# Applications of QoS

## 6.1 Operating System Support for Quality of Service

When considering application support for QoS, a key underlying issue will be the level of QoS support in the operating system for packet marking. Earlier sections of this guide have stated that it is preferential to move the packet marking process as close to the source as possible. As such, the ability to do this within the host itself would be the ultimate outcome here. This section will present an overview of QoS support in the major end user operating systems including Microsoft Windows, Linux and UNIX based systems.

### 6.1.1 QoS Support in Microsoft Windows

As described in the Quality Windows Audio-Video Experience - qWave [qWave] - QoS mechanisms in Windows operating systems have undergone significant changes in the last few releases to adapt to changes in the prevalent QoS approach and support bandwidthintensive applications.

Windows 2000 first introduced the Generic QoS (GQoS) application programming interface (API) as a framework for QoS. The GQoS API provided access to QoS mechanisms that were available as part of the networking stack based on the Resource Reservation Protocol (RSVP) for signalling and reserving resources on the network; traffic shaping and filtering mechanisms; and layer 2 and layer 3 priority-marking mechanisms. Windows 2000 also provided tools, such as the Subnet Bandwidth Manager and QoS policy control. The WinSock2 API provided application programmers with access to windows QoS functionality from Windows 2000 onwards. Figure 6-1 below shows the (RSVPbased) QoS interface in WinSock2.

```
typedef struct _flowspec
{
uint32       TokenRate;           // In Bytes per sec
uint32       TokenBucketSize;     // In Bytes
uint32       PeakBandwidth;       // In Bytes per sec
uint32       Latency;             // In Microseconds
uint32       DelayVariation;      // In Microseconds
SERVICETYPE  ServiceType;         // Guaranteed, Predictive
uint32       MaxSduSize;          // In Bytes
uint32       MinimumPolicedSize;  // In Bytes
} FLOWSPEC;
```

*[1]*

In Windows XP, the focus shifted towards prioritization and traffic shaping mechanisms.

Although GQoS continued to be the application interface for accessing QoS prioritisation, the reservation mechanisms had been removed as DiffServ packet marking replaced RSVP signalling. The kernel component that implements prioritisation and traffic shaping is the QoS Packet Scheduler (Psched.sys), which is accessible through the GQoS API and through a lower-level application interface called the Traffic Control (TC) API. The TC API provides control of QoS mechanisms (such as prioritization and shaping) at the host level, rather than at the application level, but it requires administrative privileges to be invoked. The QoS mechanisms provided in Windows XP support enterprise QoS needs for wired networks. In Windows XP Service Pack 2 (SP2), the GQoS mechanisms allow the application to set layer 3 priorities only.

In the Windows Vista release, these QoS mechanisms will continue to be supported for enterprise QoS needs, but now that home networks and wireless technologies have become common, consumers may wish to support QoS in the home. In the Windows Vista release, qWave provides this support. In addition, some existing QoS mechanisms, such as Psched.sys, have also been extended to address home network scenarios.

### 6.1.2 QoS Support in Linux

Support for Quality of Service was available for Linux kernel versions from 2.1.90 onwards and support for DiffServ was formally integrated into Linux from kernel version 2.4 [LinuxDS]. Linux is an interesting case when considered next to MS Windows, as it is typically used both as an operating system in end-hosts and network devices. As such, in addition to packet marking support, Linux incorporates QoS handling functionality necessary for routing devices.

The QoS handling support in Linux consists of the following three basic building blocks:

- queuing discipline
- class based queuing
- filters/policers/classifiers

Linux QoS traffic control is supported through the availability of a number of components that map onto traditional TC functionality:

- **Scheduling** - supported by *qdisc*. Default is FIFO, but other algorithms that control the rearranging of packets is supported
- **Shaping** - provided by the *class* component
- **Classifying** - performed by the *classifier* and *filter* components. Users can select packets based on attributes of the packet
- **Policing** - a *policer* requires one action above and another action below a specified rate
- **Dropping** - as part of the *policer*, any rule can have a drop action. The *policer* can be configured to drop all traffic matching a particular pattern
- **Marking** - an element of *qdisc*. Packets can be marked to represent priority

Extensive documentation is available online that provides guidance for configuring and using DiffServ DSCP traffic marking on Linux end user operating systems. Two popular tools for doing this are iptables and iproute2.

### 6.1.3 QoS Support in other Operating Systems

To date there is little in the way of formal QoS support in UNIX-based end systems, including Apple OS X, beyond that developed for Linux.

