



Report of Quality of Service Think Tank

July 2001

**ND/STG/QOS/DOC/010
Version 1.0 Release**

UKERNA manages the networking programme on behalf of the higher and further education and research community in the United Kingdom. JANET, the United Kingdom's academic and research network, is funded by the Joint Information Systems Committee (JISC).

For further information please contact:

JANET Customer Service

UKERNA
Atlas Centre, Chilton, Didcot
Oxfordshire, OX11 0QS

Tel: +44 (0) 1235 822 212
Fax: +44 (0) 1235 822 397
E-mail: service@ukerna.ac.uk

Copyright:

This document is copyright The JNT Association trading as UKERNA. Parts of it, as appropriate, may be freely copied and incorporated unaltered into another document unless produced for commercial gain, subject to the source being appropriately acknowledged and the copyright preserved. The reproduction of logos without permission is expressly forbidden. Permission should be sought from JANET Customer Service.

Trademarks:

JANET[®], SuperJANET[®] and UKERNA[®] are registered trademarks of the Higher Education Funding Councils for England, Scotland and Wales. The JNT Association is the registered user of these trademarks.

Disclaimer:

The information contained herein is believed to be correct at the time of issue, but no liability can be accepted for any inaccuracies.

The reader is reminded that changes may have taken place since issue, particularly in rapidly changing areas such as internet addressing, and consequently URLs and e-mail addresses should be used with caution.

The JNT Association cannot accept any responsibility for any loss or damage resulting from the use of the material contained herein.

Availability:

Further copies of this document may be obtained from JANET Customer Service at the above address.

This document is also available electronically from:

http://www.ja.net/development/qos/qos_tt_report.pdf

Contents

Executive Summary	3
1 Introduction	4
1.1 Consultation	4
1.2 Quality of Service.....	5
1.3 Approach.....	5
1.4 SuperJANET & its constituency.....	6
1.4.1 JANET, SuperJANET and UKERNA	6
1.4.2 Current Network Operation	6
1.4.3 Regionalisation.....	6
1.4.4 Extending access	6
1.4.5 e-Science and JANET.....	6
2 Requirements.....	8
2.1 Videoconferencing	8
2.2 Other video-based services.....	9
2.3 Voice over IP.....	9
2.4 Outreach aspects.....	10
2.5 GRID and e-Science	10
2.5.1 Astronomy and astrophysics	11
2.5.2 Biological Sciences.....	11
2.5.3 Particle Physics	12
2.5.4 Access Grid.....	13
2.6 Managed Bandwidth Service.....	13
References for section 2	14
3 Technology review.....	15
3.1 Elements of QoS technology	15
3.1.1 Integrated Services (IntServ).....	15
3.1.2 Resource Reservation Protocol (RSVP)	15
3.1.3 Differentiated Services (DiffServ)	17
3.1.4 Less than best effort.....	20
3.1.5 QoS Support in the End-System.....	20
3.2 Provisioning and traffic engineering	21
3.2.1 Multiprotocol Label Switching (MPLS).....	21
3.2.2 Bandwidth Brokering	24
3.3 Traffic engineering via application engineering and proxies	24
3.4 Summary	25
3.4.1 IntServ & RSVP	25
3.4.2 DiffServ.....	25
3.4.3 LBE.....	25
3.4.4 MPLS.....	26
3.4.5 Bandwidth Brokering	26
3.4.6 Conclusions	26
References for section 3	26
4 Policy.....	28
4.1 Requirement	28
4.2 Inter-domain issues.....	28
4.2.1 Requirements	28
4.2.2 Provisioning	29
4.2.3 Other projects.....	30
4.3 Policy enforcement.....	32
4.4 Initial policy model	32
5 Proposed 'road map'	34
5.1 Technology Recommendations	34
5.2 Prototype service testing.....	35
5.2.1 Videoconferencing.....	35
5.2.2 Voice over IP.....	36
5.2.3 Traffic Engineering for GRID & former MBS applications	36

5.2.4	<i>Outreach</i>	36
5.2.5	<i>Advance booking: on-demand vs. scheduled use of Premium IP</i>	37
5.2.6	<i>Monitoring</i>	37
5.2.7	<i>Timescales</i>	37
5.3	Issues for on-going study.....	37
5.3.1	<i>Authentication & authorisation</i>	37
5.3.2	<i>Policy model development</i>	38
5.3.3	<i>GMPLS</i>	38
5.4	Related issues	38
5.4.1	<i>Firewalls</i>	38
5.4.2	<i>Privacy & encryption</i>	38
	Glossary.....	39
	Appendix A — Think Tank Membership	42
	Appendix B — Acknowledgements	43
	Annex — Terms of Reference.....	44

Executive Summary

This report addresses the issues of how to introduce Quality of Service support into SuperJANET4 in order to support services sensitive to network delay and jitter. It also considers the related question of introducing traffic engineering support both for certain of these applications and for requirements associated with specific programmes in the UK Research and Higher and Further Education sectors. The report is divided into five sections, the first of which serves to introduce the general nature of the requirement, the method of working adopted by the Think Tank, and its Terms of Reference (annexed to this report).

Section 2 surveys the requirements identified by the Think Tank in consultation with the three sectors above. They cover videoconferencing and videocasting, voice over IP, GRID and e-Science, issues associated with the need to extend JANET services beyond the conventional workplace, and issues arising from the closing of the Managed Bandwidth Service formerly made available over the SuperJANET3 ATM infrastructure.

Section 3 provides a review of technologies available for QoS and traffic engineering support. In respect of QoS support, the Differentiated Services technology is proposed as the basis for initial deployment of QoS support, primarily on the grounds of avoiding potential scaling problems in the network core, and also because it is not predicated upon deployment of support for signalling in end-user systems. Multi-Protocol Label Switching is recommended as the major tool to assist with traffic engineering.

Section 4 provides an initial assessment of some of the issues relating to policy, including particularly the need to address multi-domain networking, both within the UK network and in international networking. It should be pointed out that the mechanisms for QoS and traffic engineering support proposed here are consistent with those being studied and adopted in Europe and the USA.

The main recommendations of the report are contained in Section 5. Section 5.1 consists of a set of 'Technology Recommendations', and Section 5.2 proposes a number of application-led areas which together constitute a proposed programme to be the subject of a call for proposals to carry through development and testing of an initial 'QoS-enabled' network service. Section 5 also draws attention to some areas which are related to the introduction of QoS but not specifically within the remit of the Think Tank.

1 Introduction

Integral to SuperJANET4 is the development programme necessary to ensure that the network infrastructure is capable of meeting the expanding range of requirements of the education and research sector. A draft document outlining the major elements of this programme was published by UKERNA in December 2000: *SuperJANET4: Development of Network Support for Applications in Learning, Teaching and Research* (J Sharp). Specific actions proposed by this report were:

“to harness the expertise available within the academic community and from the supplier of the switching equipment used within SuperJANET4 to create a roadmap for the development of network QoS;”

and:

“to trial at an early stage the emergent technologies, albeit they will at the outset offer only large granularity of control.”

In response primarily to the first of these, a ‘think tank’ was established under the chairmanship of Chris Cooper, Rutherford Appleton Laboratory, charged with three objectives:

1. *assessing the requirement for QoS on JANET;*
2. *developing a policy framework on which to implement QoS on JANET; and*
3. *providing initial recommendations for the technical mechanisms by which to implement QoS on JANET*

The terms of reference containing some additional background are annexed; and Appendix A contains the list of members of the think tank.

1.1 Consultation

Whilst it is now generally well understood that there is a need to deploy network support for a wider range of services than hitherto, the think tank consulted with a number of people from the HE and FE sectors in order to gauge requirements more specifically and to complement the representation already present within the think tank membership.

In particular, the think tank has consulted with representatives from

- the Scottish MANs and the Scottish MANs Video Conferencing Network (SMVCN);
- the Welsh MANs and Welsh Video Network project (WVN);
- Network North West; and
- the Further Education sector, including representatives from six FE Colleges, NILTA, and Regional Support Centre Technical Advisor.

The think tank has also consulted a variety of documents, on requirements:

- an early draft of *Bandwidth requirements of Scottish HE networks* (SHEFC C & IT Programme);
- Report of Infrastructure Task Group to Further Education Information and Learning Technology Committee;

and technical issues:

- GÉANT Deliverable D9.1: *Specification and Implementation Plan for a Premium IP service;*
- SEQUIN Deliverable D2.1: *Quality of Service Definition;*
- a substantial collection of documents made available by Cisco Systems Ltd;
- H.323 IP Videoconferencing demonstrator (‘VIP demo’) project deliverables.

Selected papers and RFCs of the IETF are also referenced in Section 3, Technology Review.

1.2 *Quality of Service*

Currently, JANET is a *single-service* network: all traffic is treated in the same way, essentially equally and all packets receive so-called *best effort* service, whereby packets are accepted for delivery to the destination specified in the header, but no guarantees of either an absolute or probabilistic nature are made by the network in regard to delivery. All that the network undertakes is to make ‘best efforts’ to deliver a packet. It does not undertake to deliver (packets may be dropped to relieve congestion), to deliver in order (packets may not all follow the same route, and routes are not necessarily of the same length), to deliver undamaged (bit errors are neither detected nor corrected by the network), or to deliver to any particular timescale (either in terms of end-to-end delay or delay variation between one packet and the next). If timing aspects are not crucial, recovery of packets lost, damaged, or misordered in transmission can be achieved by action, including retransmission, undertaken by end systems. If overall delay is not crucial, receiver buffering can compensate for delay variation (jitter).

