

# Quality of Service in the Tellabs® 8600 Managed Edge System

Deploy IP/MPLS technology at all network provisioning layers to unify your infrastructure and save costs.

## Executive Summary

Today's telecommunications users require a variety of services which have differing QoS requirements in terms of delay, jitter, packet loss and bandwidth. IP/MPLS technology is ideal for delivering such a wide array of services with varying QoS requirements as it supports both guaranteed and best-effort services. This technical solution description discusses the QoS functionalities of the Tellabs 8600 managed edge system.

In this state-of-the-art IP/MPLS router, all the QoS parameters can be controlled using a user-friendly network management system that enables large scale deployment of services. Using the latest semiconductor technology enables a cost-efficient implementation of all the QoS functionality in hardware, bringing down the overall cost of equipment. This enables service providers to deploy IP/MPLS technology at all layers of their provisioning network: core, regional and access. Such a unified network infrastructure benefits the service provider with major savings in both capital and operational expenditures.

## Introduction

Connectivity services can be grouped into four main service classes based on the QoS requirements of the end user applications as shown in Figure 1.

These service classes have contrasting QoS requirements as follows:

### 1. Real Time

Voice and multimedia applications are intolerant to delay and jitter. Consequently they place the strictest performance requirements on the network.

### 2. Priority Data

Interactive or transactional business applications such as database access, enterprise resource planning (ERP) and customer relationship management (CRM) systems require controlled maximum delay to ensure a predictable performance for all users.

### 3. Business Data

Common intranet services such as file and print servers require guaranteed performance in terms of packet loss and data integrity.

### 4. Best Effort

The best-effort service class typically does not provide any performance guarantees. However, a reasonable performance in terms of delay and packet loss is normally expected. The main reason for using the best-effort service class is for cost-efficient bandwidth applications such as Internet access and e-mail services.

	Service Class			
	Real Time	Priority Data	Business Data	Best Effort
Delay (ms)	<25	<50	N/A	N/A
Jitter (ms)	<10	<20	N/A	N/A
Loss	<0,1%	<0,1%	<1%	N/A

Note: Delay and jitter one way

Table 1. Example of service class QoS characteristics.

Table 1 shows an illustrative example of the QoS characteristics of these service classes. The actual values will depend on the network topology (i.e. geographical distances and the number of network elements), the speed of the network links and the amount of traffic loading on each of these links.

Traditionally these services have been provided using separate networks in order to meet the different performance and cost requirements. Today the latest IP/MPLS technology enables provision of services with different QoS requirements over a single converged network.

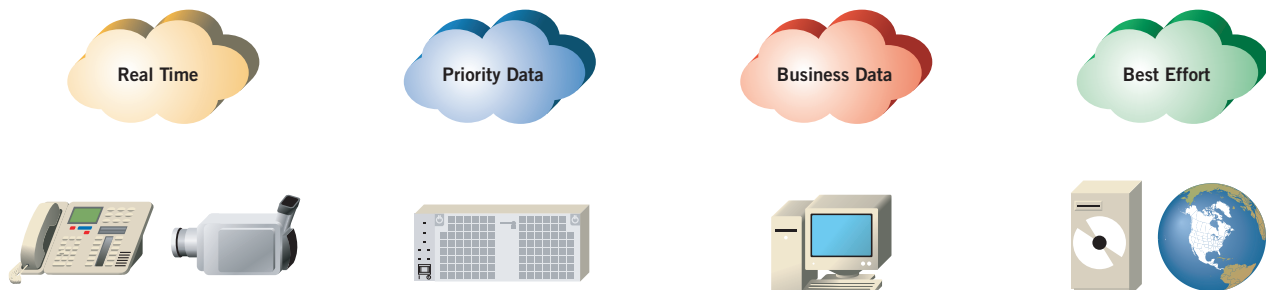


Figure 1. QoS requirements for connectivity services.

From the cost point of view, the latest semiconductor technology has enabled a high degree of functional integration. This has resulted in the reductions in the cost of such feature-rich equipment. Indeed, highly integrated IP/MPLS routers are now able to compete against basic ethernet switches with MPLS capability. This has enabled the deployment of IP/MPLS routers in the access and regional network, bringing the QoS-aware network even closer to the end-user customers.

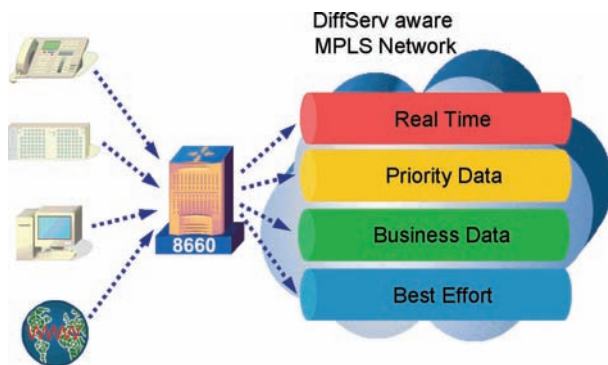


Figure 2. Support for all service classes in a DiffServ-aware MPLS network.

## QoS Parameters

QoS is typically defined by 4 parameters:

1. Packet delay,
2. Delay variation or jitter,
3. Bandwidth,
4. Packet loss.

The following sections provide a brief overview to each of these parameters.

### Packet Delay

Packet delay between 2 network end points is a combination of a number of factors:

- The signal propagation delay is the time it takes for photons to carry information in a physical fiber. This is typically 5 ms per 1000 km. in an optical fiber. In an optical fiber the speed of light is about 65% of the speed in a vacuum.
- The forwarding delay is the time it takes to receive a packet, make a forwarding decision and then begin transmitting the packet. This is typically 10–20  $\mu$ s for modern switch/routers with a hardware-based forwarding plane.
- The serialization delay is the amount of time it takes to transmit the bits of a packet on the outgoing link. The serialization delay depends on the speed of the interface and in general becomes insignificant on higher speed links above 100 Mbps or more. For instance, the serialization delay for a 1500 byte packet on a gigabit ethernet link is about 12  $\mu$ s.

- The queuing delay is caused by statistical multiplexing and is the time a packet has to wait in a queue before it is sent out. The queuing delay can range from zero to hundreds of milliseconds.

The signal propagation delay depends only on the physical distance and it cannot be improved if the distance is fixed. In today's state-of-the-art routers and switches, the forwarding delay is normally in the same range; but overall it is small compared to the other major factors that cause delay. The serialization delay can be improved by using high-speed network interfaces such as 100 Mbps or higher. Higher speed links can also cause the delay factor to become insignificant.

In conclusion, the main factors which can cause delay in packet networks are the signal propagation delay and queuing delay. Since the signal propagation delay is fixed, the queuing delay is the only factor that can be controlled by network planning and by the deployment of traffic management techniques. Therefore it is critical for a service provider to set correctly those parameters which affect the queuing delay.

### Jitter

Delay variation or jitter is a critical parameter for interactive voice and video applications as well as for delivering emulated TDM services. Generally speaking, time insensitive services such as file transfer or streaming services are not affected by jitter. Jitter occurs as the result of one of the following factors:

- Alterations in signal propagation delay due to changes in network topology caused by equipment or link failures.
- Variations in serialization delay caused by different packet sizes.
- Fluctuations in queuing delay caused by changes in the length of the queues.