### 6.1.4 Operational/Practical Issues for QoS on End Hosts

Despite the availability or otherwise of QoS support in the end user operating systems, users/applications wishing to employ traffic marking without prior consent from the network administrator are unlikely to actually experience an improvement in service. This is because, depending on the provider's QoS policy, traffic with unauthorised prioritisation/marking or coming from unauthorised destinations will typically find that this is ignored or even stripped out of the packet when it enters the provider network. In extreme cases, this traffic might even be dropped outright. As we have seen in section 3, in the case of JANET it is possible that packets marked in this way will simply traverse the network transparently and not be affected. As such, in addition to ensuring that your end system and applications support QoS marking, it will still be necessary to contact your network provider to gain authorisation for any QoS provisioning before traffic is sent. This aspect is covered from the JANET perspective in Section 7 of this guide.

## 6.2 Voice over IP (VoIP)

This section discusses VoIP in the scope of formal campus-supported deployments rather than specific applications such as Skype et al. VoIP is a generic term for the routing of voice conversations over the public Internet or any other IP network and is becoming increasingly common [TerenaVoIP]. Whilst there are many different signalling protocols and manufacturers, most forms of VoIP boil down to voice packets encoded at an ITU encoding method and being transported using Real Time Protocol or Real Time Control Protocol.

The signalling/control of the conversation and the voice packets themselves can be viewed as separate entities. The signalling, whether it is H.323, SIP, MGCP or something completely different, is used for call setup, transformation and teardown and thus whilst it is important that the delivery of these packets is guaranteed and timely, the criticality of this is not typically apparent to the end user. The voice conversations themselves are generally encoded using the ITU-specified G.711 or G.729 codecs. G.711 has a payload of 64Kbit/s, G.729 has a payload of 9Kbit/s. ITU G.114 examines delay in voice applications and specifies the following parameters for one-way delay.

| **Range in Milliseconds** | **Description** |
|---|---|
| 0-150 | Acceptable for most user applications. |
| 150-400 | Acceptable provided that administrators are transmission time and the impact it has on transmission quality of user applications. |

| | |
|---|---|
| Above 400 | Unacceptable for general network planning<br>However, it is recognized that in some exce<br>this limit is exceeded. |

RFC 4594 [RFC4594] suggests differing treatment for signalling and voice packets when considering QoS within an IP network. It goes on to define two service classes to identify the difference between control and voice; the Telephony Service Class and the Signalling Service class.

The Telephony Service Class specifies the following types of traffic to use it:

- VoIP (G.711, G.729 and other codecs)
- voice-band data over IP (modem, fax)
- T.38 fax over IP
- circuit emulation over IP, virtual wire, etc.
- IP VPN service that specifies single-rate, mean network delay that is slightly longer then network propagation delay, very low jitter, and a very low packet loss.

It also defines the traffic characteristics as:

- mostly fixed-size packets for VoIP (60, 70, 120 or 200 bytes in size)
- packets emitted at constant time intervals
- admission control of new flows is provided by telephony call server, media gateway, gatekeeper, edge router, end terminal, or access node that provides flow admission control function.

Packets matching these conditions should be treated as Expedited Forwarding (EF), with packets being marked either by the application itself or at the nearest router to the source of the application flow. The resulting network service offered to the application should be that of an enhanced best-effort service with a controlled bandwidth, very low delay and very low loss. The nature of voice traffic is that it does not tolerate packet loss.

The Signalling Service Class specifies the following types of traffic to use it:

- peer-to-peer IP telephony signalling (e.g. using SIP, H.323)
- peer-to-peer signalling for multimedia applications (e.g. using SIP, H.323)
- peer-to-peer real-time control function
- client-server IP telephony signalling using H.248, MEGACO, MGCP, IP encapsulated ISDN, or other proprietary protocols
- signalling to control IPTV applications using protocols such as IGMP
- signalling flows between high-capacity telephony call servers or soft switches using protocol such as SIP-T. Such high-capacity devices may control thousands of telephony (VoIP) calls.

It also defines the traffic characteristics as:

- variable size packets, normally one packet at a time
- intermittent traffic flows
- traffic may burst at times

- delay-sensitive control messages sent between two end points.

Packets matching these conditions should be treated as Class Selector 5 (CS5) with packets again marked either by the application itself or at the nearest router to the source of the application flow. The resulting network service offered to the application should be that of an enhanced Best Effort service with controlled rate and delay. Whilst applications in this service category do not normally react well to loss, the protocol usually has mechanisms to deal with this and the timely delivery of signalling packets is more important in this case.

Finally, in the next few years there may also be a need to consider supporting QoS on wireless networks to support the so called Voice over Wireless LAN (VoWLAN) model when the 802.11n standard is released.

## 6.3 Videoconferencing

Videoconferencing is particularly demanding on the network. It has a reasonably high bandwidth requirement, is real-time and is truly synchronous; few other applications have all of these needs. In this section we will consider some of the issues relating to network engineering for ITU-T compliant H.323 videoconferencing systems. Other types of videoconferencing systems may have differing requirements, particularly those that utilise multicast. For the purposes of this section we will consider ‘Quality of Service’ in a broad context which covers major network performance issues and relates to techniques that can be utilised to ensure good quality transmission and reception of videoconferencing media.