However, if end-to-end or round-trip delay is important, then measures need to be taken to limit queue lengths everywhere in the network switching elements or, if not all traffic is sensitive to delay, to ensure that delay-sensitive traffic receives suitable preferential treatment. That a variety of service regimes may be offered by a single network is the central concept of a multi-service network. While a suitable level of provisioning can be used to alleviate some of the effects of queuing delays, in a general purpose network with demand-driven patterns of traffic, provisioning alone is insufficient to avoid the effects of temporary output port congestion, nor in practice is it generally possible to ensure a uniformly adequate level of provisioning throughout the network, particularly since it is rarely the case that the network is provisioned through a single administration or operated by a single management domain.

The definition of the range of services to be offered by a packet network, together with the techniques deployed within the network to deliver this range of services, can be said to constitute *quality of service* within a packet network. It is with this restricted interpretation of the topic that this report is primarily concerned.

Wider interpretations of the term quality of service may include other attributes of network service, such as availability. Although this report does not specifically address this issue, it is recognised that this is now a crucial aspect of network service, brought about in the first instance by an increasingly common assumption in current working practice that the network is ‘always there’. In the process of putting together this report, it has become clear that two additional features of emerging usage will serve to emphasise still more strongly the need for high availability of the network.

- Network endeavours such as the GRID concept, in which large numbers of components are required to act in concert in ‘loose’ (elastic?) synchronism, also place constraints on the network itself: it is required that many links and components are all available simultaneously, implying that a greater probability of availability is required of the network components than might be necessary for less complex use of the network.
- Scheduled, synchronous activities amongst groups of people carry with them the implicit assumption that the complete infrastructure will be available with very high probability at the appropriate time. For activities such as teaching, research collaboration, and so on, mediated by the network, it is a *sine qua non* that the network be available. These types of emerging uses serve to emphasise strongly the necessity for a highly available network infrastructure.

Whilst not specifically addressing this issue, the think tank recognises the importance of this requirement and endorses the necessity of meeting it at all levels in the network infrastructure.

1.3 *Approach*

The overall approach taken by the think tank has been application service led. A number of existing and emerging end-user requirements have been identified, explored, and confirmed. These are used as a set of objectives to provide a firm context for guiding the choices and recommendations made by the think tank. It is also suggested that the same generic approach should be taken to developing and deploying network quality of service support. By concentrating on engineering the support of specific application services the developments can be focused on achieving these specific objectives, as well as ensuring that the end results meet the expressed needs of users of the JANET portfolio of services.

1.4 *SuperJANET & its constituency*

1.4.1 *JANET, SuperJANET and UKERNA*

JANET (the Joint Academic NETwork) is the wide-area network that was originally created in 1984 to serve the needs of the higher education and research sectors in the United Kingdom. In 2000, following the successful establishment of FEnet in Wales (1996), this constituency was extended to include the whole of the further education sector. UKERNA is responsible for overall management, operation and development of JANET through a service level agreement that defines the services and service levels provided to its customers. JANET has grown from an X.25-based network in 1984 connecting 50 sites, to an IP-based network currently supporting more than 700 direct connections. *SuperJANET* was the name given to the broadband capability of JANET, now the backbone component, currently SuperJANET4, which came into operation at an initial 2.5Gbps in April 2001, superseding its 155Mbps predecessor, SuperJANET III.

1.4.2 *Current Network Operation*

The delivery of JANET services to customer sites is provided largely through a number of regional networks developed since 1997. SuperJANET provides international and national connectivity through its links to the USA and mainland Europe, together with peering connections to commercial networks through the *LINX* (London InterNet eXchange). The main operational centre for JANET is the Network Operations and Support Centre (NOSC), which operates under contract to UKERNA, and is situated at the University of London Computer Centre (ULCC).

1.4.3 *Regionalisation*

Over the last few years, a number of initiatives by the Higher Education Funding Councils have led to the establishment of regional academic networks in most areas of the United Kingdom. Some of these networks are being (or are about to be) re-procured, and at least one is currently under construction. These networks are expected to develop over the coming years in line with the various regional agendas that are being established, so that each becomes a major player in the delivery of education and research services within its region. UKERNA manages the operation of the backbone and, through contracts with the entities managing each regional network, is responsible for the levels of service provided by each regional network. The backbone is also the means of providing external (Internet) connectivity to each regional network.

1.4.4 *Extending access*

One of the original visions of SuperJANET4 was to extend access to the network beyond the traditional base in HE, RC, and FE institutional sites into the home and external workplaces. To achieve this, both fixed and wireless network technologies will be utilised. Users at the end of these links will still expect access to the applications and services that they would normally use from the traditional working place, which implies that QoS services will also need to be extended.

1.4.5 *e-Science and JANET*

In 2001 funding was made available from the government for an e-science initiative. E-science encompasses the GRID concept and a decision was made that JANET will be the infrastructure on which the UK GRIDs will be built.

As of mid-2001, the e-Science programme is in the process of being formed. The demands the programme will make upon the network infrastructure can at this stage only be described in general terms, based on what is known generally about the GRID concept and the formative work so far within the scientific communities intending to exploit GRID technology. Section 2.5 describes this requirement, so far as it is currently known.

The traditional constituency of JANET characterised as the education and research institutions already embraces the major fraction of the community anticipated to participate in the programme. However, the whole emphasis of the technology and mode of operation embodied in the GRID concept is such as to require many sites and resources to co-operate over extended periods on a single task or a closely related set of tasks. Initial indications are that experimentation and exploitation will be pursued primarily amongst the research community. Whether successful aspects will be taken up within the teaching context remains to be seen.

The impact on the network support of this requirement for multiple participation in a single activity is likely to be far-reaching, and has already been touched upon in Section 1.2. It seems likely that it will stress the

interdomain aspects of network service, since many domains will be required to operate in concert to enable effective GRID operation. Within the UK these domains include as a minimum most campuses, all MANs, and the SJ4 backbone. For some communities, it also includes international domains. In consequence of this, this report takes into consideration (some of) the corresponding activities in Europe and the USA in developing the introduction of quality of service.

2 Requirements

This section describes the end-user requirements for those applications and services which require a level or quality of network service other than the traditional 'best effort' in order to function satisfactorily.

2.1 Videoconferencing

Consultations carried out by the Think Tank have confirmed a strong demand from users and their representatives for a well supported, Internet-based videoconferencing (VC) service. It is clear from our studies, and those of the VIP-demo project, that there is a requirement for more than one class of VC system. Many users, for instance the Welsh Video Network (WVN), will be installing high quality, studio-based IP VC network facilities. The Scottish MANs Video Conferencing Network (SMVCN), along with the Scottish MANs, will shortly be reprocurring and will need to replace the existing ATM-based VC infrastructure with one based on IP. The WVN, SMVCN, and other similar users will have made substantial financial investment and will expect their VC services to operate with high quality and high reliability. Such users will expect the VC facilities to be critical services for the functioning of teaching and research activities. A large number of other users are likely to become frequent users of VC services which are provided as facilities of desktop computers. For instance, recent installations of Microsoft operating systems include a copy of NetMeeting as a free component, installed by default. Such users are likely to be less demanding in terms of their requirement for the transmission of high quality video and audio but may grow in future to dominate studio usage in terms of numbers of users. Other users will have expectations between the above extremes. It is evident that VC services need to be supported at multiple levels of quality.

The VIP-demo project and others have identified certain network characteristics as crucial for the successful support of high quality videoconferencing. The characteristics of latency, jitter and loss are particularly important. From a user viewpoint, only application end-to-end measures of these parameters are relevant: latency is latency, a user does not care whether this occurs within end equipment, within the links of the network, or within network plant. There are both user reasons and technical reasons for needing latency to be small. Technical reasons include meeting the requirements for successful echo suppression; user reasons include the important requirement to maintain a truly interactive feel to inter-personal communication.

Within the context of interactive or conversational speech, there is a range of opinion about suitable goals for values of end-to-end (one-way) application delay. Any such figure is a compromise between application usability and technological capability. ITU guidelines reported in [1] broadly categorise as follows the acceptability of one-way delay as perceived by an end user:

- 0 – 150 ms: good interactivity;
- 150 – 400 ms: tolerable;
- more than 400 ms: unacceptable.

Consistent with this, Vegesna [2] suggests, in the context of IP QoS, that a target of 100 ms for user-perceived end-to-end, one-way delay is a desirable goal to maintain the interactive nature of speech communication. Kumar *et al.* [3] suggest that to maintain full-duplex voice conversation it is desirable for the one-way delay to be kept below 300 ms.

In principle, it may be expected that the same broad categorisations might apply to videoconferencing. However, a major factor in video transmission is the contribution to delay introduced by compression. For example, for current generation H.323 codecs, the encoding/decoding delay is approximately 240ms [4]. In addition to this, each multipoint control unit (MCU) in the H.323 architecture contributes a delay of 120 – 200 ms. Even for audio, depending on the degree of compression and the size of packets used, there can be appreciable compression and packetisation delay.

Two issues arise in translating these end-user application requirements into network performance requirements. Size of playout buffer may be used to trade delay for jitter. This choice is dependent on what bounds on jitter can be offered by the network, which in turn is dependent on the level of provisioning and the techniques deployed to support quality of service. The other issue is that of multiple domains. A single management domain can determine the level of provision and the techniques to be deployed, and thus offer specific bounds on performance. Within the UK, end-to-end performance typically depends on five domains: two sites, two regional access networks, and the backbone. The situation when international use is considered is compounded by the addition of a further level in the hierarchy.

The foregoing discussion suggests that the more stringent requirement on one-way network delay arises in the context of voice over IP, and that VC may perhaps exploit the same QoS, at least in respect of delay

performance; and as compression delays reduce, so VC performance will be enhanced. Although some general guidance has been offered on the problem of determining suitable performance bounds for the network in support the VC application, it is suggested that the operational QoS performance parameters for JANET should only be determined in the light of practical experience.