The variation in the queuing delay is the major source of jitter when compared to the other factors. Therefore it is essential that queue lengths are kept as short as possible.

### Bandwidth

Bandwidth services are defined by four parameters specified as follows:

- CIR is the committed information rate that is always guaranteed to be available.
- PIR is the peak information rate and determines the maximum average rate that the application is allowed to send.
- CBS is the committed burst size and specifies the amount of data that an application is allowed to send at a higher rate than CIR.
- PBS is the peak burst size and is similarly the number of bytes that an application is allowed to send at a higher instantaneous rate than the PIR.

CBS and PBS have a direct effect on queuing delay and jitter. Increasing the CBS and PBS values will multiply the oscillation of the queue length, which in turn will result in increased delay and jitter.

Generally it is not necessary to specify all of these bandwidth parameters for each traffic contract. Table 2 shows how these parameters can be used with real time, priority data, business data and best effort service classes.

Service Class				
	Real Time	Priority Data	Business Data	Best Effort
CIR	X	X	X	
CBS	X	X	X	
PIR			X	X
PBS			X	X

Table 2. Use of bandwidth parameters with different service classes.

### Packet Loss

If we assume that there are no transmission failures at layer 1, that addresses are properly set and that the frame checksum (FCS) does not indicate any errors, then packet loss is normally the result of policing and queue management. Since packet loss due to layer 1 transmission failures or FCS errors is statistically insignificant in modern networks, then the main source of packet loss is caused by policing and queue management.

One of the components of the switch/router responsible for traffic contract enforcement is the policer. The policer may drop a packet if it exceeds the agreed traffic contract. The CIR, PIR, CBS and PBS parameters control how the policer decides which packets are marked and when they should be dropped. The policer's marking and dropping actions will depend on the service class. Also, the queue manager may decide to drop a packet if the length of the queue exceeds a certain limit. Packet loss can therefore be decreased by allocating more bandwidth for the queue in the scheduler, by increasing the maximum queue length allowed and by reducing the RED or WRED drop probability parameters.

### Controlling Delay, Jitter and Packet Loss

Queuing is necessary in order to handle the differences between incoming and outgoing port rates in packet networks. There can be short periods of time when the incoming rate exceeds the service rate of the output port. The resulting level of congestion can be controlled with policers, queue management, scheduling and shaping.

The critical task for the network planner is to allocate adequate resources to reasonably guarantee the QoS of each service class. This reasonable guarantee depends on the nature of the service and can range from best effort with no guarantees, to real-time with very strict delay guarantees. In practice, the best effort service class still needs a level of QoS that will reasonably meet most customers' expectations. Otherwise they will change their service provider causing customer churn. As the best effort service class specifies the peak rate (PIR and PBS) only, overall performance is controlled by adjusting the oversubscription ratio.

It is important to determine the maximum queue length allowed as no single high or low limit will be optimal for all service classes. If the upper limit is set too low, then the queue cannot accommodate traffic bursts without dropping packets excessively. If the upper limit is too high, then memory resources are wasted. The purpose of the queue is to accommodate occasional bursts and this is only possible when the average queue length is short compared to its maximum length. In a well-designed network the average queue should only contain a few packets. A large number of packets in a queue increases delay and jitter, causes packets to be dropped and makes it difficult to use statistical multiplexing.

In order to guarantee the network performance shown in Table 1, we first have to understand the network element performance guarantees. To calculate the level of delay, jitter and packet loss for a network element, a number of assumptions have to be made. These include determining the maximum packet size, maximum CBS, maximum CBS/CIR ratio, maximum load a service class places on an egress port and the egress port bandwidth. Once these element level performance figures are determined, the network topology should be analyzed to calculate the physical distances and number of hops. When all this information is available, the QoS characteristics for each service class can be calculated.

Table 3 and Table 4 give illustrative examples of the network element level performance values for the real time and priority data service classes respectively. These tables show the element level queuing delay for a specified packet loss level i.e. the number of packets exceeding the one way delay limit.

	100 Mbps	STM-1	1 Gbps	STM-16
Max Packet Size (bytes)	1500	1500	1500	1500
Max RT Load on Egress Port	25 %	45 %	70 %	70 %
Packet Loss	10-5	10-5	10-5	10-5
Max One Way Delay (ms)	1	1	0,250	0,250

Note: RT = Real-time service class

Table 3. Element level delay for the real time service class.

	100 Mbps	STM-1	1 Gbps	STM-16
Max. RT + PD Load on Egress Port <sup>(1)</sup>	80 %	75 %	50 %	80 %
Packet Loss	10 <sup>-6</sup>	10 <sup>-6</sup>	10 <sup>-6</sup>	10 <sup>-6</sup>
Max One Way Delay (ms)	3	2	0,3	0,3

Note: RT = Real-time, PD = Priority data service class  
 (1) RT load on egress port <20%

Table 4. Element level delay for the priority data service class.

Table 5 shows the resultant CBS and CBS/CIR settings required to achieve the performance values calculated in Table 3 and Table 4.

	Service Class			
	Real Time	Priority Data	Business Data	Best Effort
CBS (bytes)	≤1500	≤21875	N/A	N/A
CBS/CIR (ms)	N/A	≤17,5	≤30	N/A
Design Target	Delay	Delay	Loss	N/A

Note 1: Maximum RT+ PD load on the priority data service as shown in Table 4.

Note 2: These values are valid when the egress port speed ≤ 100 Mbps.

Table 5. CBS and CIR design rule example.

The queuing delay and queuing memory usage (i.e. packet loss due to congestion) for the priority data service class are controlled by adjusting the upper limits of the CBS and the CBS/CIR ratio. For business data traffic, the queuing memory usage is controlled by adjusting the upper limit of the CBS/CIR ratio so that the drop probability for green packets is less than 0,01%. Since the upper limit of the CBS for the real time service class is one packet, the upper limit of the CBS/CIR ratio has no relevance. The maximum packet size in the real time service class is related to the maximum load on egress port and the one way delay values of Table 3. Decreasing the maximum packet size will increase the maximum allowed load on an egress port if the packet loss and delay are kept fixed.

### Functions of a QoS-aware IP/MPLS Router

The Tellabs 8600 managed edge system is based on a switchless architecture. This means that there is no dedicated switch card; the switching is handled by each of the line cards. The line cards are connected in a full mesh as shown in Figure 3. A switchless architecture means lower initial costs, as the service provider does not have to invest in dedicated switch cards. Instead, the switching capacity increases as the number of interfaces is incremented.

Redundancy is also a built-in feature of a switchless architecture. Since the router does not have any dedicated switch cards, redundancy comes at no extra cost. The switchless architecture supports multicast, MSP 1+1 and LSP 1+1 protection schemes since all packets can be replicated to all working and protecting units over the backplane at the same time. There are a maximum of 16 physical ports on each line card; the actual number depends on the physical interface type.