Overall ‘quality’ in relation to videoconferencing is a multi-faceted area, much of which is beyond the scope of this guide. Factors include quality of capturing equipment (e.g. cameras/microphones), quality of reproduction equipment (e.g. displays/speakers), conversion (e.g. analogue to digital and vice versa), coding/decoding algorithms and transmission.

The discussion here concerns standard definition (SD) videoconferencing. High definition (HD) videoconferencing is emerging and requires a faster network connection to operate effectively. Where screen shots have been taken from live videoconferencing systems some information may have been concealed to protect security.

### 6.3.1 General Videoconferencing Requirements

H.323 videoconferencing, like all IP-based applications, utilises a packet-switched network. By nature a packet-switched network is shared between many different services, for example e-mail, web browsing, etc. It can therefore be difficult to predict when network utilisation will result in congestion on the network. Congestion leads to packets being discarded, delayed or re-ordered, all having an effect on the perceived quality of the videoconference by the end user.

In H.323 the streams of packets that represent the media (sound and pictures) are implemented using User Datagram Packets (UDP). With UDP, if packets are lost they are lost forever, thus affecting the quality experienced by the receiving users. However, each packet has a unique identifier so H.323 does provide a mechanism for a receiving station to identify when packets are missing from a sequence or if the packets are received in the wrong sequence. The receiving station can therefore notify the transmitting station that this has

occurred. When this happens some videoconferencing equipment can assume that there is congestion on the network and attempt to downspeed, i.e. transmit at a lower call speed. It may then begin to increase the call speed until it again receives notification that packets are being lost, and then downspeeds slightly again to find an optimum level. This is adaptive rate control.

### 6.3.2 General Network Considerations

A comprehensive guide to engineering an IP network to support videoconferencing traffic is available for free download from JANET(UK)'s Video Technology Advisory Service [JANETVidAS]. The primary aim of network engineering for videoconferencing is to minimise delay jitter and loss. We need to minimise delay to ensure that the real-time nature of the conference is maintained. If the delay between transmission and playback at the receiving station is too high then natural communication between the participants will be stifled. Loss needs to be minimised as each packet carries part of the sound or picture and so the receiving station will not be able to reproduce the media accurately, leading to gaps in the sound and pixilation of the picture.

To help maintain quality, videoconferencing systems should be connected to switched network segments and not to repeated segments (e.g. avoid using hubs). UTP cabling is advised. Repeated or shared network segments usually utilise Carrier Sense Multiple Access (CSMA) techniques which can increase delay and affect jitter unpredictably. Experience has shown that some lower-end switches may not be suitable for use with videoconferencing systems even though they claim to switch at line speed. Investing in good quality managed switches is likely to ease deployment of videoconferencing and make the process of troubleshooting problems significantly easier.

One frequent cause of quality problems for videoconferencing is speed and/or duplex mismatches within the network, sometimes caused by relying on speed/duplex autonegotiation. Speed/duplex mismatches lead to packet loss. It is recommended that the entire network path serving videoconferencing equipment is manually set to the highest speed and duplex setting as applicable (most videoconferencing end stations are usually 100Mbit/s full duplex devices). In this respect it is important to consider the whole of the network path and not just the portion between the videoconferencing system and the first switch.

Managed switches generally allow the network administrator to set the speed and duplex settings. Where unmanaged switches are used, this is generally not possible and autonegotiation has to be relied upon. Similar issues apply to external media converters, e.g. those that convert between UTP and fibre. Most modern videoconferencing systems will also permit the administrator to set the speed and duplex settings in the configuration manually. If the switch can permit the network operator to signify that an end device and not another switch is connected to the port, this should be set to prevent issues such as spanning tree updates from interfering with operation of the port. A managed switch can also be useful during troubleshooting as the network administrator may be able to display port statistics that indicate packet loss and also display the actual speed/duplex settings currently in use. As each network device that the videoconferencing traffic has to travel through increases the end-to-end delay, limiting the number of devices will ultimately assist in ensuring that the videoconferencing signals are transmitted as quickly as possible. Where possible, physically connecting your videoconferencing system to a location near the boundary of your network

nearest to your upstream network provider minimises the amount of devices within the network that have to process the videoconferencing traffic.

In addition to traditional network devices such as switches and routers, special consideration should be given to security devices such as firewalls, as the inspection process can often cause additional delay. The ability of security devices to process H.323 videoconferencing traffic appropriately is beyond the scope of this document.

Videoconferencing is a synchronous technology; generally as much information that is received is also transmitted. As such, videoconferencing is best suited to a synchronous network. Other types of network technology may not be entirely for the deployment of videoconferencing, e.g. ADSL, although some modern videoconferencing systems allow you to specify that an ADSL network is in use and the system will receive better quality than it sends. Other general network considerations that should be considered in any shared network (such as domestic broadband) which may affect quality are issues such as overall bandwidth and contention ratio.