Figures for acceptable packet loss are very application technology dependent. H.323 video is known to be sensitive to both jitter and packet loss. There are some IP audio applications which use FEC redundancy techniques that allow up to 20% loss with no noticeable effect on end user perception. Unfortunately, no H.323 implementations of which the Think Tank is aware currently use these techniques. The initial recommendation is that packet loss rates should not exceed 5% in order for current H.323 applications to maintain acceptable quality. Evidence (particularly from VIP demo) suggests that acceptable loss rates for H.323 video are in all cases smaller than the acceptable figures for H.323 audio. This arises from the fact that the audio packets are independent, whereas the compressed video packets are not: loss of even one or two may cause loss of synchronisation, which it may take an appreciable time to re-establish.

2.2 *Other video-based services*

In addition to videoconferencing, there is a significant requirement for other video-based services, primarily those commonly described as *video streaming*. Such services fall into two groups: real-time dissemination of 'live' events, and 'streamed' access to audio-video servers. The first is typified by potentially large numbers of users watching transmissions of live events such as news feeds, remote experimental observation, rocket launches, etc. The second is typified by a client application used to gain access to pre-recorded audio and video streams available from a media server. Such material might be used to support teaching activities, for in-service staff training, or to report research results. This type of service needs to support on-demand individual access to material and is likely to incorporate support for user-initiated *pause, rewind and replay*, and *browsing*.

All of these services have the common characteristic that they do not involve interactive inter-personal communication. The user has little care about very tight bounds for latency in either of these scenarios. Of course, the user will expect relatively rapid response to the pressing of a *pause* button, but acceptable responses might typically be of the order of one second rather than 100ms. The more relaxed demands on latency also reflect on the requirements for jitter. Suitable receive-side buffering, together with appropriate algorithms imply that higher levels of jitter, perhaps up to 500ms, are acceptable. While the user requirement for quality implies that a low loss rate is important at the application level, depending on the length of the receive buffer, automatic repeat request (ARQ), for example, may be exploited by the application to enable satisfactory operation in the presence of higher network packet loss.

For support of services involving multiple users watching a live event — real-time *videocasting* — multicasting is likely to be important to enable scaling to larger numbers of users, both in respect of network usage and server load. If services such as *BBC News 24*, or 'live' videocasts of significant scientific events, for example, were to prove to be popular to large groups of users, then it will be important that this type of service is delivered in a manner which does not make unscalable demands upon network resources. (Use of multicast in support of individual access to stored media requires the use of more sophisticated on-demand scheduling by the media servers.)

There is a wide range of requirements for the general 'quality' of these 'other video-based services'. Some services already in common use, such as the use of RealVideo to make BBC news feeds available, offer reasonable quality audio, but the accompanying video feed typically offers quite a low frame rate and small image size. These services, however, operate using perhaps just 30Kbps of network bandwidth. Other services, such as the remote viewing of cinema-quality material, such as might be required for teaching or research, will require much higher quality and thus will consume much higher network capacity, easily of the order of tens of megabits per second, and perhaps hundreds of megabits per second.

2.3 *Voice over IP*

There is an emerging demand for telephony or voice over IP (VoIP). A number of JANET-connected sites are currently or will shortly be re-procuring telephony facilities. Many manufacturers, both in the traditional telephony market and in the IP networking market, now have products that support VoIP. Consideration of the support of VoIP evidently needs to form an integral part of near term plans for JANET services.

Most serious products in this area are essentially voice-only implementations of H.323 in respect of media transport. However, many of the products support a range of Intelligent Network (IN) capabilities, such as

call-forwarding, call-redirection, re-direct on busy, voice mail, caller-id and so on (as, for example, so-called “star services” of BT and other operators).

For conversational voice transport, the general requirements have already been discussed in Section 2.1, Videoconferencing, above. The discussion of one-way delay requirements relates directly to VoIP. Experience through trials of VoIP will need to include as objectives the practical determination of operational bounds on delay and jitter in both single- and multiple-domain contexts, and packet loss bounds. If more than one VoIP technology needs support, trials will be required for each. Demands for IN services are not yet clear: this area requires evaluation and assessment of future needs. More generally, it is perceived that use may grow in two ways: through casual use between personal desktop or portable systems; and through installation of site telephony systems which enable telephony over IP. Emerging use of both need to be monitored, both in respect of performance and to determine the level of provisioning in future.

2.4 Outreach aspects

There is an emerging need for “outreach” services. Many JANET connected institutions expect to deliver services to consumers who are not directly located at their own, or other, JANET-connected sites. Partnerships between the HE and FE sectors and with the school sector are contributing to this requirement. Government agendas concerned with education/training for work and life-long learning are other potential drivers in this field.

Evidently, services such as videoconferencing and video-streaming will increasingly have users located off-campus, in village halls, community centres, SMEs, and private homes. Some of these user locations may indeed be connected by high-speed leased circuits or technologies such as ADSL / SDSL etc, but others may be connected at much lower speeds.

Two aspects arising from the requirement for outreach support need further investigation: these may be categorised as support for heterogeneity, and extending support for QoS beyond JANET. The first of these relates to the operation of a multi-party session where some users enjoy high quality connections (possibly associated with high-performance devices) while others suffer lower quality connections (and perhaps low-performance devices). Investigation into how to support such requirements is needed, and may include transcoding gateways, as well as layered encoding approaches to providing incremental quality. The second area, extending beyond JANET, needs investigation of how to extend interworking with other service networks with which JANET has peering agreements to include QoS. This includes other ISPs, international services, and services such as the new ADSL services.

The user requirement in this outreach area is simply that services work to these outreach locations. Outreach of itself probably introduces no new services but it does introduce a requirement to deliver these services from JANET connected sites to non-JANET connected locations. It is clear that many users see this outreach as a critical part of their future missions and business plans.

2.5 GRID and e-Science

The GRID will be used by a range of research communities each with differing needs and expectations from the network with respect to the provision of QoS. The different research communities will include (but will not be limited to) the following broad areas of science and technology:

- earth observing;
- atmospheric science, including weather forecasting;
- collaborative engineering;
- astronomy and astrophysics;
- particle physics; and
- the biological sciences.

In general, GRID traffic will include bulk data transfer; high(er) priority access to remote datasets; video and audio exchanges including videoconferencing and (medical) imaging; interactive visualisation; and GRID control traffic (see below). This being the case, unless there is bandwidth overprovision or some form of traffic segregation, bulk data transfer will directly compete with other traffic, including traffic having more demanding characteristics of loss and/or delay and delay variation (jitter).

GRID control traffic will be used by resource information services to access metadata catalogues and will also include authentication and authorisation transactions. This is expected to be low volume traffic but may be sensitive to cumulative delay when multiple requests are being made. Without the successful transmission of GRID control traffic, the concept of the GRID will fail.

In the course of processing, a GRID application is likely to access remote objects, such as files; small pieces of scattered data, for example, geometry or calibration information; or some simple information service, to provide answers to questions like: “Where was the satellite at ...?”. Alternatively, an application may access remote processing facilities and expect the required data to be available at that remote location.

For GRID-based research, the following requirements of the network will be central to its success:

- the capability to support sustained large-scale bulk file transfers; and
- the capability to distinguish and accommodate traffic exhibiting different traffic characteristics and profiles, by using different priority mechanisms and/or traffic segregation techniques.

The following description of requirements provides the best information to date for a subset of the science areas identified above.

2.5.1 *Astronomy and astrophysics*

Work in this area will include massive sky surveys and correlating data from different facilities at different wavelengths.

The GRID model under discussion considers a large number of small, dispersed data sources; multiple access points; interpretation at a small number of sites, but use by many distributed clients.

The traffic characteristics are expected to include

- fast access to small remote objects;
- fast access to data information services; and
- multiple database queries by applications.

The AstroGrid is the federation of astrophysics (including solar physics) in the UK and its intent is to provide tools to locate, process and retrieve data. This is expected to work closely with its European counterpart, the Astrophysical Virtual Observatory (AVO) [see <http://www.eso.org/projects/avo/>], and with the National Virtual Observatory (NVO) [see <http://www.srl.caltech.edu/nvo/>] in the US. The requirements of the network are twofold, firstly archiving observations which implies data retrieval from observatories. This is currently running in the order of tens of gigabytes per week which over the next few years is expected to rise to 1–2 terabytes per day with the largest component from the Solar Dynamic Observatory at 1.3 TB/day. However it is recognised that with the advent of Grid technology, existing working practice may change with the data left at the observatories to be replaced by retrieval as required based on excellent cataloguing capabilities. The second requirement is to do with serving data requests through the interactive searching of catalogues. Whilst network capability within the UK is probably adequate, both international bandwidth and QoS capabilities will become important. If the data is located at source, i.e. the observatories, there is a requirement for good, reliable network access with the ability to return tens of megabytes in under a minute at any time.

2.5.2 *Biological Sciences*

Genome databases etc.

The data associated with gene expression work using, for example, 100,000 genes, 320 cell types, 2000 stimuli, 3 time points, 2 concentrations and 2 replicates generates $\approx 8 \times 10^{11}$ data points with each point equivalent to ≈ 10000 bytes. This equates to a total experimental data level of 10^{15} bytes. This data needs to be accessible for the usual data manipulation and for visualisation.

The GRID model under consideration includes a central database to which a new sequence is added at a rate of one every 10 seconds, with the entire database currently in excess of 11 million bases. These will be relatively small but it is expected that in excess of 250,000 queries per day from distributed clients will be made to this data.