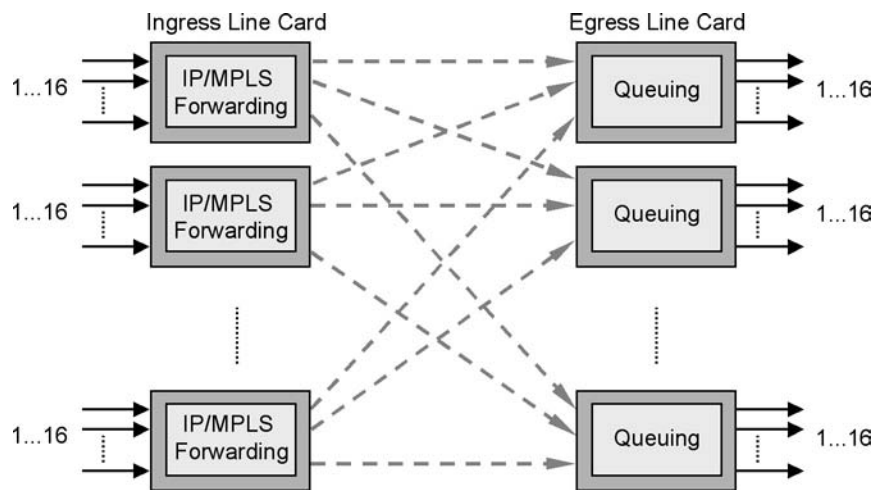


Figure 3. The Tellabs 8600 switchless architecture.

Since both the ingress and egress line cards use the same hardware, the functions of the data path depend on the direction of traffic. Figure 4 illustrates the data path inside the Tellabs 8600 managed edge system. A classifier identifies the packet flow based on L2, L3 and L4 information. A policer marks or drops the packets based on the per hop behaviour (PHB) of the packet.

Queue management decides if the packets are allowed to enter the queue. It also determines the correct queue based on the PHB Scheduling Class (PSC) of the packet and checks the size of the queue. The shaper limits the rate at which packets leave the queue. A shaper may delay the arrival of a packet to the scheduling phase. The scheduler decides which of the queues associated with the egress port is to be served next. There is one scheduler per egress port and it uses 2 scheduling algorithms: strict priority and weighted fair queuing. Note that there is no queuing on the ingress side as the system only requires egress queues.

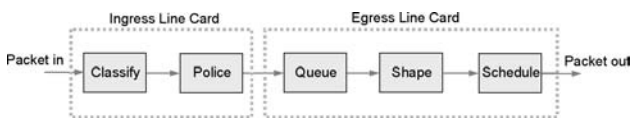


Figure 4. Implementation of QoS functions in the Tellabs 8600.

Policing, queue management, shaping and scheduling can operate on different types of traffic flows. The finest granularity is achieved with a classifier that can identify a flow based on the L3 and L4 information in a packet. A flow can also be identified by L2 information for a VLAN or LSP. In these cases the flow is identified based on the VLAN identifier or the MPLS label. QoS information for the flow is then extracted from the VLAN PRI field, MPLS EXP bits or MPLS label value (L-LSP).

The coarsest granularity is achieved by identifying flows to a physical ingress port. In this case all packets arriving on a particular port will be assigned to the same flow. The Tellabs 8600 managed edge system treats traffic flows in the same way from the QoS point of view. This means that both L2 and L3 traffic can make use of the same QoS functionality.

Figure 5 shows the internal data format in the Tellabs 8600 managed edge system. Although the data arrives in various formats with different sizes, internally it is handled in fixed-size containers. After the L1 and L2 headers have been stripped, the data container holds the MPLS labels, the IP header and the IP data packet. If the traffic comes from a non-MPLS network, the container will not include MPLS labels.

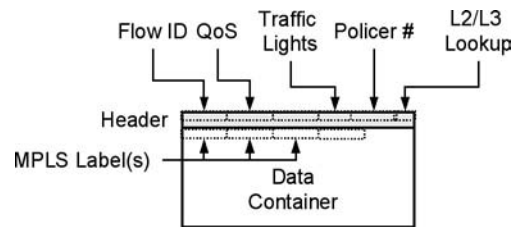


Figure 5. Tellabs 8600 internal data format.

The header field contains the information needed to determine the treatment of the container within the system. The most important fields from the QoS point of view are shown in Figure 5; note that the size of the fields are not drawn to scale. An ingress flow ID identifies each traffic flow. The ingress flow ID can be extracted from a number of sources, as shown in Figure 6. The ingress flow ID can relate to the VLAN, LSP, ATM VP/VC or physical port on which it arrived. The ingress flow ID can then be used for L2 forwarding (MPLS switching or L2 tunneling) and policer selection.

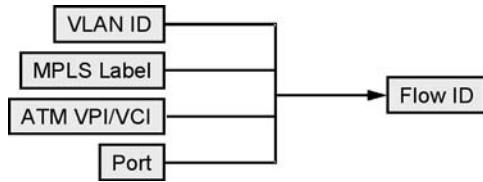


Figure 6. Ingress flow ID.

The QoS and traffic lights fields store the quality of service and drop precedence information for incoming traffic. They also carry the PHB information throughout the router. These fields may contain information from the VLAN PRI, ATM CLP, MPLS EXP or IP DSCP fields. Alternatively, they may be set by a L3/L4 classifier. MPLS L-LSP also sets the QoS field depending on the label value.

On the egress side, the information from the QoS and traffic lights (TL) fields are mapped back to the corresponding fields of the data depending on the protocol used as shown in Figure 7. The QoS field carries the PHB Scheduling Class that is used in the scheduling phase to select the queue. The traffic lights field is used in the policer and WRED queue management module. The policer sets the traffic lights field with the packet color: green, yellow or red depending on the traffic contract. The policer number field contains the address of the policer descriptor memory that contains the marking rules for the traffic flow.

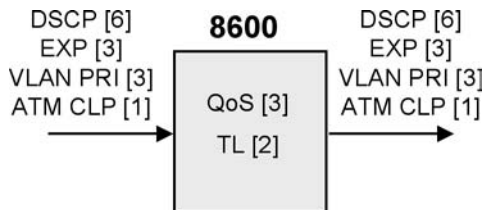


Figure 7. QoS information mapping.

The L2/L3 lookup field indicates if the forwarding is based on MPLS lookup or IP lookup. The Tellabs 8600 managed edge system also supports the tunneling of L2 traffic or port-to-port tunneling. Regardless of the layer at which the forwarding decision is made, the QoS functions of the Tellabs 8600 managed edge system are identical.

**Traffic Classification**

The Tellabs 8600 managed edge system classifies the packets at the edge of the differentiated services (DS) domain. The DS domain can be an IP network using DS or a DS-aware MPLS network. The concept of the DS architecture is that traffic is classified, policed, marked and possibly shaped at the edge of the DS domain as shown in Figure 8. The DS interior nodes (i.e. the IP/MPLS routers) select the queuing and scheduling treatment based on the PHB of the packet. This classification based on the PHB of the packet is called behavior aggregate (BA) classification.

