an heterogeneous environment this is not true. Moreover, also bursty traffic is critical in order to guarantee quality of service parameters.

User traffic characterization

The two groups in which user traffic can be divided are:

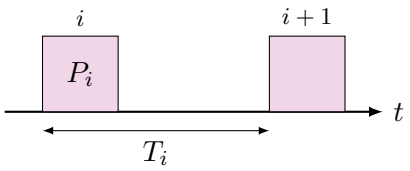
. costant bit rate sources (CBR);

. variable bit rate sources (VBR).

Costant bit rate sources

This kind of sources are not used in reality but is possible to force a variable source to generate traffic as a costant bit rate source: the most important CBR source is the phone call but, human voice is not a costant source (silence periods); to collect samples of a phone call, human voice is sampled every 125 μs: in this way a phone call generates costant bit rate.

Graphically, a costant bit rate source can be represented:



where:

. T_i is the inter packet generation time;

. P_i is the packet size.

If P_i is fixed, the costant bit rate source has:

$$T_i = T \qquad \forall i$$

with T fixed period.

Is possible to define the bit rate as:

$$R_b = \frac{P_i}{T_i}$$

Since P_i and T_i are fixed, the bit rate R_b is constant. This feature is very important because due to that fact the traffic is predictable, therefore it is possible to provide QoS parameters easily. Although there is perfect knowledge about the start of a call, how does long the call takes or in which moment the call is done are two parameters unknown and they can only be studied and modelized statistically (thanks to a stochastic approach).

Variable bit rate sources

Variable bit rate sources are the most present sources nowadays and they have variable bit rate:

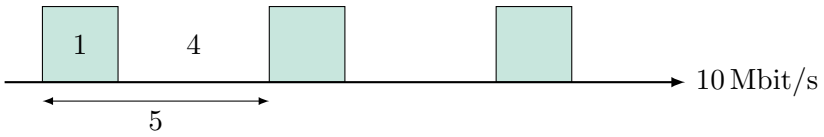
$$R_b(t) = \frac{P_i}{T_i(t)}$$

This fact is due to the variable inter packet generation time.

In order to be characterized, VBR sources need more parameters than CBR sources, such as average rate, peak rate and burstiness. The more knowledge is possible have, the controll on the network will be higher. But it is not possible to collect lots of parameters otherwise complexity will be so higher that management will be impossible.

Examples In the following examples parameters cited before will be discussed.

Example 1



all parameters are normalized.

. Average rate:

$$\frac{1}{5} \cdot 10\,\text{Mbit/s} = 2\,\text{Mbit/s}$$

. Istantaneous (peak) rate:

$$\frac{1}{5} \cdot 10\,\text{Mbit/s} = 2\,\text{Mbit/s}$$

This is a CBR source because average and peak rate are equal. The peak rate is the rate in which are trasmitted two consecutives packets (silence periods are not considered) and it is always equal or greater than the average rate.

Example 2

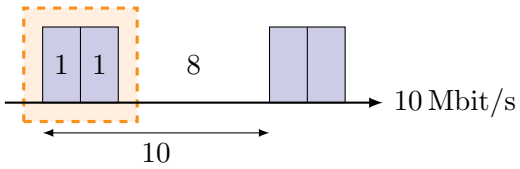
all parameters are always normalized.

. Average rate:

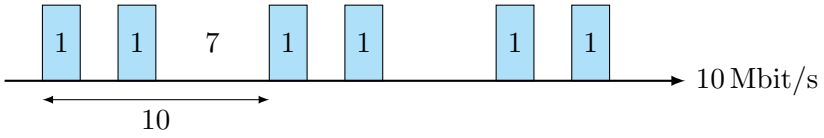
$$\frac{2}{10} \cdot 10\,\text{Mbit/s} = 2\,\text{Mbit/s}$$

. Istantaneous (peak) rate: 10 Mbit/s

The peak rate is at the highest value of speed because two consecutives packets are not separated by a silence period as shown in **example 1**:



Example 3



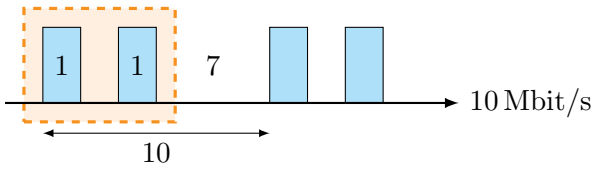
. Average rate:

$$\frac{2}{10} \cdot 10\,\text{Mbit/s} = 2\,\text{Mbit/s}$$

. Istantaneous (peak) rate:

$$\frac{1}{2} \cdot 10\,\text{Mbit/s} = 5\,\text{Mbit/s}$$

In this example the peak rate is identified by packets shown in the following picture:



Burstiness The relation between average and peak rate is called burstiness: it describe how long is possible send packets at link speed. In formula:

$$Burstiness = \frac{Peak\,rate}{Average\,rate}$$

This parameter is critical because small is burstiness easier is the control: performances of a source who transmitt for 1 second at the link speed and for another 1 second is silent are very different from another source who transmitt only for 0.1 second at link speed and then is silent until the period of 2 second elapsed. Notice that, in the long-term, the transmission speed will be the average rate, so a short transmission with an high speed is balanced with a long silence.

Networks examples

In this section some practical examples will be discussed; they are:

. telephone network: POTS;

. Internet;

. B-ISDN (considering the end to end scenario).

POTS

Fixed telephone network is a CBR source, so traffic characterization is completly known, which provides a circuit switching service (no congestion problem, no store & forward delay, no loss probability). In general blocking probability is negligible, a part from some particular events in which a lot of users want to access to network; this property, instead, is not true for mobile phone network (for example handover fail if ther are not enough resources, or directly a call refuse).