2.5.3 Particle Physics

LHC (Large Hadron Collider) Analysis.

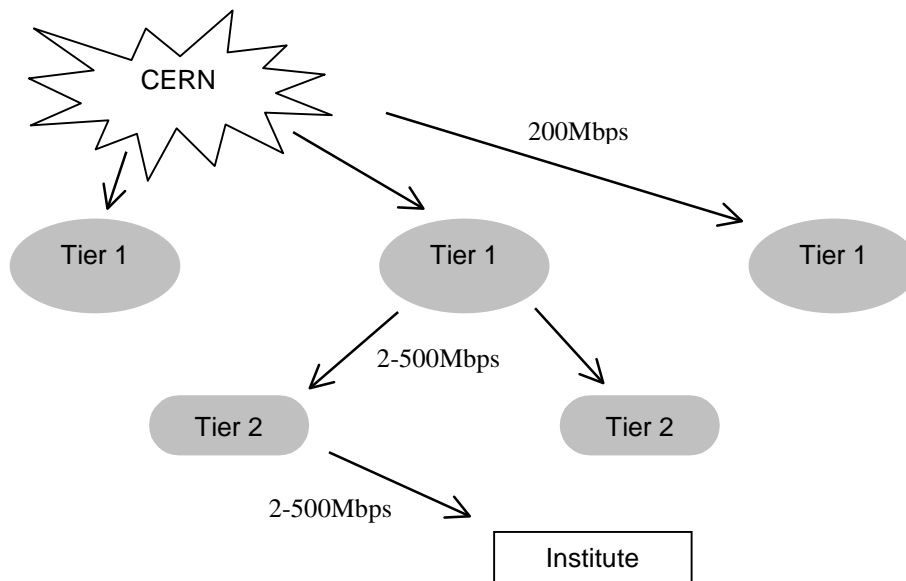
The GRID model under consideration is hierarchical, based on CERN, with branches throughout the world. In total, LHC is expected to generate in the region of 1–7 petabytes of data per year (equivalent to a continuous rate of 300Mbps – 2.1Gbps).

Good network connectivity will be required both among collaborating institutes and regional centre(s), as well as between them and CERN. Within regional centre(s) and UK institutes a transmission technology such as Gigabit Ethernet or better would be fast enough to link analysis facilities to site backbones. The adequacy of SuperJANET4 (at 2.5Gbps now, increasing to 10Gbps in 2002, and to 40–80Gbps subsequently) to offer adequate forwarding bandwidth for connectivity within the UK will depend on the extent to which utilisation by other communities grows within SuperJANET4. In addition, the ability of the existing transport protocols (tcp and udp) and the applications that make use of them to deliver the sustained data transfer rates necessary to meet such requirements are under investigation (see, for example, [5], <http://www.web100.org>, and http://www.acm.org/sigcomm/sigcomm98/tp/abs_25.html).

The LHC GRID traffic requirement is based on the concept of several storage tiers. In the UK there will be a single Tier 1 site, a small number of Tier 2 sites, and a far greater number of collaborating Institutes — shown diagrammatically below. Transmission capacity requirements have been calculated as equivalent continuous rates.

Whilst this model shows the logical flow of data with respect to the experiment, the traffic within the network core will be very different and is calculated as 0.5-1.0Gbps with substantially higher peak rates.

The traffic characteristics: will include non-time critical bulk data replication; a small number of remote queries; fast access to small remote objects; and fast access to resource and data information services.



Babar Requirements

The UK requirements of the SLAC Babar experiment include the transfer of real data from SLAC to the UK and of Monte Carlo (MC) production data from the UK back to SLAC. Babar has a model for its network needs similar to the hierarchical model of LHC. In the UK, the ‘Tier A’ site is expected to be located at RAL whilst the MC production is expected to be located at 8 university ‘compute farms’. The data requirements for the Tier A centre is expected to be as follows:

	2000	2001	2002	2003	2004	2005
ftp (TB/yr)	6	74	163	290	439	688
rate (Mbps)	2	20	24	34	40	66

[with the proviso that 2001 data transfer has yet to start].

For the MC production, by the end of summer 2001 it is expected that 40M events per month will be calculated which equates to 20 Tbytes per month or 32 Mbps (including compression). In the future this production will increase such that each of the compute farms will generate 20 Mbps, i.e., 140 Mbps all destined for SLAC. In summary,

	2001	2002	2003	2004	2005
(TB/yr) compressed	60	192	84	552	648
(Mbps)	32	51	102	147	172

Currently, both the transfer of real data and of MC production data has been continual, but this will change. The UK can afford to fall one or two weeks behind with the real data transfers and the disk storage capacity to buffer no more than a few days of MC production data. In addition there is the likelihood of transferring Babar data between European centres.

2.5.4 Access Grid

Within the UK e-Science programme, recent attention has focused upon the videoconferencing paradigm promulgated by the Access Grid project [see <http://www-fp.mcs.anl.gov/fl/accessgrid/>], in which informal (as well as formal) discussion for project collaboration, teaching, or other similar purpose is supported by linked ‘spaces’ — coffee lounge to lecture theatre — which are available at all times (though some sessions may be booked for specific purposes). In general, video streams from all participating sites are projected on a wall at each site, as are screen projections of shared applications.

While the benefits for working practice remain to be explored, it is anticipated that UK trial of this technology will form a part of the initial e-Science programme. Technically, video and audio streams require support, as well as suitable data service to support interactive shared applications. However, in contradistinction to H.323-based services using MCU technology, Access Grid exploits network multicast. A feature of current prototype use is that sites (at least potentially) receive each other's video streams at all times at sufficient quality to support open discussion. The scaling implications of this will require investigation, especially as each Access Grid node is expected to be ‘always on’ and to insert an estimated 4Mbps into the network. It should also be noted that since the activity is an international one, UK participation implies a requirement not only for UK-wide multicast but international multicast.

2.6 Managed Bandwidth Service

During the 1990s there was considerable deployment within Europe, nationally and internationally, of both prototype and service networks based on ATM technology. This infrastructure was used to support a number of national and international research activities. As the technology matured towards the end of the decade, national and international research and education networks offered a service which was IP-based but delivered over an ATM bearer service. The ATM bearer retained the capability of providing an end-to-end bearer service. Although most of the underlying transmission capacity was defined for use in support of the

IP service, a proportion—typically around 10%—was made available for use by end-user consortia on a relatively informal basis both nationally and internationally.

Typical usage was in the form of a *virtual path* or *channel connection* (VPC or VCC), which was configured to limit peak rate, and in effect provided a form of temporary point-to-point circuit. A consortium which arranged for a set of these could manually construct a primitive form of *virtual private network* (VPN).

The uses to which such a service could be put may be categorised as follows:

1. dedicated link bandwidth, perhaps as a means of avoiding congestion points within the service network;
2. network experiments at the IP layer or above, which would place the service network at risk;
3. use by application-level experiments which demand a quality of service not available within the service network.

Towards the end of the 1990s, a *Managed Bandwidth Service* (MBS), based on the underlying ATM infrastructure, was made available, nationally and internationally, to meet such requirements. Demand continues for a service to meet all three of the above categories.

The current technology underlying the SJ4 core network cannot offer a native MBS and there is not currently a solution to providing an equivalent service in the immediate future. There is the possibility that traffic engineering techniques, based for example, on MPLS, will be available to provide parts of this service but technical developments will be required for this to happen.

Since the service would be introduced in the service network at the IP layer, the service cannot in principle be equivalent to the previous MBS in respect of the second category above. Such a service is also predicated on the assumption that provision of MBS through congestion points is determined by policy: either extra provisioning is required to increase capacity or MBS capacity is taken out of (already congested) existing capacity.

References for section 2

- [1] V. Reijs, *Perceived quantitative quality of applications*, http://www.heanet.ie/Heanet/projects/nat_infrastruct/perceived.html, 2001.
- [2] Srinivas Vegesna, *IP Quality of Service*, Cisco Press, 2001 (ISBN 1-57870-116-3).
- [3] V. Kumar, M. Korpi, and S. Sengodan, *IP Telephony with H.323*, Wiley, 2001 (ISBN 0-47139-343-6).
- [4] P. Schopis, *The Reality and Mythology of QoS and H.323*, <http://www.mega-net.net/megaconference/presentations/PaulSchopis.ppt>, Megaconference 2001.
- [5] W. Feng and Tinnakornrisuphap, *The failure of TCP in High-performance computational Grids*, US DoE contract W-7405-ENG-36, Los Alamos National Laboratory, 2000.

3 Technology review

This section provides a review of a number of the major QoS technologies currently available at the IP layer. Section 3.1 provides an appraisal of the Integrated Services model (IntServ) and the associated Resource Reservation Protocol (RSVP), and the Differentiated Services (DiffServ) architecture. Section 3.2 discusses the issues surrounding provisioning for traffic engineering within a multiservice network, including Multiprotocol Label Switching (MPLS) and Bandwidth Brokering. Support for QoS associated with specific layer-2 technologies is not considered here.

3.1 Elements of QoS technology

3.1.1 Integrated Services (IntServ)

Overview

In the Integrated Services (IntServ) architecture [1], three classes of services are proposed, based on applications' delay requirements. These are: the guaranteed-service class which provides for delay-bounded service agreements; the controlled-load service class which provides for a form of statistical delay service agreements (nominal mean delay) that will not be violated more often than in an unloaded network; and the well-known best-effort service which is further partitioned into three categories: interactive burst (e.g. WWW), interactive bulk (e.g. FTP) and asynchronous (e.g. e-mail).