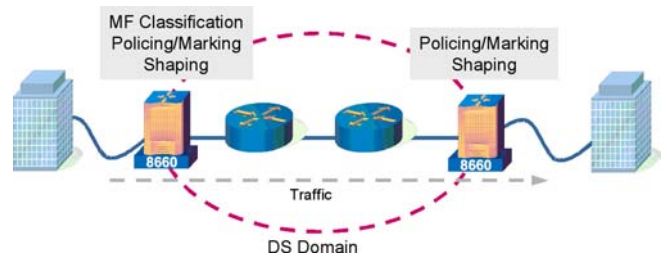


Figure 8. Differentiated services model.

The identification of customers at the edge of the DS domain is based on multi-field (MF) classification in which the L3 and L4 fields of an incoming packet are inspected. The fields that can be used in a classifier are shown in Table 6. Mathematical operands (e.g. equal, not equal, greater than, less than, within range, out of range) can be given for L4 source and destination ports to specify a range of port values that can generate either an accept or a deny decision. The matching classifier contains the PHB for the packet. It may also contain a pointer to the policer descriptor memory which contains the CIR, CBS, PIR and PBS parameters for the packet. The other alternatives for selecting a policer are the physical port, logical port (VLAN) or IP next hop lookup.



Filtering rule parameters
IP Source Address
IP Destination Address
IP ToS/DSCP
IP Fragmented Packet
L4 Protocol
L4 Source Port
L4 Destination Port
L4 Code Bits

Table 6. Parameters for a filtering rule.

The PHB for a packet is selected as a result of the classification. An example of PHB allocation is shown in Table 7. Currently the IETF has defined six PHB groups: EF, AF1, AF2, AF3, AF4 and BE. While real time traffic is marked with EF and best effort with BE, priority and business data can be assigned to the one of the AF groups. The remaining AF groups can be assigned to other types of traffic as required.

	PHB
Real-time	EF
Priority Data	AF41
Business Data	AF31      AF32      AF33
Best-effort	BE

Table 7. PHB assignment for service classes.

A classifier rule may consist of multiple filter entries. Each filter entry consists of a set of fields, as shown in Table 6, and an operand for L4 source and destination ports. Figure 9 shows the steps before and after the classifier. The classifier takes as input the L3/L4 fields of the packet and the ingress flow ID or the virtual route forwarding (VRF) number. If a packet comes from a non-MPLS network, the flow ID is used. If a packet comes from an MPLS network with an inner label indicating a VRF, the VRF number is used. The flow ID can represent a VLAN, ATM VPI/VCI or MPLS label. If the classification takes place at the edge of the MPLS network, the flow ID typically indicates a VLAN ID. If the flow ID is used as input to the VRF lookup, the flow ID represents an MPLS label. The matching classifier may return a PHB (QoS and traffic lights bits) and a policer number, together with an accept or a deny decision for the packet.

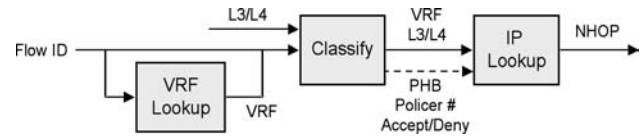


Figure 9. Classifier function.

**Policer and Marker**

In order to police traffic that is characterized by the CIR, PIR, CBS and PBS parameters, a two-rate, three-color marker (tr-TCM, RFC 2698) is required. The policer checks the incoming packet against the CIR and PIR token buckets associated for this packet flow, and marks the packet either with green, yellow or red. Instead of marking, the policer can also drop non-conforming packets. The decision whether to drop or mark packets can be set for each service class.

A suggestion on how to define these is shown in Table 8. If the traffic is in the real time or priority data service class, all the packets arriving above CIR are dropped. For the business data service class, packets are marked with green, yellow or red;. Alternatively, they may be dropped depending on the configuration. For the best effort service class, all packets arriving above PIR are dropped. This color coding is carried internally with the packet in the traffic lights bits and it will be written to the drop precedence bits of the DSCP or to the EXP field of the L-LSP or E-LSP on the egress side of the router before the packet is sent out.

	Service Class			
	Real Time	Priority Data	Business Data	Best Effort
$f \leq \text{CIR}$	–	–	Green	–
$\text{CIR} < f \leq \text{PIR}$	Drop	Drop	Yellow	–
$f > \text{PIR}$	Drop	Drop	Red/Drop	Drop

f: Rate of the traffic flow

Table 8. Policar function for service classes.

The working principles of a token bucket policer are described in Figure 10. The CIR is used to define the rate at which tokens are put into the CIR token bucket.  $\Delta t_{CIR}$  is the time between token arrivals, therefore  $CIR \sim 1/\Delta t_{CIR}$ . Similarly, the PIR is used to define the rate at which tokens are put into the PIR token bucket.  $\Delta t_{PIR}$  is the time between token arrivals, so that  $PIR \sim 1/\Delta t_{PIR}$ . CBS defines the maximum size of the CIR token bucket and PBS defines the maximum size of the PIR token bucket.

Initially both the token buckets are full. Each time a packet arrives, it is tested against the CIR and PIR buckets and the number of tokens corresponding to the size of the packet are removed from each bucket. A packet conforms to the CIR rule if there are enough tokens in both the CIR and PIR buckets; a packet conforms to the PIR rule if there are enough tokens in the PIR bucket. If there are enough tokens in both the CIR and PIR buckets, tokens are removed from both of the buckets. If there are only enough tokens in the PIR bucket, then only the PIR bucket is reduced. The policer marks or drops packets depending on which service class they belong to as shown in Table 8. The service class of a packet is identified by the PHB group that was set by the classifier at the ingress edge of the DS domain.

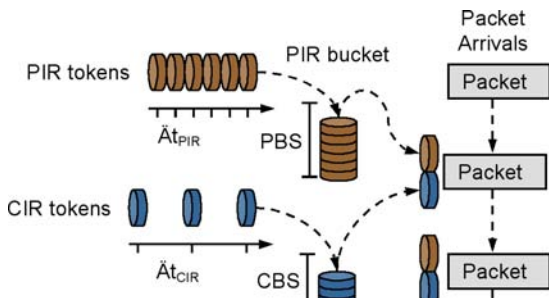


Figure 10. Token bucket policer.

Figure 11 illustrates an example of the packet marking process for the business data service class. Packets are marked with green as long as there are tokens in both buckets. If the CIR bucket is empty but there are still tokens in the PIR bucket, packets are marked with yellow. If both the buckets are empty, packets are marked with red or they are dropped.

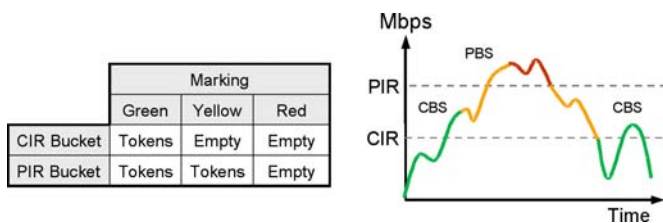


Figure 11. Packet marking vs. token availability for business data.

The CIR, PIR, CBS and PBS parameters are stored in the policer descriptor memory. The alternatives for the policer selection are:

- Classifier: the matching classification rule contains a policer descriptor address.
- IP next hop (NHOP) lookup: the IP DA or VRF + IP DA is used as input to the lookup table that returns the policer descriptor address.
- L2 flow (MPLS, Ethernet, ATM) + QoS: The flow ID and QoS bits may directly address policer memory. QoS (E-LSP, L-LSP, ATM CLP, VLAN PRI) bits are used together with the L2 flow ID to divide the flow into a maximum of eight priority levels. Therefore each priority level can have its own policer.
- Ingress port + QoS: Each ingress port can be associated with a policer. QoS bits give eight priority levels that can be used together with the physical port number.