When the communication is started, QoS parameters are always guaranteed because each call is completly independet from others. From the user point of view is good because his satisfaction is high but is possible that network resources are lowly used.

Internet

This is a VBR source in which user's behavior si completly unknown and the service provided is packet switching (datagram): it means that all network resources are shared among users, bit rate and delay are unpredictable, loss probability is significant while error probability is not relevant and blocking probability is null. Congestion problem is possible and it is dangerous from the user point of view (so also for QoS) but from network point of view is good: the implication is that network resources are well used.

QoS parameters are difficult to provide because datagram service is not able to recognize flows among the network and they are largely dependent form the behavior of users, which satisfation can be, sometimes, very low.

B-ISDN

This is the intermediate situation between POTS and Internet: infact the source's behavior is know (at least statistically) and the service provided is packet switching with virtual circuit. It means that network is able to recognize flows since packets are forced to follow a circuit.

Some parameters are negotiable such as bit rate, delay and loss probability is negligible as blocking ad error probability.

QoS depends either form user's behavior and algorithms used to manage network resources and the two parameters, user satisfaction and network utilization, are traded.

Design to obtain QoS

Designers have to trade different purposes given features like:

. topology of the network: number of nodes, link speed;

. traffic and user characterization.

All resources are well known and defined; the **two purposes** to reach are:

. high network utilization;

. QoS to each user connection.

Taken individually, these purposes are trivial and there already are several optimal solutions; it is very critical trade to obtain good performances. One solution is *overprovisioning* which disadvantage is the very high cost.

There are lot of parameters that designers have to take into consideration to reach the two purposes before mentioned:

. time scale;

. network design and planning (how resources are distributed) based on:

  . traffic estimates;

  . cost constraints;

  . algorithms, routing criteria and traffic engineering;

. network management (how the network runs) based on:

  . measurements made on different time scale (if the service is circuit switching measure are not very important but for datagram service they are fundamental);

  . fault and recovery management (two different phases: protection and restoration);

. connection management;

. data unit transport.

QoS parameters are studied assuming that the network has already been designed, is properly managed and have always avaiable resources; with these conditions algorithms for traffic control problem are taken into consideration. The two main families of control are:

. preventive control: this idea was developed from the phone network, so try to avoid completly congestion (low utiliziation of the network);

. reactive control: idea developed from internet network react only when congestion problem occur (high utilization of the network).

Traffic control

The essential elements in order to control traffic in network are:

. possibity of detecting flows, so the approach should be connection oriented;

. declaring an interface, some preliminaries information in which user and network are able to negotiate parameters (traffic characterization and QoS parameters);

. resource allocation (bit rate, buffers size).

There are some algorithms implemented in order to have traffic control:

. CAC (Connection Admission Control) and routing;

. scheduling (performing decision about priority packets) and buffer management;

. confermance verification (policing);

. traffic shaping;

. congestion control.

Connection oriented network allows to know on over kind of network algorithms have to be performed: if the purpose is user satisfaction services provided by network will be circuit switching or packet switching with virtual circuit (these two tecniques allows largely QoS); instead, if the purpose is the high utilization of the network datagram service will be implemented.

The higher is the knowledge of traffic network the higher will be the control, if complexity does not increases much. QoS parameters can be negotiate in different ways:

. call basis;

. contract basis.

The first approach allows to change *on line* parameters while the second method is negotiated at first (although is possible re-negotiation). In principal network the negotiation is:

. POTS: contract basis;

. Internet: no negotiation;

. B-ISDN: both contract and call basis;

. Internet extended (with QoS): negotiation through SLA (Service Level Agreement).

Algorithms: routing and CAC The routing decisions are taken based on the choice of the best path, where the best path very often is the shortest path. Notice that, from the QoS point of view, not always the shortest path it is the right choice. Imagining a scenario in which chosing the best path have like conseguence not fixed delay; probably it is better choose another path (longer than the shortest one) which have fixed delay guarantees.

CAC determines if a connection can be opened or not. First the routing decision is taken, then CAC checks if traffic required can be accepted and if QoS parameters can be guaranteed; after all network status is checked. In any cases, when some QoS parameters are already guaranteed, it is not possible accepted a new communication reducing quality. Moreover, connections are immediatly refused if network can be overloaded or congestioned.

These algorithms are mainly preventive, but can become reactive.

Algorithms: scheduling and buffer management Scheduling algorithms have to choose which is the data unit to be transmitted among all data unit stored in the buffer.

Buffer can be managed in different ways: allocation may be exclusive or shared, partial or total (principally buffer are not completly fully: data unit are accepted until about the 95% of the total size). When resources are over dropping policies have to be implemented. If traffic is homogenous one method can be FIFO (if there are packets avaible they are served time by time) but in heterogeous environment this is not true (some queue can be fully and others no, so globally there is not good utilization).

These algorithms are both preventive and reactive.

Algorithms: policing and shaping Policing is traffic verification: network controls the behavior of users in order to have a check of traffic characterization.

Shaping algorithm condition the characterization of the traffic: adapt it in order to be conformant to a given previous negotiation between users and network.

These algorithms are mainly preventive, but they can become reactive in a special case: if QoS level change over time.

Algorithms: congestion control Congestion is a traffic excess over a channel. From the network point of view is good because links are always fully so network utilization is high, but for users is very dangerous expecially because is not possible provide well QoS parameters in presence of this problem.

There are two possibilies that can generate congestion:

. short term traffic variability that in not the most critical;

. allocation policies that share resources in order to increase network utilization: this is the worst case from user's point of view.

Congestions effects are mainly:

. increasing buffer occupancy;

. increasing of delays;

. increasing the quantity of data lost.

This algorithm is reactive: it operates only when congestion problem is present.