### 6.3.3 General bandwidth (or speed) considerations

Videoconferencing systems are usually capable of operating at different call speeds; generally the greater the call speed the better the quality of experience. Where part of the information is lost, videoconference systems do not have sufficient detail to reproduce the sound and pictures accurately and the user will notice glitches in the sound and pixilation of the videoconferencing image. One technique to help ensure a better user experience is to ensure that videoconferences are run at a speed that the network is likely to be able to handle and not at higher speeds. It should be noted that videoconferencing systems usually exceed the call speed selected, which is an average over time, so a 768Kbit/s call can actually peek at a rate above 768Kbit/s. Typically, the figure quoted by various bodies for this headroom ranges between 10% and 50% above the stated call speed, so a 768Kbit/s call could peek at 1.5Mbit/s.

It should also be noted that increasing the call speed does not always increase the quality, even on an appropriately engineered network, as videoconferencing systems that are capable of using the newer audio and video coding protocols are much more efficient. Therefore a call using H.261 at 768Kbit/s is not necessarily better than running a call using H.264 at 384Kbit/s.

Some videoconferencing systems offer error concealment whereby the videoconferencing system will attempt to mask any lost information from the end user. This improves the quality of the user experience but it should be noted that this is a reactive measure which only seeks to conceal lost information. Error concealment therefore reduces the true accuracy of the reproduced media but smoothes over any glitches that would otherwise be evident to the user.  6.3.4 Quality of Service for Videoconferencing

Local Area Networks aside, generally, the further away from the backbone of the JANET network the videoconferencing system is deployed the less bandwidth is available, e.g. for a multi-site organisation the lowest bandwidth connection is typically between the remote site and the main site (taking into account the connection between the main site and the Regional Network and also the connection between the Regional Network and the JANET backbone). If the connection between the remote site and the main site is congested, this is perhaps one of the biggest challenges to maintaining videoconferencing quality. A number of techniques are

available to support videoconferencing but will typically utilise DSCP values to allow a network to prioritise the traffic.

In order for H.323 to benefit from a QoS-enabled network, the H.323 packets need to be marked with the appropriate values in accordance with the QoS scheme implemented by the network operator. This can be achieved in several ways:

- Some (not all) videoconferencing endpoint equipment can mark the packets appropriately itself prior to placing them onto the network. This requires the administrator of the system to enter the values into the appropriate section of the videoconferencing device configuration. Sample screen shots from a selection of typical systems are included in Figure 6-2 below. Some videoconferencing systems also allow different values to be configured for audio, video and/or control packets. Many experts believe that the quality of the videoconferencing audio is more important than the video because without audio (or with poor quality) communication is difficult; other experts believe that for a truly reliable videoconferencing experience the quality of the video is at least as important as the audio. An advantage of allowing the videoconferencing system/application to mark packets with the appropriate QoS value is that only H.323 traffic is marked. This may be useful if a multi-purpose device such as a PC-based videoconferencing system is used, as the network will also be serving other applications which will not be marked for QoS.
- If the videoconferencing system connects directly to a network infrastructure device, such as a switch, it may be possible for the switch to remark packets that originate on a particular port to an appropriate QoS value. This relies on an assumption that the device is able to remark packets on a particular port. In this scenario it is likely that the device will only be able to mark all traffic on the port and will therefore not be able to discern between audio/video packets and any other traffic from that port.
- An item of network infrastructure equipment elsewhere in the network such as a core switch or a router may also be able to remark the QoS values of videoconferencing traffic. It should be noted however that, by this stage, the traffic may already have passed through several pieces of network infrastructure equipment without any prioritisation. Once the traffic reaches the device that remarks the traffic, it will have been aggregated with traffic from other devices. The device performing the remarking will therefore need to be capable of remarking based on source MAC or IP address. It is unlikely that such a device will be capable of remarking only H.323 traffic and so all traffic from the source device would be prioritised.
- If a dedicated VLAN is in use for videoconferencing devices then it may be possible to prioritise the VLAN by marking all traffic assigned to that VLAN appropriately or by using the VLAN ID as an identifier.
- Some videoconferencing infrastructure equipment such as application layer gateways, video proxies, routers, firewalls or firewall traversal solutions may have the ability to mark H.323 traffic with QoS values. This may be useful if the devices are deployed at the boundary of two networks which use different QoS schemes, as the device can be utilised to remark traffic heading in each direction (i.e. onto different networks) with an appropriate (different) QoS value for the destination network (according to their QoS scheme).

It should be noted that where a QoS scheme is in use on a network, it is also necessary to police other devices on the network to ensure that they are not marking their own traffic

inappropriately in order to receive priority on the network.