The main point is that the guaranteed-service and the controlled-load classes are based on quantitative service requirements and both require signalling and admission control in network nodes. These services can either be provided per flow or per flow aggregate, depending on flow concentration at different points in the network. Although the IntServ architecture need not be tied to any particular signalling protocol, RSVP (described in section 3.1.2) is often regarded as the signalling protocol in IntServ. The best-effort service, on the other hand, being the default service, requires no signalling.

Discussion

The major advantage of IntServ is that it provides service classes that closely match the different application types for the wide variety of applications that need to be supported within a multiservice networking environment. For example, the guaranteed-service class is particularly well suited to the support of time-critical, intolerant applications. On the other hand, time-critical, tolerant applications and some adaptive applications can generally be efficiently supported by controlled-load services. Other adaptive and elastic applications are accommodated in the best-effort service class. IntServ leaves the best-effort service class unchanged, and also leaves the forwarding mechanism in the network unchanged. This allows for an incremental deployment of the architecture, while allowing end-systems that have not been upgraded to support IntServ to be able to receive data from any IntServ class (with, of course, a possible loss of guarantee).

End-to-end service guarantees cannot be supported unless all nodes along the route support IntServ. This is obviously so because any "pure" best-effort node along any route can treat packets in such a way that the end-to-end service agreements are violated.

3.1.2 Resource Reservation Protocol (RSVP)

Overview

RSVP is based on the concept of the session [2]. A session is composed of at least one data flow and is defined in relation to a *destination* (more precisely, as the triplet [destination address, destination port, protocol id]). As the destination address can be a multicast address, the destination can thus be either a group of receivers or a single receiver. A flow is defined as any sub-set of the packets in a session, or in other words, as a sub-set of the packets sent to a given destination. A flow is therefore simplex. Theoretically, the sub-set of packets making up a flow may be arbitrary, but in the current state of the RSVP specification, a flow is defined as the set of packets emitted from a given *source* (identified by the pair [source address, source port]).

RSVP works as follows [2, 3]:

Path messages are periodically sent towards the destination and establish a “path state” (including reverse signal path) per flow in the routers. *Resv* messages are periodically sent towards the sources, and they establish the required reservations along the path followed by the data packets. The style of reservation in RSVP is thus “receiver oriented”, since it is the receivers that initiate the requests for resources to be reserved. In order to reduce the overhead associated with RSVP, any *Path* or *Resv* message that does change the states held by a router is not forwarded immediately by that router. Instead, each router periodically issues its own *Path* and *Resv* messages carrying information about the flows it holds. A lifetime *L* is associated with each reserved resource. This timer is reset each time a *Resv* message confirms the use of the resource. If the timer expires, the resource is freed. This principle of resource management based on timers is called *soft state*. Soft state is also applied to the path state in the routers (in this case, the timer is reset upon reception of a *Path* message).

Teardown messages (*PathTear* and *ResvTear*) are available for immediate release of the corresponding states (path state and reservations). Teardown requests can be initiated by a sender, or by a receiver, or by any intermediate RSVP router (upon state timeout or service pre-emption). It is worth noting that all the messages described above are delivered unreliably: because of the protocol reliance on soft-states, the concept of acknowledgement is not used in RSVP.

RSVP allows several styles of reservation: distinct resources may be assigned to given flows while several flows may share some resources. The selected style for a given flow is expressed in the *filter spec* associated with that flow. There exist three types of filter:

- Wild-card filters: every flow of a session shares the associated resource.
- Shared-explicit filters: explicitly identified flows share the same resource.
- Fixed filters: ensure that one flow is granted the exclusive use of a resource.

Benefits

RSVP has been designed to be able to operate across non-RSVP networks. It is extremely difficult (if not impossible) to guarantee end-to-end services in such a case. Nevertheless, this allows for a progressive deployment of the protocol associated with a steady improvement of the end-to-end best-effort service seen by flows exploiting RSVP in the parts of the Internet where it is supported. The soft-state mechanism is a very simple self-stabilising mechanism to keep the nodes of the network in a consistent state. It provides “natural” recovery from node crashes, as well as preventing “resource leaks” by reclaiming resources made obsolete by various events external to the reservation scheme (such as, route changes, loss of teardown messages, users leaving a multicast group without explicitly releasing resources).

RSVP as a receiver-driven protocol scales to large numbers of participants in multicast groups. The reservation request from a receiver does not have to propagate all the way to the sender in most situations. If the reservation request encounters an existing reservation in one of the RSVP routers along the route which is equal to or greater than its own reservation request, then it merges with the existing reservation at that router and does not travel any further. This also allows for “incremental” reservations whereby some receivers behind a bottleneck can hold partial reservations and then regularly poll the network hoping for completion of full reservations.

The different reservation styles proposed in RSVP tend to improve resource usage efficiency in the nodes of the network. For instance, shared reservations are well suited to scenarios where multiple sources are unlikely to transmit simultaneously (for example, audio sources in conferencing applications), because, in such a case, the size of the shared reservation is essentially independent of the number of sources. It should be noted that shared reservations provide for an “overall” gain in reservation efficiency. Unless all the sources transmit with the same “requirements”, and the shared reservation is exactly equal to the requirements of a single source, resource waste may occur near the sources while a resource gain occurs everywhere else in the network (compared to having simultaneous individual reservations for each source).

Shortcomings

The operation of RSVP is based on periodic messages exchanged unreliably. This can result in possibly long establishment latencies, if RSVP messages are lost during the establishment phase of a reservation. This is because the average time to recover from such losses is equal to the message refresh period whose value is several tens of seconds.

RSVP was designed to scale in terms of the size of the groups of receivers it can support on individual flows. However, the reservation model in RSVP can itself represent a threat for scalability of the protocol, especially in parts of the network where flow concentration is high (such as in the core of the Internet). It is so because this reservation model induces both state overhead and message overhead that are, at best, linear in terms of the number of sessions established. The state overhead resides in both the slow path of routers (reservation states in the RSVP daemon) and, more importantly, the fast (data) path of routers (filtering/classification, scheduling, and possibly policing states). Flow aggregation techniques are currently being proposed to try to solve this state scalability problem.

There is also considerable overhead in terms of the message overhead caused by the periodic refresh of soft-states. The message overhead consumes bandwidth and, in most cases (even with aggregation), processing resources in the routers. Furthermore, although soft-state is a very simple mechanism, it has proved slow to react to some network conditions (e.g., node failures, route changes).

The receiver-based approach, as used in RSVP (where reservation request and specification travel upstream toward the source), may not be well suited to all types of application. Indeed, some applications fit a model where it is the sender(s) that fix(es) the quality of transmission (for example, Internet telephony, and digital TV/VoD services). The receiver-based scheme of RSVP also leads to fairly static reservations, which in some cases can be wasteful. For instance, consider the case where sources of a multicast session use different coding schemes (giving different communication requirements) and are unlikely to be simultaneously active. A receiver that wants to receive the different data flows without any loss of quality will have to request a reservation matching the characteristics of the most demanding coding scheme, because it never knows when each sender is going to be active. In contrast, if the sender could propagate reservation messages downstream, the reservations could be up-dated whenever the sources are active.

3.1.3 Differentiated Services (*DiffServ*)

Overview

By recognising that most of the data flows generated by different applications can ultimately be classified into a few general categories (i.e., *traffic classes*), the differentiated services (*DiffServ*) architecture [4] aims at providing simple, scalable, service differentiation. It does this by discriminating among the data flows and treating each according to its traffic class, thus providing a logical separation of the traffic in the different classes.

In *DiffServ*, scalability and flexibility are achieved by following a hierarchical model for network resource management:

- A. Inter-domain resource management: unidirectional service levels, and hence traffic contracts, are agreed at each boundary point between a customer and a provider, for the traffic entering the provider network.
- B. Intra-domain resource management: the service provider is solely responsible for the configuration and provisioning of resources within its domain (i.e., the network). Furthermore, service policies are also left to the provider.

At their boundaries, service providers build their offered services with a combination of traffic classes (to provide controlled unfairness), traffic conditioning (a function that modifies traffic characteristics to make it conform to a traffic profile and thus ensure traffic contracts are respected) and accounting (to control and balance service demand, through billing or administrative policy). Provisioning and partitioning of both boundary and interior resources is the responsibility of the service provider and, as such, is outside the scope of *DiffServ*. For example, *DiffServ* does not impose either the number of traffic classes, or their characteristics, on a service provider.

Although traffic classes are nominally supported by interior routers, *DiffServ* does not impose any requirement onto interior resources and functionalities. For example, traffic conditioning (i.e. metering, marking, shaping or dropping) in the interior of a network is left to the discretion of the service providers.

If each packet conveyed across a service provider's network simply carries in its header an identification of the traffic class (called a DS code point) to which it belongs, the network can provide a different level of service to each class. It does this by associating an appropriate Per-Hop Behaviour (PHB) with each traffic class (DS code point) and treating each packet accordingly, by dropping it or imposing suitable scheduling behaviour (delaying, expediting, etc).

Two important PHBs for forwarding defined in DiffServ are Expedited and Assured:

- **Expedited Forwarding (EF)** specifies a minimum departure rate for packets. Provided this is not exceeded, then associated queues in routers will be short or empty. This forwarding behaviour is intended for support of services offering bounded loss, delay, and jitter.
- **Assured Forwarding (AF)** specifies a group of related behaviours. Each group is allocated forwarding resources. Packets within a group are marked with a drop precedence: packets with lower precedence are dropped (probabilistically) in favour of those with higher precedence. Three levels of precedence are defined.

It must be noted that DiffServ is based on local service agreements at customer/provider boundaries. Therefore, end-to-end services will be built by concatenating such local agreements at each domain boundary along the route to the final destination. It may be noted that monitoring and policing is in principle necessary to observe and enforce adherence to each of these unidirectional agreements on the receiving side of each interdomain network boundary. [Indeed this approach is proposed for Géant in GN1 (Géant) Deliverable D9.1-Ad1, 'Implementation architecture specification for a Premium IP Service'.]