Figure 12 shows the policer selection hierarchy. The first selection is whether to use a physical port number or a flow ID as a policer address. The selection between a physical port and a flow ID is made based on the flow ID value; a range of flow IDs can directly address the policer descriptor memory. The QoS bits will be used together with the physical port or the flow ID, so that the policer can be selected based on the QoS requirements of the packet. The policer can also be selected by the classifier or IP next hop lookup, which will override the flow ID or the physical port based policer selection.

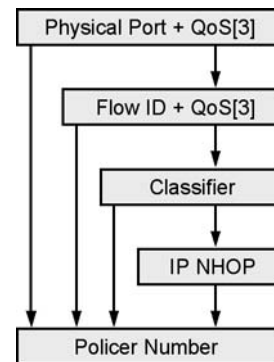


Figure 12. Policer selection hierarchy.

### Queue Management

Queue management decides which packets are allowed to enter the queue based on the level of queue occupancy. Active queue management means that a router may drop packets before a queue becomes full. RED and WRED are commonly regarded as effective queue management algorithms for traffic that uses a transport protocol with end-to-end flow control such as TCP. TCP operates by increasing and decreasing the transmission rate until the maximum rate without packet drops (or set ECN bits) is found.

As TCP increases the transmission rate, the queues in the routers along the path become longer. This in turn increases the packet drop probability of the RED/WRED algorithms. TCP will notice these packet drops and slow down the transport rate. In this way excessive queue lengths and, in the worst case, total network congestion, can be avoided. Instead of dropping the packet, WRED may also set the ECN bit in the IP header to indicate the congestion. The benefit of using ECN is that instead of dropping the packet, the sending application gets an indication that there is a congestion in the network. It can then use this information to slow down the transmission rate. Since packet drops or ECN marking tend to occur randomly in TCP flows, the oscillation of a large volume of TCP flows is avoided. This can prevent the problem of global TCP synchronization. Also, the queues remain short as RED and WRED start dropping or marking before queue lengths build up.

Real time services usually use UDP as a transport protocol, while most of the business data and best effort services typically use TCP. The challenge for network planners is to optimize the network utilization while minimizing the transit time for the delay-sensitive traffic. This is achieved by separating delay-sensitive real time traffic to a queue that uses strict priority queuing. TCP-based flows can then be allocated to the other queues depending on the service class. Bandwidth allocation for each queue is then set based on the desired QoS performance and the amount of guaranteed traffic allocated for each queue. The over-subscription ratio for a service class can also affect the QoS performance and the overall utilization of the network.

Figure 13 shows the architecture of the queue management and the scheduling phase. There is an individual queue management function for each queue. The selection of a queue is based on the scheduling class indicated by the PHB Scheduling Class (PSC). The PSC (which is part of the PHB information) is carried in the DSCP field of an IP packet, in the EXP field of an E-LSP or it is implicitly derived from the label value of an L-LSP. A classifier is responsible for setting the PHB for packets coming from a customer's network.

There are eight queues per egress port, so that the total number of queues per line card depends on the number of physical interfaces. In addition to these queues, each line card contains 1000 additional queues that can be used for per VLAN or per LSP queuing. These queues have similar queue management, scheduling and shaping features as the queues described in Figure 13.

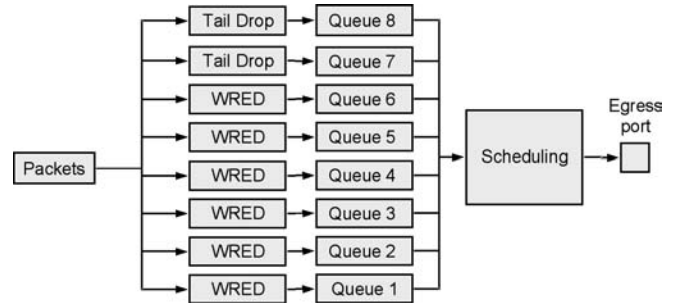


Figure 13. Queue management for each queue.

Table 9 shows the queue allocation proposed for each service class. The priority data service class uses AF41 PHB only, so in practice the queue management is RED. The control traffic requires one queue (number 7) and the service classes require four queues. Thus there are three queues available for further use. Queue assignments are adjustable in case there is a need to change these settings.

Service Class	PSC	Queue #	Queue Management	Scheduling
Real-time Priority	EF	8	Tail Drop	Strict
Priority Data	AF4x	4	WRED	WFQ
Business Data	AF3x	3	WRED	WFQ
Best-effort	BE	5	RED	WFQ

Table 9. Queue allocation to service classes.

The RED drop probability function is specified by three parameters: minth, maxth and maxp. Minth defines the length of the queue at which point the drop probability function starts to increase. If the length of the queue is below minth, all packets are allowed to enter the queue. Maxth defines the maximum length of the queue. All the packets entering a queue are dropped if the average queue length is at maxth. Maxp is the drop probability at maxth. With these three parameters we can define a drop probability function shown in Figure 14. The queue length is an average value which is calculated by the formula:

$$Avg_n = Avg_{n-1} + w * (Q - Avg_{n-1})$$

where Avg is the average queue length, Q is the actual queue length and w is the weight that defines how fast the average value responds to changing Q values. A typical value for w is in the range of 0,001...0,002 (~2<sup>-10</sup>...2<sup>-9</sup>).

When a packet enters a queue, a random number is generated and it is tested against the drop probability function. Packets are dropped randomly, but the long term drop probability average will follow the function. However, minth and maxth are strict limits in the sense that a packet is never dropped if the queue length is below the minth value and a packet is always dropped if the queue length is at the maxth value.

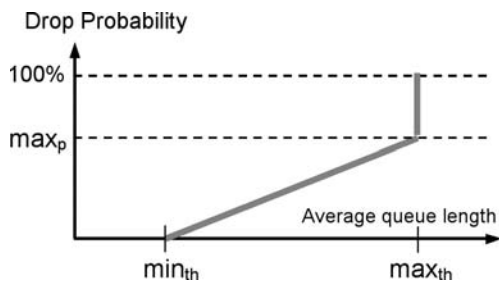


Figure 14. RED drop probability curve settings.

WRED works in the same way as RED, but minth, maxth and maxp are individually defined for each packet color: green, yellow and red. This leads to three different drop probability functions that are shown in Figure 15. This example shows a queue length at a particular instant. If a packet entering the queue is green, it is allowed to enter. If the packet is yellow, there is a very small probability that it is dropped. If the packet is red, there is a significant likelihood that the packet is dropped.