### 6.3.5 Janet Videoconferencing

The multipoint control units deployed to provide the Janet Videoconferencing service are either co-located at a C- PoP on the JANET backbone or are located at positions near to the JANET backbone network. The bandwidth utilisation at the locations of the MCU and in fact the entire backbone network are continually monitored and managed to ensure that they are suitably over-provisioned in relation to available bandwidth, to serve the needs of the entire JANET community comfortably.

Sites which are registered to use Janet Videoconferencing also undertake a regular Quality Assurance (QA) test or assessment. Network performance between the videoconferencing system and the Janet Videoconferencing MCU is one element of the test; the Janet Videoconferencing operator will provide details of any issues identified during the session. The automated Janet Videoconferencing-Check system which is available to all registered Janet Videoconferencing sites also provides information on network performance and can be used to identify when issues arise. The statistics are displayed via the videoconferencing system as part of the session and also via the integrated Janet Videoconferencing booking


Polycom PVX


Polycom VSX8000

*nferencing applications*

Tandberg MXP

[2]

The MCUs which provide the Janet Videoconferencing service currently do not mark outbound IP traffic with any QoS values. The IP packets originating from the MCU are marked as ‘default’. As the traffic destined for videoconferencing systems originates from the MCU, all videoconferencing traffic will enter the JANET backbone marked as ‘default’ traffic. It will therefore arrive at the boundary of the Regional Network marked as such and, if not remarked by the Regional Network Operator, will arrive at the organisational boundary with its original QoS values.

It is possible to obtain a list of the MCU IP addresses from Janet Videoconferencing. This would permit each Regional Network Operator or each organisation to remark traffic that has originated from the Janet Videoconferencing MCUs with a QoS value which is appropriate for their own scheme. This permits different organisations and different Regional Network Operators to deploy their own QoS marking scheme and to prioritise videoconferencing traffic appropriately. Regional and organisational network operators therefore have the ability to deploy QoS based on a model that best suits their network, for example using over provisioning, static provisioning or some form of dynamic allocation.

## 6.4 Streaming Applications

Streaming applications typically represent the delivery of multimedia content to heterogeneous user terminals from a remote source. This class is unique as, in contrast to the other QoS applications listed here, they do not involve two-way communication and do not necessarily depend on the (near) real-time delivery of ‘live’ streams. Moreover, it is harder to categorise this application class as it must support different terminal devices with requirements ranging from low-quality video for mobile terminals through to high-definition content for large, public displays.

This section will attempt to set out the general issues for supporting streaming applications including the types of content that might be supported for various terminals, and the general network requirements in each case. We will then attempt to broadly define the QoS requirements for this class of application and the tolerances involved in this case. Finally, this section will outline some optimisation techniques that may be employed to supplement QoS provisioning for streaming applications.

### 6.4.1 General Multimedia Streaming Requirements

Multimedia streaming applications are categorised by the one-way delivery of content to a user terminal from a remote content source. Due to the recent explosion of feature-rich multimedia mobile devices (e.g. IPhone, etc.), a spectrum of user terminals will expect to make use of multimedia streaming ranging from limited-capacity mobile devices, through reasonably-powerful PCs, laptops and home entertainments systems, to high-definition public displays, etc. These heterogeneous terminals will implement a range of media player applications supporting a variety of media codes and formats.

A number of popular media codecs are currently available and in common use representing a range of qualities and media formats, including MPEG-2, MPEG-4, H264, DivX/XVid [VideoCodecs] and more. These codecs are supported by the majority of both open and propriety media players such as QuickTime, Real Media, MPlayer, etc. and so are available on most operating systems and user terminals. The requirements of these codecs all vary

slightly in terms of their performance but the defining factor (in terms of QoS requirements) will always be the quality of the content being delivered.

The general network requirements for multimedia streaming are a sufficiently provisioned end-to-end path from the source to the user terminal but, again in contrast to the other classes discussed here, this provisioning is not necessarily straightforward since the endpoints cannot be easily predetermined in advance (at least on the terminal side) and may need to scale depending on the number of users being supported.

### 6.4.2 QoS Requirements

As discussed above, streaming applications will have varying bandwidth requirements depending on the content (format) being delivered and the user terminal in question but the following rates give an indication of the bandwidth usage based on several popular video codecs:

| | MPEG-2 (Mb/s) | MPEG-4 DivX (Mb/s) | MPEG- |
|---|---|---|---|
| Low Quality (Mobile Terminals | 4 | 2 | |
| Medium Quality (SD video) | 15 | 8 | |
| High Quality (HD 720) | 60 | 20 | |
| High Quality (HD 1080) | 80 | 20+ | 1 |

Regardless of these bandwidth requirements, streaming applications will all exhibit the following QoS requirements [Szigeti]:

- packet loss at no more than 2-5%
- latency should not exceed 4-5 seconds
- no specific jitter requirements depending on buffering capabilities.