Benefits

The DiffServ architecture is an elegant way, within the constraints of the network provision, to provide much needed service discrimination within a commercial network in which use of a differentially better class of service results in a higher bill. Customers willing to pay more will, in a loaded network, see their applications receive a better service than those paying less will. This scheme exhibits an "auto-funding" property: "popular" traffic classes generate more revenues that can be used to increase their provisioning. Even in a non-commercial network, the accounted usage for each traffic class may still be capable of influencing the provisioning policy for each class.

A traffic class is a pre-defined aggregate of traffic. Traffic classes in DiffServ are accessible without signalling, which means that they are readily available to applications without any set-up delay. Consequently, traffic classes can provide *qualitative* or *relative* services to applications that cannot express their requirements quantitatively. This conforms to the original design philosophy of the Internet. An example of qualitative service is "traffic offered at service level A will be delivered with low latency", while a relative service could be "traffic offered at service level A will be delivered with higher probability than traffic offered at service level B". *Quantitative* services can also be provided by DiffServ. A quantitative service might be "90% of in-profile traffic offered at service level C will be delivered". As the provisioning of traffic classes is left to the provider's discretion, this provisioning can, and in the near future will, be performed statically and manually. Hence, existing management tools and protocols can be used to that end. However, this does not rule out the possibility of more automatic procedures for provisioning in the future.

The only functionality actually imposed by DiffServ in interior routers is packet classification. This classification is simplified compared to the one in RSVP because it is based on a single IP header field containing the DS codepoint, rather than multiple fields from different headers. This has the potential of allowing certain functions performed on every packet, such as traffic policing or shaping, to be done at the boundaries of domains, so that forwarding is the main operation performed within the interior of the provider network.

Another advantage of DiffServ is that the classification of the traffic, and the subsequent selection of a DS codepoint for the packets, need not be performed in the end-systems. Indeed, any router in the stub network where the host resides, or the ingress router at the boundary between the stub network and the provider network, can be configured to classify (on a per-flow basis), mark and shape the traffic from the hosts. Such routers are the only points where per-flow classification may occur, which does not pose any problem because they are at the edge of the Internet, where flow concentration is low. The potential non-involvement of end-systems and the use of existing and widespread management tools and protocols allows for a swift and incremental deployment of the DiffServ architecture.

Shortcomings

Simultaneously providing several services with differing qualities within the same network is a very difficult task. Despite its apparent simplicity, DiffServ does not make this task any simpler. Instead, in DiffServ, it was decided to keep the operating mode of the network simple by pushing as much complexity as possible onto network provisioning and configuration. Of course, network provisioning and configuration have been performed since the creation of the very first communication networks, and thus they benefit from long experience and from available tools and traffic models. However, so far, large networks have mainly offered a single type of service (e.g. best effort service in the Internet, interactive voice in telephone networks, etc.).

The provisioning of networks providing multiple classes of service at the same time is therefore a rather new area which requires much more research to study the added complexity due to possibly adverse interactions between different classes of service. The construction of end-to-end services by concatenating local service agreements is also a non-trivial research issue.

The key to provisioning is the knowledge of traffic patterns and volumes traversing *each* node of the network. This also requires a good knowledge of network topology and routing. The problem with the Internet is that provisioning will be performed at a much slower time scale than the time scales at which traffic dynamics and network dynamics (e.g. route changes) occur. This problem can be illustrated with the simplest case of a single service provider network whose service agreements with customers are static. Although the amount of traffic entering the domain is known and policed, it is impossible to guarantee that overloading of resources will be avoided. This can happen in two ways.

- The entering packets can be bound for any destination in the Internet and may thus be routed towards any border router of the domain (except the one where it entered). In the worst case, a *substantial proportion* of the entering packets might all exit the domain through the same border router. (This is exactly analogous to output port contention in a switch or router.)
- Route changes can suddenly shift vast amounts of traffic from one router to another.

Unless resources are massively over-provisioned in both interior and border routers, traffic and network dynamics can cause momentary violation of service agreements, especially those relating to quantitative services. On the other hand, massive over-provisioning results in a very poor statistical multiplexing gain and is therefore inefficient and expensive. To increase resource utilisation in its network, a service provider can trade generality and robustness for efficiency. For example, to limit the amount of expensive resource dedicated to the support of quantitative services, service providers can limit quantitative service contracts to apply between any pair of border routers in the domain. In such a case, the service would apply only to packets entering the domain at a designated ingress router and leaving the domain at a designated egress router. This helps solve the first problem described above at the cost of generality, since only packets bound for destinations “served” through the egress router can benefit from the service. Of course, to ensure that the egress router is in the route to any given destination, the inter-domain routing entry for that destination must be statically fixed in the ingress router. Even for a fixed ingress-egress pair, intra-domain routing dynamics can still occur. This means that the set of internal routers visited by the packets travelling between the ingress and the egress routers can still suddenly change. However, the “directionality” of the traffic considered here is such that the number of possible routes is considerably reduced compared with the general case, and so is the resulting and necessary over-provisioning. A service provider could, however, reduce to a minimum the over-provisioning of quantitative services offered between pairs of border routers by “pinning” the intra-domain route between those routers. Fixing the egress router for a given destination and/or pinning internal routes between border routers nevertheless incurs a loss of robustness.

Alternatively, a service provider might wish to use dynamic logical provisioning and configuration (i.e. sharing of resources between classes) as an answer to the problems of network and traffic dynamics. However, depending on the type of service agreement (qualitative, relative or quantitative) and the QoS parameters involved in the agreement, dynamic logical provisioning might require signalling and admission control.

From the point of view of a flow, the class bandwidth is not a meaningful parameter. Indeed, bandwidth is a class property that is *shared* by all the flows in the class, and the bandwidth received by an individual flow depends on the number of competing flows in the class as well as on the fairness of their respective responses to traffic conditions in the class. Therefore, to receive some quantitative bandwidth guarantees, a flow must “reserve” its share of bandwidth along the data path, which involves some form of end-to-end signalling and admission control (among, at least, logical entities called *bandwidth brokers*). This end-to-end signalling should also track network dynamics (i.e., route changes) to enforce the guarantees, which can prove very complex.

On the other hand, delay and error rates are class properties that *apply* to every flow of a class. This is because, in every router visited, all the packets sent in a given class share the queue devoted to that class. Consequently, as long as each router manages its queues to maintain a relative relationship between the delay and/or error rate of different classes, relative service agreements can be guaranteed without any signalling. However, if quantitative delay or error rate bounds are required, then end-to-end signalling and admission control are also required.

End-to-end signalling and admission control would increase the complexity of the DiffServ architecture. The idea of dynamically negotiable service agreements has also been suggested as a way of improving resource

usage in the network. Such dynamic service level agreements would require complex signalling, since the changes might affect the agreements that a provider has with several neighbouring networks. The time scale at which such dynamic provisioning could occur would be limited by scalability considerations, which in turn may impede its usefulness.

3.1.4 *Less than best effort*

The availability of enhanced QoS mechanisms throughout a network or path of networks implies the availability of the desired QoS mechanisms (e.g., particular DiffServ implementations) in the routers along that path.

The Internet 2 Scavenger project [<http://qbone.internet2.edu/qbss/>] takes a different approach to enabling a better QoS for high priority applications. Rather than classifying high priority traffic, sites are voluntarily able to mark low priority traffic. The aim of the scheme is to give a better quality of service for the more important data by channelling lower priority, non-critical traffic into a “less than best effort” (LBE) class. This is analogous to the ‘renice’ facility on a shared Unix system, where background tasks can voluntarily be given lower scheduling priority.

Any traffic marked as LBE is discarded first when congestion occurs at a router, or the router may offer a very small minimum departure rate, or give the traffic a low probability of timely forwarding.

The classification of LBE traffic can be done at a site’s egress router, or if the application supports it, by the user directly. The facility is being promoted within the Internet 2 community as an issue of “netiquette”. One of the interesting aspects of Scavenger is the possibility to enable new classes of application, where LBE data is used heavily, but in the knowledge that it is always discarded first when congestion occurs.

The implementation on Internet 2 is currently done using the DiffServ codepoint of “001000” to indicate LBE traffic. For the LBE property to propagate on the network the codepoint should be immutable, i.e., no other router on the path should clear or alter the LBE marking. A router may give the traffic regular best effort treatment if it is not LBE-aware. The scheme can be applied only at congestion points on a network, provided the immutability property is honoured.

Scavenger is in an early deployment state on Internet 2, but is worthy of tracking. If SJ4 adopts the idea, it should use similar implementation mechanisms to Internet 2 (and GÉANT, if that adopts the scheme) for interoperability and common understanding of LBE marking.

3.1.5 *QoS Support in the End-System*

For some time now, we have been used to applications that are able to provide certain levels of QoS capability in order to support “time sensitive” applications. For example, there are applications that support video codecs that are able to adapt to the current loading on the network, and make the most of the current effective bandwidth availability. There are applications that use buffering techniques to allow the storage of a certain amount of audio / video data, in order to maintain a smooth playback, despite fluctuations in delay.