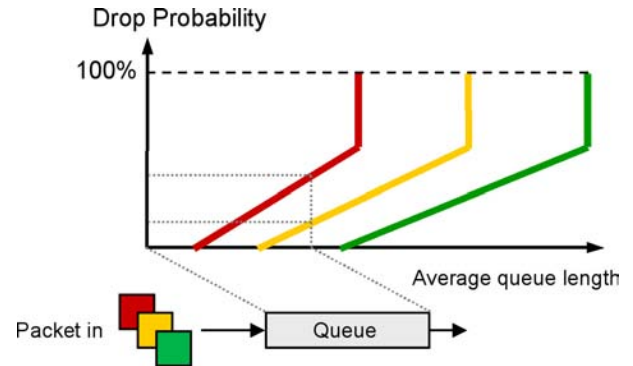


Figure 15. WRED drop probability functions for green, yellow and red packets.

Tail drop is another alternative for queue management. With tail drop only the maximum length of the queue is defined. Packets are allowed to enter the queue if the queue length is below the maximum. If the queue reaches the maximum length, all subsequent packets are dropped. Tail drop is well-suited for real time applications that use UDP as a transport protocol since UDP does not have end-to-end flow control like TCP. As the real time traffic uses strict priority queuing, this will always take precedence. Packet drops can therefore be avoided by setting the CIR, CBS and queue length appropriately depending on the speed of the physical interface. The network delay experienced by the real-time traffic will depend on the CBS and the actual traffic load on the links.

The recommended settings for the maximum queue length allowed, min<sub>th</sub>, max<sub>th</sub> and max<sub>p</sub> are shown in Table 10. For the priority data service class, the maximum queue length allowed is based on the following assumptions:

- A packet entering a queue that is so heavily congested that the packet would be delayed for more than 6 ms is dropped. This is because the priority data service class specifies an upper boundary for delay.
- The average size of a packet is 500 bytes.
- The speed of an egress link is 100 Mbps.

The activity of the real time service class queue also affects the performance of the priority data service class. This is because the real time traffic uses strict priority queuing. If the real time service class queue is constantly occupied, the priority data service class, as well as other service classes, are delayed until real time traffic is served. This means that by setting a maximum queue length allowed for the priority data we cannot guarantee that any priority data packet will not be delayed for more than 6 ms. The maximum allowed queue length values for business data and best effort data are based on the WRED and RED settings. The maximum allowed queue length (200 packets) is chosen so that there is a very small probability for the case in which the queue is controlled by the maximum allowed queue length value instead of RED/WRED.

The selection of RED and WRED parameters is based on the following principles:

- Keep the average queue length low to avoid an increase in round trip delay due to long queues.
- Avoid global synchronization of TCP sources.
- Differentiate the treatment of packets with different drop precedence in WRED.

The setting of the maximum queue length allowed for the priority data service class depends on the preferences of the service provider and end users. There are two possible approaches:

- 1) drop all the packets entering a congested queue if they will exceed the delay limit as proposed in Table 10, or
- 2) avoid packet drops even though they exceed the delay limit.

In the latter case the queue length can be the same for all service classes (e.g. 200 packets).

	Priority Data	Business Data	Best Effort
Queue length (Packets)	150	200	200
Minth Green (Packets)	30	30	N/A
Maxth Green (Packets)	90	90	N/A
Maxp Green	10%	10%	N/A
Minth Yellow (Packets)	N/A	5	N/A
Maxth Yellow (Packets)	N/A	15	N/A
Maxp Yellow	N/A	10%	N/A
Minth Red (Packets)	N/A	5	5
Maxth Red (Packets)	N/A	10	15
Maxp Red	N/A	20%	10%

Table 10. RED and WRED settings for a 100 Mbps egress link.

For the real time traffic where tail drop queue management is used, only the maximum queue length has to be set. The length of the queue depends on the amount of queuing delay each router is allowed to add. To meet the QoS requirements of Table 1, the settings shown in Table 11 are recommended. The table shows the queue length in milliseconds and the corresponding queue size in bytes. The packet drop probability with these settings is in the order of 10-5.

	100 Mbps	STM-1	1000 Mbps	STM-16
Queue length (ms)	1	1	0,25	0,25
Queue length (bytes)	12500	18750	31250	77500

Table 11. Queue length for real-time service class.

## Queue Scheduling

The purpose of the queue scheduling algorithm is to decide which is the next packet to be served. Networks that implement differentiated services can assign service classes that support both guaranteed (or committed) and excess traffic components. An important feature of a queue scheduling algorithm is to guarantee that the committed rate service can be delivered. This means that each queue should receive at least its fair share of the total output port bandwidth as indicated by its weight in every time interval. This must also be independent of the previous usage of the bandwidth.

The algorithm should also protect against misbehaving traffic flows in other service classes. This means that the available share of the bandwidth is not reduced by traffic flows in another service class using more than their allocation. The algorithm should also allow the other service classes to use the excess bandwidth when all the bandwidth allocated to a service class is not used. And finally, the algorithm should be such that it can be implemented efficiently in hardware. This is very important in order to guarantee wire-speed operation in high-speed networks.

There are a number of different queue scheduling algorithms available and each of them attempts to find the optimal solution by balancing fairness and computational complexity. There is no industry standard algorithm for queue scheduling. Fortunately it has been proven that a class of fair throughput scheduling algorithms guarantee the overall fairness and provide an upper bound on end-to-end delay in the network, even in a multi-vendor environment when using different implementations. An end-to-end fairness guarantee is especially important when offering service level agreements to customers.

The Tellabs 8600 managed edge system uses start-time fair queuing (SFQ) which is a variation of the fair throughput algorithm WFQ. In SFQ two tags, a start tag and a finish tag, are associated with each queue:

$$S(p_j) = \max\{v(A(p_j)), F(p_{j-1})\}, j \in M$$

$$F(p_j) = S(p_j) + (L_j/r), j \in M$$

Where:

$S(p^j)$  is the start tag of the packet  $p^j$   
 $A(p^j)$  is the arrival time of the packet  $p^j$   
 $F(p^j)$  is the finish tag of the packet  $p^j$   
 $v(A(p^j))$  is the virtual time (defined as the start tag of the packet in service at the arrival time of packet  $p^j$ )  
 $L^j$  is the length of packet  $p^j$  and  $r$  is the scheduling weight of the queue serving the packet flow  $p$ .

Packets are scheduled in increasing order of the start tags of the packets. The calculation of virtual time  $v(t)$  is very simple in SFQ. Whereas the other variations of WFQ require a more complex calculation.



Below is an example of a SFQ calculation. Table 12 and Table 13 represent two packet flows in which: L is the length of a packet, A is the arrival time of a packet, S is the start tag, F is the finish tag and V(t) is the virtual time. The order column shows the order in which the packets are sent out. The first packets in both queues arrive at the same time, so their start tags are the same. In this case the scheduler decides arbitrarily which packet is sent out first. At each turn, the start tag and the finish tag are calculated for the next packet to be sent out in each queue; the packet with the smallest start tag value is selected.

	L	A	S	F	V(t)	Order
P11	2	0	0	2	0	1
P21	2	1	2	4	2	4
P31	2	10	5	7	5	9
P41	2	11	7	9	7	11
P51	2	12	9	11	9	14
P61	2	13	11	13	11	15

Table 12. Packet flow 1,  $r_1 = 1$ .