There are also various suggestions when it comes to DSCP marking for streaming applications, including CS1 for entertainment content and CS4 for general-purpose content, but in a general sense, anything at a better than Best Effort rate should be sufficient.

### 6.4.3 Optimisations

Fortunately, since streaming applications are uni-directional and do not necessarily deal with real-time content, a number of network optimisations can be applied to improve their

performance and scalability to supplement QoS provisioning. For example, some form of multicast can be deployed to support synchronous content delivery to multiple users (but this has scalability issues in itself) and some form of multimedia caching can be used both within the network and site to supplement the delivery process. In the commercial area, content distribution networks such as Akamai [Akamai] can be utilised to reduce network load and resource contention.

## 6.5 Control Traffic Prioritisation

When building a QoS-enabled network, it is important to consider not only the user-to-user traffic running over it but also the router-to-router traffic, since if this is interrupted then all traffic will be disrupted regardless of priority or classification, rendering QoS irrelevant. RFC 791 [RFC791] specifies two precedence values to be used to mark such traffic: Internetwork Control (precedence 6) and Network Control (precedence 7 or CS7). RFC 791 states that:

> ‘The Network Control precedence designation is intended to be used within a network only. The actual use and control of that designation is up to each network. The Internetwork Control designation is intended for use by gateway control originators only. If the actual use of these precedence designations is of concern to a particular network, it is the responsibility of that network to control the access to, and use of, those precedence designations.’ RFC 2474 [RFC2474] and RFC 2475 [RFC2475] enhance the 3-bit precedence value definitions into 6-bit DSCPs in a manner that maintains backward compatibility with precedence 6 mapping to DSCP CS6 and precedence 7 mapping to CS7. It should be noted, as stated within RFC 2474:
>
> ‘IP Precedence is widely deployed and widely used, if not in exactly the manner intended in [RFC791]. This was also recognized in [RFC1122], which states that while the use of the IP Precedence field is valid, the specific assignment of the priorities in [RFC791] were merely historical.’

Looking at a specific router vendor as an example of how network control in the QoS context is addressed, we see for example that Cisco® IOS ‘marks Interior Gateway Protocol (IGP) traffic such as Routing Information Protocol (RIP/RIPv2), Open Shortest Path First (OSPF), and Enhanced Interior Gateway Routing Protocol (EIGRP) to DSCP CS6’ [CiscoQoS1] (note that in this quote it mentions DSCP CS6, but actually Cisco® works with precedence values and so in this case are actually setting precedence 6 albeit the same binary value). In addition to this, Cisco® also uses a proprietary, non-administratively configurable mechanism within their routers for granting internal priority of important control data called PAK_PRIORITY [CiscoQoS2] (a mechanism that encapsulates this traffic within the router with a small header that contains a PAK_PRIORITY flag to indicate the relative importance of control packets). Cisco® does not use the PAK_PRIORITY with EGP traffic (e.g. BGP), but does still set precedence 6.

Implementing QoS support for network control traffic will vary from vendor to vendor and the specific routing/switching platforms involved (and may not be necessary in some environments where some network equipment internally prioritizes such traffic or places it directly to the head of an output queue, for example Cisco® IOS-based routers where ‘IGPs are usually adequately protected with the Cisco® IOS internal PAK_Priority mechanism’ [CiscoQoS1]), but in general this traffic could be given its own queue, or threshold within a queue, or even be reclassified to an existing ‘higher-than-BE’ class (e.g. Premium IP) to

ensure that a small amount of the bandwidth on the link is always available for this type of traffic (Cisco® recommends that ‘EGPs such as BGP have an explicit class for IP routing with a minimal bandwidth guarantee’ [CiscoQoS2]).

## 6.6 Grid / E-Science Applications

The main motivation behind grid computing is to harness geographically dispersed resources, such as processing power and storage space, to achieve computationally intensive tasks. The users of such systems are allowed to employ resources seamlessly that they neither own nor have direct access to, but are instead owned by other organisations or individuals. This offers a large communal pool of resources to be used for resourceintensive computations. Examples of such computations include simulations, data analysis, data management and virtualisations.

Similar to other distributed applications, scheduling tasks in a grid computing environment requires the allocation of numerous remote resources simultaneously. However, grids are different from other distributed systems in that their resources are neither centrally-owned nor controlled. Nevertheless, the expectation for the performance of grids is typically just as high as those of cluster systems. This raises the need for strong guarantees of network performance in grid networks.

The performance of data delivery has been and still is a driving force in networking research. Demands for shorter delays, less jitter, more bandwidth etc. are common for many application models, yet it is particularly important for grid applications. Both advance and on-demand resource reservation in grids is essential [Foster] as without such abilities, the only thing that could guarantee high network performance is over-provisioning. Not only is this an expensive measure but it is also only a temporary fix as application demands are constantly increasing. Nevertheless, over-provisioned networks are still prone to unpredictable behaviour. On top of the high performance requirements, grid applications need to predict the duration of each task accurately, whether it is local or remote. This need for determinism further raises the QoS requirements of grid applications.