As discussed in previous sections, IntServ / RSVP, and DiffServ have different relationships with the end-systems aiming to take advantage of the QoS support available within the network. One of the features of DiffServ is that the end-system does not under normal circumstances take part in any of the marking of packets. This implies that it may be easier to introduce DiffServ into an existing network. However, it is possible that there will be some interaction between the end-system and the point in the network at which the packet marking is carried out, in order to allow an application’s requirements to be reflected in the network resource provisioning. This could be of a static nature where the border router marking the packet has rules based on the application sending the traffic (typically deemed from the port number), or of a more dynamic nature, where the end-system has the capability to signal its requirements for the application to the border router (implying that the end-system application is “QoS aware”). The IntServ / RSVP model places considerable emphasis on the end-system carrying out the “signalling” of its requirements to the network (based around the idea of receiver-oriented resource reservation), and therefore inherently relies on the application being aware of its QoS requirements.

The advent of the Internet-based QoS mechanisms discussed earlier has led to the development of software packages to support them, for use with popular operating systems. For DiffServ (despite end-systems not being involved with packet marking), marking capabilities have been developed for a number of operating systems, examples of which include both Linux and Microsoft Windows 2000. The specification of the RAPI (an RSVP Application Programming Interface) [5] and SCRAPI (a simplified form of an RSVP API) [6]

interfaces within the IETF have provided a base standard API for the use of RSVP from an end-system point of view.

To use Microsoft Windows 2000 as an example, through the Winsock2 API and a number of built-in QoS modules, a range of QoS support facilities are accessible to the application programmer, at the IP layer and layer 2: although layer 2 QoS support is not considered in this report, nevertheless the appearance of these facilities serves to indicate the emergence of support for QoS in end systems. Aggregates of traffic are supported in terms of DiffServ packet marking and at layer 2 by the provision of expedited traffic capabilities as defined in 802.1D [13]. Per-flow end-to-end signalling is supported in terms of IntServ/RSVP. IntServ service is supported by a number of traffic control modules such as a packet classifier and packet scheduler. From a management perspective, support is available for COPS (Common Open Policy Service) [14], SNMP and a Subnet Bandwidth Manager (SBM). (SBM provides RSVP-style admission control for use with 802-style networks [15].) Applications such as NetMeeting automatically provide support for RSVP, in order to allow users that are connected to networks that support RSVP to take advantage of the available resource provisioning capabilities, now.

3.2 Provisioning and traffic engineering

The idea of traffic engineering is to allow traffic flows to be organised such that the network can be optimised in one way or another. This optimisation may be in the form of effectively utilising network resources, or in terms of minimising congestion in certain parts of the network. The multiservice nature of future IP-based networks implies that service differentiation is also a key requirement when considering the issue of network and traffic engineering.

The remainder of this section covers two key technologies in the area of traffic engineering. Multiprotocol Label Switching (MPLS) defines an approach to seamlessly integrating layer 3 routing with layer 2 switching techniques. Bandwidth Brokering is a technique that, in conjunction with the concepts defined within the DiffServ architecture, enables the concatenation of agreements across multiple domains.

3.2.1 Multiprotocol Label Switching (MPLS)

Overview

The IETF MPLS working group is addressing the issues of scalability of routing, provision of more flexible routing services, increased performance and more simplified integration of layer 3 routing and circuit switching technologies, with the overall goal of providing a standard label switching architecture [7].

Each MPLS packet has MPLS-specific header information that is either encapsulated between the link layer and the network layer, or resides within an existing lower-level header. At most, the MPLS header will contain a label, TTL field, Class of Service (CoS) field, stack indicator, next header type indicator and checksum.

MPLS defines a fundamental separation between the grouping of packets that are to be forwarded in the same manner (the Forwarding Equivalence Classes, or FECs), and the labels used to mark the packets. This is purely to enhance the flexibility of the approach. At any one node, all packets within the same FEC could be mapped onto the same locally significant label (given that they have the same requirements). However, there are instances where one may wish to engineer the network in such a way that several different labels are used (for example, when wishing to differentiate explicitly between streams). The assignment of a particular packet to an FEC is done once, at the entry point to the network. MPLS-capable routers (termed Label Switched Routers or LSRs) then use the label and CoS field alone to make packet forwarding and classification decisions. Label merging is possible in instances where multiple incoming labels are assigned the same FEC.

MPLS packets are able to carry a number of labels, organised in a last-in first-out stack. This can be useful in a number of instances, for example where two levels of routing are taking place across transit routing domains. Regardless of the existence of the hierarchy, in all instances, the forwarding of a packet is based on the label at the top of the stack. In order for a packet to traverse an MPLS network by a particular path, the entry node at the transmitting end of the path pushes a label relating to the path to be taken by the packet onto the stack, and sends the packet to the next hop in the path.

A collection of LSRs forming the series of hops along a path in an MPLS network constitutes a Label Switched Path (LSP). Two options are defined for the selection of a route for a particular forwarding class. Hop-by-hop routing defines a process where each node independently decides the next hop of the route. Explicit routing is where a single node (often the ingress node of a path) specifies the route to be taken (in

terms of several or all of the LSRs in the path). Explicit routing may be used to implement network policies, or to allow traffic engineering in order to balance the traffic load.

There are three main approaches for identifying traffic to be switched. First, path creation can be control or topology driven, where labels are pre-assigned in relation to normal routing control traffic. Here, the network size dictates the load and bandwidth consumed by the assignment and distribution of label information. Second, request-based control traffic from protocols such as RSVP can trigger path creation relating to individual flows or traffic trunks. Here, the number of labels and computational overhead will depend entirely on the number of flows being supported. Finally, data traffic driven label assignment is where the arrival of data that is recognised as a “flow” activates label assignment and distribution “on the fly”. This approach implies that there will be latency while the path set-up takes place. Overheads in this case will be directly proportional to traffic patterns.

MPLS is able to work in an environment that uses any data link technology, both connection-oriented and connectionless. MPLS also provides the potential for all traffic to be switched, but this depends on the granularity of label assignment, which again is flexible and depends on the approach used to identify traffic (discussed above). Labels may be assigned per address prefix, for example a destination network address prefix, or set of prefixes, and can also represent explicit routes. On a finer grained level, labels can be defined per host route, and also per user. At the lowest level, a label can represent a combined source and destination pair, and in the context of RSVP, can also represent packets matching a particular filter specification.

MPLS needs a mechanism for distributing labels in order to set up paths. The architecture does not assume that there will be a single protocol (known as a Label Distribution Protocol or LDP) to complete this task, but rather there will be a number of approaches that can be selected depending on the required characteristics of the LSPs. In the instance where paths relate to certain routes, label distribution could be piggybacked onto routing protocols [8]. Where labels are allocated to the packets of a specific flow, distribution can be included as part of the reservation protocol. New protocols have been developed for general label distribution [9] and the support of explicitly routed paths [10]. MPLS label distribution requires reliability and the sequencing of messages that relate to a single FEC. While some approaches use protocols that sit directly over IP (thus implying that they are unlikely to be able to meet these reliability requirements), a number of the defined LDPs solve this issue by operating over TCP.

Within the MPLS architecture, label distribution binding decisions are generally made by the downstream node, which then distributes the bindings in the upstream direction. This implies that the receiving node allocates the label. However, there are also instances (especially when considering multicast communications) where upstream allocation may also be useful. In terms of the approach to state maintenance used within MPLS, a soft-state mechanism is employed, implying that labels will require refreshing in order to avoid timeouts. Approaches to this include the MPLS peer keep-alive mechanism, and the time-out mechanisms inherent within routing and reservation protocols (in instances where they are used to carry out label distribution).

In terms of support for QoS, MPLS provides the Class of Service (CoS) field that enables different service classes to be offered for individual labels. As noted earlier, MPLS is able to provide QoS support on a per-flow basis either using flow detection or request-based control traffic from protocols such as RSVP to trigger label assignment. For the support of “higher-level” traffic engineering approaches, the CoS field could be ignored, using a separate label for each class. In this instance, the label would represent both the forwarding class and the service class.

More recently, there has been a great deal of interest in the area of Generalized Multiprotocol Label Switching (GMPLS), also often referred to as Multiprotocol Lambda Switching. The idea behind GMPLS is that there is support not only for packet switching, but also for devices that perform switching in the time, wavelength and space domains. Here, the intention is provide the bridges required between the IP and photonic layers, in order to enable the continued growth / scaling of the IP world. Work currently underway in this area includes the development of a protocol known as the Link Management Protocol, used to control channel management, link connectivity verification, link property correlation and fault isolation. Also, there is work on the modification / extension of existing MPLS protocols such as CR-LDP and RSVP-TE in order to support GMPLS. For an overview of this work, see Banerjee, 2001 [10].

Benefits

MPLS allows efficient packet forwarding to assist high-speed data transfer. Although the link-layer to be used is not specified, the approaches all provide the scenario where it is possible to fully integrate and couple traditional datagram routing concepts with link-layer switching devices supported within the telecommunications industry. MPLS functionality is now being supported directly within hardware, with

routing and switching mechanisms being combined at the chip level in order to provide integration at high speeds, thus increasing its viability.

MPLS-capable devices are able to provide additional functionality beyond the best-effort packet forwarding found within a router. This flexibility means that in principle it is possible to support ideas such as Quality of Service differentiation. The fundamental separation between forwarding class and label assignment provides a great deal of flexibility in terms of the way in which traffic can be engineered.

Alone, IP does not lend itself to the idea of traffic engineering (i.e., the ability to manage bandwidth and routes in order to provide equal loading of resources within the network). Until now, it has been reliant on other technologies (such as ATM) and associated encapsulation techniques in order to offer this functionality. MPLS provides support for traffic engineering through the deployment of constraint-based routing. Stemming from the idea of QoS routing, constraint-based routing not only provides routes that are able to meet the QoS requirements of a flow, but also considers other constraints including network policy and usage. Label distribution protocols supporting label switching for end-to-end constraint-based paths [11] allow traffic characteristics to be described in terms of peak rate and committed rate bandwidth constraints, along with a specified service granularity (which can be used to define the delay variation constraint).