	L	A	S	F	V(t)	Order
P12	2	0	0	1	0	2
P22	2	1	1	2	1	3
P32	2	2	2	3	2	5
P42	2	3	3	4	3	6
P52	2	4	4	5	4	7
P62	2	5	5	6	5	8
P72	2	11	6	7	6	10
P82	2	12	7	8	7	12
P92	2	13	8	9	8	13

Table 13. Packet flow 2,  $r_2 = 2$ .

As can be seen from Figure 16, even though the first packets arrive at the same time in both the queues, queue 2 gets twice the bandwidth because of the scheduling weights  $r_2 = 2$  and  $r_1 = 1$ . The weights can be arbitrarily selected as it is their ratio which controls the bandwidth allocation. Note that on turns 4 and 5 as well as on turns 11 and 12, the start tags are again equal. In these cases the scheduler decides randomly which packet is sent out first. The function of virtual time is seen on turn 9: in this case the start tag for the packet 3 in flow 1 is 5 (not the previous finish tag = 4). The selection of the larger value (virtual time in this case) prevents flow 1 from stealing all the bandwidth from flow 2, because flow 1 did not send any packets for some period of time. On turns 14 and 15, queue 1 is the only one with packets, so it now has all the bandwidth available. If a service provider wants to limit the bandwidth in a certain queue, shaping can be used. Shaping delays the arrival of packets to the start tag and finish tag calculation phase.

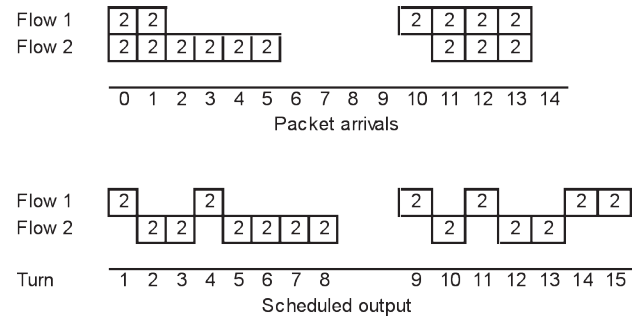
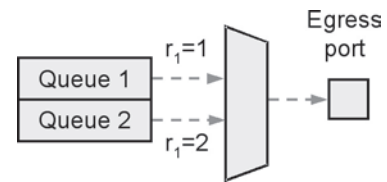


Figure 16. Packet arrivals and scheduled output.

SFQ meets all the requirements of a good scheduling algorithm:

- it enables the allocation of egress port bandwidth to a number of queues;
- each queue gets a guaranteed minimum bandwidth depending on the scheduling weight if they have packets to be served;
- queues can utilize excess bandwidth if other queues are empty;
- ill-behaving traffic flows do not steal bandwidth from other queues;
- prior use of bandwidth does not affect the future allocation of bandwidth;
- it has an efficient implementation in hardware.

A combination of strict priority queuing (SP) and WFQ is required to implement the service classes of Table 1. SP queues have the highest priority: if there are packets in the SP queues, they are sent out immediately after the packet currently in service has been sent out (non-pre-emptive scheduling). SP queues are used for the real-time traffic as it requires the smallest possible delay. The maximum delay per router is controlled by setting the SP queue length as shown in Table 11. Packets marked with EF PSC will be subject to SP scheduling as shown in Figure 17.

The second alternative is to use WFQ that allocates the available bandwidth based on relative weights. A share of the available bandwidth can be given to each service classes by choosing the weights accordingly. The minimum bandwidth allocated for each queue is dictated by the amount of traffic that is guaranteed to be served by the queue (guaranteed component). The excess traffic, for which bandwidth is not reserved, is subject to statistical

multiplexing. The amount of excess traffic defines the level of congestion that the router experiences. Packets marked with AF1x, AF2x, AF3x, AF4x or BE PSC will be assigned to queues that use WFQ scheduling, as shown in Figure 17.

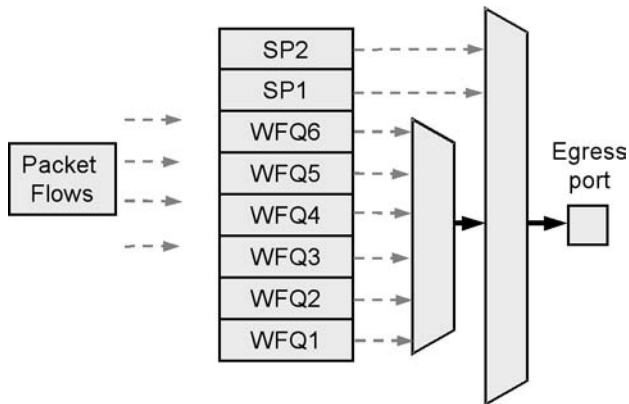


Figure 17. Scheduling.

SP and WFQ can be employed at the same time. If SP queues have traffic to send, it is always sent out first. If SP queues are empty, WFQ uses all the bandwidth. There is also a leakage parameter for SP queues that releases bandwidth for WFQ in case the SP queues are constantly occupied. The leakage parameter should be set so that the delay constraints of the real-time service are not exceeded.

Table 14 shows the weight settings for each queue. SP queues do not have any weights as they are always served first. The real time services should be policed to ensure that they do not consume all the bandwidth. The recommended policer setting (see Table 5) suggests using CBS = 1 packet for the real time service. The bandwidth usage of the SP queue is then effectively the sum of the CIRs of the real time services using that particular port. The use of a leakage parameter is not recommended as it simply adds delay to the real time service. The weight of the priority data service should be set high to ensure that the priority data queue is served next after the SP queues. The weight for the business data services should be in proportion to the expected maximum sum of guaranteed (CIR) traffic in a port. The remaining weights are set based on the required performance.

Service Class	Queue	Rate <sup>(1)</sup>	Scheduling weight $r_n$ setting
Real-time	SP2	B (2)	
Control	SP1	B (3)	
	WFQ6	$(r_6/R)B$	
Best-effort	WFQ5	$(r_5/R)*B$	
Priority Data low delay.	WFQ4	$(r_4/R)*B$	$r_4=8000$ <sup>(4)</sup> to ensure
Business Data	WFQ3	$(r_3/R)*B$	Reservation of guaranteed bandwidth.
	WFQ2	$(r_2/R)*B$	
	WFQ1	$(r_1/R)*B$	

$B$  = port bandwidth,  $R = r_1 + r_2 + r_3 + r_4 + r_5 + r_6$

(1) Minimum rate if the queue is constantly backlogged.

(2) Leakage parameter gives transmit opportunities to WFQ.

(3) The amount of control traffic is low, so it will not consume all the bandwidth.

(4) The value range of weight  $r = 1 \dots 8000$ .

Table 14. Weight settings for WFQ.

## Shaping

The function of the shaper is to limit the output rate by delaying the moment when a packet is allowed to leave the queue for the egress port. If there is no shaping, a packet will be sent out as soon as the scheduler gives it permission. Shaping is set on a per queue basis, which allows the setting of different bandwidth limits for each service class. Shaping is useful for smoothing out bursty traffic profiles. This improves the overall network performance, as congestion caused by traffic bursts is less likely to happen. Shaping is implemented with token buckets in a similar way to policers, by defining CIR, PIR and CBS.