### 6.6.1 QoS on Grid Networks

The need for QoS management in grid networks extends to various levels: specifying the application requirements, mapping them to the resource capabilities and availability, agreeing an SLA with the resource owners (called virtual organisations, VOs) and with the clients (application users), and inspecting/monitoring the QoS parameters whilst resources are allocated [Al-Ali]. Different grid applications may have different network requirements, such as minimum bandwidth, time-sensitivity, implicit data paths and added-on transport layer services. Moreover, the requirements of different flows of the same application vary. Hence, grid applications may require end-to-end QoS provisioning on a per-flow basis.

Different techniques have been suggested and used. Some techniques, such as that in [Yang], integrate DiffServ and IntServ: DiffServ addresses the diverse flow demands, while IntServ is used to solve the end-to-end problem. Other techniques use an active networks paradigm to reserve network resources dynamically in order to obtain the highest possible overall satisfaction level. In [Munir], the shortage of available bandwidth triggers the renegotiation of the ‘average required rate’ for each flow. Other techniques use requirements aggregated over virtual organisations. In [Keahey], resource reservations are made on a VO level, not on an individual user level. The responsibility of the VO using such a system is two-

fold: to collect information about the resource reservation requirements of its users, and then to negotiate the aggregate demands with the resource providers. This promotes the scalability of the grid, as it is much easier for resource providers to deal with a small number of VOs rather than a large (and changeable) number of individual users. This model also allows a certain level of resource reservation and usage separation between different VOs. Further, this model creates a distinction between the concerns of VOs and those of resource providers. VOs provide a level of aggregation at the application level which is positive in terms of QoS provisioning. However, ultimately, unless the network is over-provisioned, QoS parameters specified using one of the above techniques will need to be mapped onto a network-level solution such as DiffServ. G-QoSM is a QoS framework for service-oriented grids. It supports QoS-matching resource and service discovery, QoS guarantees and SLA management, and QoS management of allocated resources. G-QosM introduces three levels of QoS delivery: guaranteed, controlled load, and best effort.

The Globus Architecture for Reservation and Allocation (GARA) is an architecture that provides uniform end-to-end QoS and is used to reserve and allocate heterogeneous collections of resources. GARA complements the Globus Resource Management Architecture by providing mechanisms for reservation and allocation of different resources, including network, processing, storage and other resource types. GARA relies on the mechanisms implemented by the Globus Resource Management Architecture to manage and co-allocate resources using agents.

Sensitivity to delay and inter-packet arrival times is an issue for many grid applications as the level of parallelism between processes is such that their overall performance could be affected by high delay and/or jitter. MeteoAG is one such application which is being used for forecasting Alpine watersheds and thunderstorms based on parametric measurements from data collection points deployed in the field. Automatika is another grid application that imposes a deadline on the delivery of non-multimedia-stream data packets. In this case, the application relies on the prompt return of results from web service invocations. In both of these applications, the need for guaranteed packet arrival times is crucial as data that arrives late has to be discarded in order to process the data that is due. QoS could help in this case by reserving sufficient network resources to ensure that delay and jitter remain within the limits such that they do not deteriorate the overall performance of the application.  6.6.2 Example - Large Hadron Collider at CERN LHC, or the Large Hadron Collider, is a particle accelerator and collider and is also the world’s largest machine, costing a total of £2.6 billion. Located at CERN near the Switzerland-France border, LHC is expected to see its first particle collisions in May 2008. The experiment is planned to run for nine consecutive months and then cease for three months before commencing again. During its first active period, LHC is expected to trigger huge amounts of raw data in the neighbourhood of 10 petabytes. This harvested data will then be processed by the Grid and the results obtained will be compared to those of simulated experiments. ATLAS (A Toroidal LHC ApparatuS) is one of the five particle detector experiments that will run at LHC, and is ‘the largest volume detector ever constructed for Particle Physics’ [CERNLarge]. ATLAS brings together almost 2000 scientists from around the world.

The Grid infrastructure for this project, the LHC Computing Grid (LCG) Project [LHCGrid], is a collaboration between 165 scientific organisations, universities and government bodies connected together using a dedicated 10Gbit/s lightpath. These sites are organised using a three-tier distribution architecture. Tier 0 is the Particle Physics laboratory at CERN where part of the data analysis will take place. However, all Particle Physics aside, the main function

of the laboratory is to farm out the raw data over the Grid to the Tier 1 sites.

There are ten Tier 1 sites scattered across France, Germany, Italy, Japan, the Netherlands, the UK and the US. Each of these sites has a cloud of Tier 2 sites associated with it. Tier 1 sites are responsible for splitting up the raw data they receive (from Tier 0) between their respective Tier 2 sites. Each Tier 2 site processes the data upon receiving it, stores the results locally on magnetic disk, and sends a copy of the results to its respective Tier 1 site where it is stored on tape. Hence, there are always at least two copies of every result file in the Grid.

Most data is sent as chunks in the order of (a few) gigabytes at predefined time intervals. The knowledge of such transmission periods makes it possible to reserve resources in advance. Sufficient network resources need to be reserved in advance at the scheduled times in order to maximize the performance of the data transfers. DiffServ forwarding mechanisms (EF and AF) could be used in this situation to make the necessary resource allocations [Kulatunga].

Additionally, there are some datasets of special importance that need to be transferred spontaneously and regardless of the pre-scheduled times. One such situation is called ad hoc submission, where data processing results are returned from Tier 2 to Tier 1 whenever ready. Such situations may necessitate on-demand resource reservation in the network.

## 6.7 Less Than Best Effort Applications

LBE is a service with a lower priority than that of Best Effort (BE) and with a very small minimum bandwidth guarantee (typically around 1% of the link bandwidth) configured such that during periods of congestion, LBE-marked traffic is ‘squeezed down’ to make way for higher priority traffic. This minimum guarantee, however, ensures that there is always a little bandwidth available to help keep TCP sessions alive. The LBE service used within JANET is based largely on the Internet2 Scavenger Service which in turn was based on ‘A Lower Than Best-Effort Per-Hop Behavior’ internet-draft by R. Bless and K. Wehrle [LBE] and ‘A Bulk Handling Per-Domain Behavior for Differentiated Services’ internetdraft by B. Carpenter and K. Nichols [DSPHB]. Internet2 defines the scavenger service [Scavenger] as creating:

> ‘... a parallel virtual network with very scarce capacity. This capacity, however, is elastic and can expand into capacity of the normal best-effort class of service whenever the network has spare cycles. The expansion happens with a very high time granularity: everything not used by the default class is available for the scavenger class.’

RFC 3662 [RFC3662] (which followed on from Internet2’s scavenger service and the aforementioned internet-drafts) presents a more generalised definition of LBE as being:

> ‘intended for traffic of sufficiently low value (where “value” may be interpreted in any useful way by the network operator), in which all other traffic takes precedence over LE [Lower Effort] traffic in consumption of network link bandwidth. One possible interpretation of “low value” traffic is its low priority in time, which does not necessarily imply that it is generally of minor importance. From this viewpoint, it can be considered as a network equivalent to a background priority for processes in an operating system. There may or may not be memory (buffer) resources allocated for this type of traffic.’

LBE would therefore be useful for TCP-based applications that are not time-sensitive, allowing

these applications to make use of free bandwidth. Examples of such applications are:

- mirroring services (such as the UK Mirror Service [UKMirror])
- remote backups
- grid transfers.

In addition to the above, there are also other possibilities for which the LBE service might be put to use, for example:

- to mark ‘out of character’ traffic in an attempt to minimize the effects of undesirable traffic such as seen with DoS attacks. Where IDS and similar techniques are available and sufficiently evolved, if ‘out of character’ traffic is identified but is not at that point identifiable as undesirable (i.e. a DoS attack, a worm or similar) then an option exists to let the traffic onto the network, at least temporarily until further identification is possible, whilst minimizing its impact on the rest of the network
- to allow for a flexible, non-stringent policy control of non-business applications. Where applications exist that, whilst they provide no contribution to organizational objectives, are tolerated rather than blocked (e.g. gaming applications, streaming video services such as YouTube), marking of such traffic as LBE might be deemed reasonable (tolerated when free bandwidth is available but squashed during times of congestion).

In phase one of the JANET QoS project, LBE was investigated to determine that it behaved as expected and that the implementation on existing JANET/Regional/campus networks might allow it sensibly to be adopted into a production service. Net North West and the University of Manchester tested LBE in conjunction with other traffic classes under test conditions [MANP1] whilst Southampton (in partnership with Imperial College), focused on LBE, choosing a real application, i.e. AccessGrid®, as the basis of their experiment to ‘show that LBE flows could be serviced by the network without disrupting the “regular” applications in an everyday network environment (i.e. outside of a more clinical testbed environment)’ [SMTPNP1].

In phase 2, a JANET-based service was sought whose traffic might sensibly be classified as LBE. Kentish MAN identified the UK Mirror Service as a candidate and so (in partnership with NNW) configured and tested this traffic, using the conclusions and lessons drawn from the phase one trials as a basis for the configuration [KentP2].

---



**Links**
[1] http://community.ja.net/system/files/images/tg-qos-6.1.jpg
[2] http://community.ja.net/system/files/images/tg-qos-6.2.jpg