Rate limiting is an alternative method for limiting the speed of traffic. Figure 18 illustrates the difference between rate limiting and shaping. In this example the egress port is rate-limited or shaped to 4 Mbps. Let's assume that a queue has a burst size of five packets and that each packet is 1000 bytes. Figure 18 illustrates how the packets are sent out in the case of a rate-limited port and a shaped port. Rate limiting allows packets to go out immediately if bandwidth is available, as long as the upper limit (4 Mbps) is not exceeded during the measurement interval, which is 10 ms in this case. The shaper, on the other hand, calculates departure times for the packets so that the bandwidth limit is not exceeded. This effectively means delaying the packet departure times. As a result, shaped traffic is smoother than rate-limited traffic.

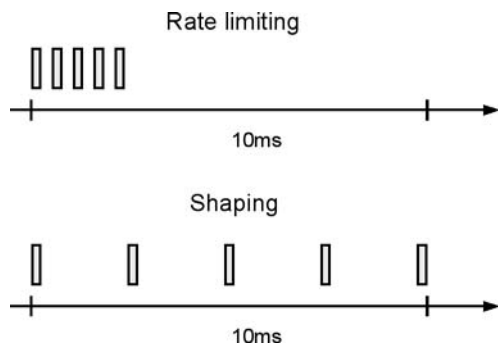


Figure 18. Rate limiting versus shaping.

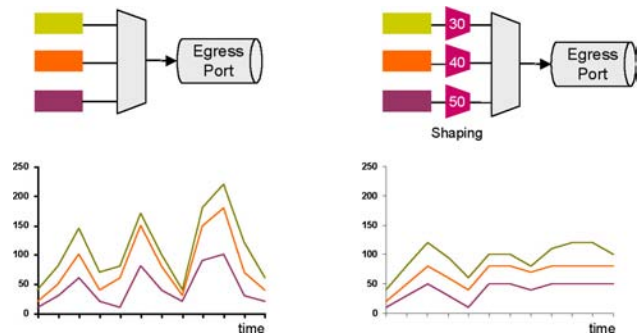


Figure 20. Shaping of individual queues.

Figure 19 and Figure 20 illustrate how shaping can be used in the Tellabs 8600 managed edge system. In Figure 19 there are three traffic flows that use a common queue. The diagrams illustrate the bandwidth (bits per second) on the egress port for these traffic flows. The diagram on the left shows the bandwidth without shaping, while the diagram on the right shows the bandwidth shaped to 120 units. Figure 20 shows the case when the three traffic flows are in separate queues. The diagram on the right in Figure 20 shows the situation when the flows are shaped to 50, 40 and 30 units.

Shaping is an optional function; it can be turned on or off for each queue depending on the service requirements. Each line card has 8 queues per egress port plus 1000 additional queues that can be associated to any egress port on a line card. These queues can be used for per-LSP or per-VLAN queuing. Figure 21 illustrates the case where the additional queues are used on a per-VLAN basis. In this case they are shaped to limit the amount of bandwidth from the customer VLAN to the network. Shaping can also be used to limit the amount of bandwidth from the network to the customer.

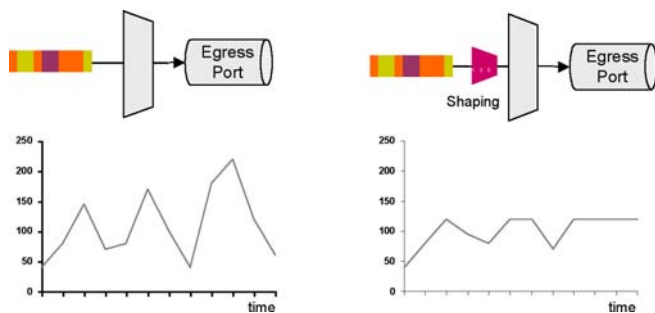


Figure 19. Shaping of a single queue.

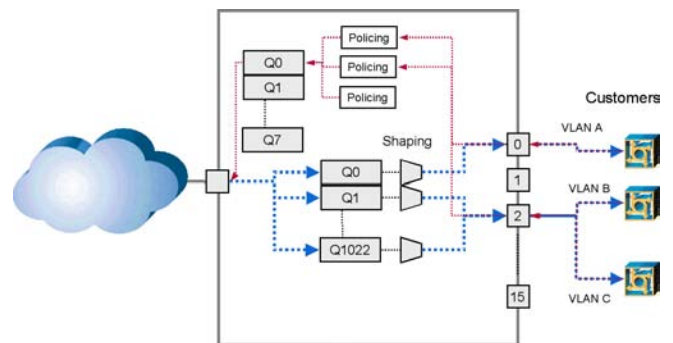


Figure 21. Per VLAN policing and shaping.

## Summary

A wide range of services with different QoS requirements can be delivered using an IP/MPLS network. An IP/MPLS router identifies and classifies packets according to customers and applications. It also marks or drops packets to limit the bandwidth according to the customer's traffic contract. In addition, a router uses the RED/WRED algorithm for queue length control, and strict priority or WFQ scheduling algorithms for bandwidth allocation between the queues. Shaping can be used to limit the bandwidth and to smooth out bursty traffic. The queuing delay and jitter depend on the packet size distribution, burstiness, maximum load of a service class on an egress port, and egress port bandwidth.

## Acronyms

Term	Explanation
BA	Behavior Aggregate
CBS	Committed Burst Size
CIR	Committed Information Rate
DA	Destination Address
DS	Differentiated Services
DSCP	Differentiated Services Code Point
E-LSP	EXP-inferred-PSC LSP
FCS	Frame Checksum
FR	Frame Relay
IP	Internet Protocol
L1/L2/L3/L4	Layer 1/2/3/4 of the OSI model
LAN	Local Area Network
LSP	Label Switched Path
L-LSP	Label-only-inferred-PSC LSP
MPLS	Multi Protocol Label Switching
PBS	Peak Burst Size
PHB	Per Hop Behaviour
PIR	Peak Information Rate
PSC	PHB Scheduling Class
QoS	QoS Quality of Service
RED	Random Early Detection
RFC	Request For Comments
SFQ	Start-time Fair Queuing
SLA	Service Level Agreement
SP	Strict Priority
VLAN	Virtual LAN
VRF	VPN Routing and Forwarding (table)
WFQ	Weighted Fair Queuing
WRED	Weighted Random Early Detection

### North America

Tellabs  
One Tellabs Center  
1415 West Diehl Road  
Naperville, IL 60563  
U.S.A.  
+1 630 798 8800  
Fax: +1 630 798 2000

### Asia Pacific

Tellabs  
3 Anson Road  
#14-01 Springleaf Tower  
Singapore 079909  
Republic of Singapore  
+65 6215 6411  
Fax: +65 6215 6422

### Europe, Middle East & Africa

Tellabs  
Abbey Place  
24-28 Easton Street  
High Wycombe, Bucks  
HP11 1NT  
United Kingdom  
+44 871 574 7000  
Fax: +44 871 574 7151

### Latin America & Caribbean

Tellabs  
Rua James Joule No. 92  
EDIFÍCIO PLAZA I  
São Paulo – SP  
04576-080  
Brasil  
+55 11 3572 6200  
Fax: +55 11 3572 6225