Explicit routing (a subset of constraint-based routing) allows the specification of the route to be taken across the network. This is enabled within MPLS by allowing a label to represent a route, without the overhead of source routing found within normal IP forwarding (that makes it too resource intensive for use in most circumstances). Different paths can be selected in order to allow traffic engineering to be carried out effectively, allowing network load to be balanced in a far more flexible manner than manually configuring virtual circuits (as with other primitive approaches to the engineering IP traffic). The engineering of paths in such a way implies a simple mechanism for measuring traffic between edge network devices making use of a label switched path.

Shortcomings

MPLS essentially attempts to overlay connection-oriented concepts onto connectionless technologies. While providing several advantages, in a number of instances this approach reduces the overall flexibility of the IP protocol, and could be branded as being somewhat heavyweight. Some of the conclusions that led to the research into MPLS, such as the fact that routers are too slow, or that routing tables are becoming too large, have been weakened by the appearance of fast and powerful gigabit routers.

The MPLS architecture [7] defines a base-level label swapping technology. As shown within the previous sections, MPLS allows for traffic to be switched under different circumstances (topology driven, flow driven etc...), using different label distribution protocols depending on the circumstances. While this implies that MPLS is flexible, it is likely to be applicable only within well-managed networks, where all components are able to provide support for MPLS and the individual distribution protocols in use. With topology driven label assignment (where labels are allocated and distributed without reference to the traffic), a full mesh of labels will be established. The overhead of this approach is essentially relative to the size of the network, and has the potential to use a vast number of labels. This can be a large overhead in instances when labels are allocated to routes where very little traffic is flowing.

In terms of the dynamic provision of varying levels of QoS, MPLS poses a number of issues.

- Label assignment based on support for traffic flows will require a path to be put in place at the moment the flow is detected, therefore implying that there will be some latency prior to a full path being in place. In this instance, the overhead will increase in relation to the number of flows being supported, and the duration of the flows. Label assignment in order to support short flows implies a large overhead. When label distribution is included as part of a reservation protocol (such as RSVP), the overheads and scalability of such a protocol must also be considered.
- Label distribution protocols must work in a reliable manner given that the loss of a control message in this instance could cause a delay in the establishment of a label path. This constitutes a serious impediment to the support of critical applications. As mentioned earlier, the use of TCP with a number of LDPs offers the necessary reliability. In the case of flow-based label assignment and the use of RSVP, the reliable transmission of the LDP information is not guaranteed due to the use of UDP.

3.2.2 Bandwidth Brokering

The resource management model associated with the Differentiated Services Architecture (discussed in section 3.1.3) recognises the distinction between intra-domain resource management, where service providers are responsible for the provisioning of resources within the domain, and inter-domain resource management, where service agreements and traffic contracts are used to determine the amount of traffic that may enter a provider's network. It does not provide any consideration of how these traffic contracts are put in place. It is likely that these agreements will initially be provisioned on a manual basis. One of the "shortcomings" of the DiffServ approach outlined in section 3.1.3 is that in order to allow these agreements to be provisioned in a more automated way (potentially on a fairly short time-scale), rather complex signalling mechanisms would be required. This is where the role of the Bandwidth Broker fits in. Briefly, each of the networks in an end-to-end path has an associated Bandwidth Broker. This broker performs admission control for reservation requests from its own users. When there is a need to alter the amount of bandwidth available for a certain class between the network and an adjoining network, it communicates with the broker managing the adjoining network (using some form of standardised protocol), in order to make the necessary changes to the available resource (subject to the underlying agreements between the two). This in turn may involve the broker of the adjoining network communicating with other brokers further down the path towards the receiver of the flow, in order to provision for resources in an end-to-end fashion. As well as performing admission control, the Bandwidth Broker therefore also has the task of providing policy control (where administrative and pricing policies are accounted for) and may perform reservation aggregation (where multiple resource requests can be aggregated). The Internet2 programme is heavily involved in the development of Bandwidth Broker mechanisms: see "The Internet2 Qbone Bandwidth Broker Advisory Council" [12] for further details.

3.3 Traffic engineering via application engineering and proxies

There are methods by which some degree of traffic engineering can be achieved by implementation choices at the application layer, or by "transparent" proxy services. For example:

- **IP Multicast:** is well suited to delivery of live real-time media transmission to a number of recipients. Transmissions are advertised and receivers are able to join the multicast group(s) that interest them. Multicast also has uses beyond media delivery, and thus the provision of multicast support on SuperJANET4 is to be welcomed and co-operative support in MANs and at campuses should be sought. The implementation and deployment of multicast applications should be encouraged. However, in cases where individuals are seeking media-on-demand, multicast may be less attractive because users are generally not prepared to wait for the "next showing" of multicast content.
- **Tiered hierarchies and mirrors:** where large amounts of data are to be accessed for computational or other requirements, the option to mirror or tier the data among a number of hierarchically organised sites may be attractive. JANET has an open FTP mirror (mirror.ac.uk) and the current proposal to make CERN GRID data available within JANET follows a tiered design. By scheduling the mirroring process at unsociable hours, more bandwidth and thus better performance is available for networked applications during regular working hours.
- **(Adaptive) content profiling:** such activity is currently favoured where content is to be delivered over a medium that does not support the bandwidth requirement, or that reports a high degree of IP loss. Examples include WAP gateways and adaptive video codecs that change their encoding methods when packet loss is detected. The latter is something that H.323 does not offer; H.323 is an example of an application that is not tolerant of anything above mild packet losses.
- **Web caches:** the JANET web cache service has been in operation for some time, and has helped raise awareness of bandwidth savings that can be made by local content caches. As a result many campuses operate their own cache services. However, caches are unable to store dynamic content, and may raise latency in responses to HTTP requests because of the extra "hop" in fetching new pages. Given that HTTP is typically 70% of the inbound traffic to a site from the US [cf. UKERNA's transatlantic billing statistics], sites are keen to minimise their chargeable element on that traffic.

While techniques can be used to reduce demand for bandwidth on a site's external network link, if the link remains congested to some extent, even if only at network "rush hours", some QoS method is highly desirable for important, network-sensitive applications.

In some instances, a many-to-one transmission service is required (the opposite to multicast, in a sense), for example feeding CCTV over IP to a central monitoring point. Such cases may not be common at present, but may become more prominent as always-on devices (cameras or other monitoring devices) proliferate. One might also expect the collection of GRID sensor data to follow this many-to-one pattern. There are no obvious “engineering” tricks to alleviate the bandwidth requirement in such a case.

3.4 Summary

3.4.1 IntServ & RSVP

The IntServ model, which encompasses quantitative services which are potentially a good fit for many of the QoS requirements described in Section 2, requires signalling and admission control in order to reserve resources within the network so as to meet the quantitative service specifications. RSVP, initially defined for signalling associated with single (simplex) flows, is the user-network signalling system of choice associated with this model. Flows are associated with destinations and may include multicast. Resources may be exclusive to a flow or shared amongst flows. Disadvantages associated with IntServ/RSVP at this time are

- state and message processing associated with flows presents a scalability problem for core networks; flow aggregation is under study, but is at an early stage; and
- end-system support for RSVP and application-awareness of QoS requirements in all end-user systems requiring network QoS support, including H.323 codecs, IP telephones, etc., is a prerequisite for deployment: although there is some early deployment of prototype support in some operating systems and applications, this is not yet widespread.

3.4.2 DiffServ

The DiffServ architecture is based on the notion that only a few basic classes of forwarding behaviour (EF and the AF groups) are actually needed: many flows can be aggregated into these few classes. It is recognised that aggregate flows may cross multiple network domain boundaries. The provision of QoS within the interior of each domain is at the discretion of individual domain management (and may exploit QoS facilities at layer 2, layer 3, or a combination). At each interdomain boundary, service level agreements (SLA) are instituted, and monitoring and policing of these on receiving sides is necessary to ensure adherence to the SLA and to prevent resources becoming over-committed. An important advantage of DiffServ for deployment purposes is that there is no requirement to change end-user applications: all marking and policing may be implemented in user-network and network-network edge routers. Some disadvantages of DiffServ, however, need to be noted:

- implementation of per-hop behaviour applies to aggregated flows: a constituent flow may not always receive all the resources it needs, unless it is the sole constituent. For example, delay bounds may be met but bandwidth requirements may occasionally not be;
- suitable provisioning and configuration is essential, specifically to reduce the instances of inadequate resources being available for flows; and
- meeting end-to-end QoS specifications is hard in principle because the services offered by each domain are not constrained to be the same.

In respect of the last of these points, possible ways to alleviate the difficulties associated with multiple domains include:

- a) domains offering similar or identical services;
- b) constraining particular classes of traffic by route pinning; and
- c) using bandwidth brokers to ensure capacity is available for key classes of applications.

3.4.3 LBE

Less than best effort (LBE) is a complementary ‘altruistic’ approach whereby sites may mark non-critical traffic as low priority. The effect is that LBE traffic is discarded first in the presence of congestion, leaving best effort a better service.

3.4.4 MPLS

Complementary to specific techniques for the provision of QoS is traffic engineering. One such technique, MPLS, seeks to exploit traffic engineering properties of connection-oriented network techniques for the benefit of connectionless IP. MPLS header information is incorporated into either an extra (MPLS) header (between layer 2 and 3) or the underlying layer-2 header. The MPLS header includes a label and a class of service (CoS) field, the former to enable LSRs to route the packet, the latter to provide service differentiation. The CoS field can in effect be used to maintain DiffServ-like QoS in an MPLS network. In deploying MPLS, it must be recognised that

- MPLS support is needed throughout a domain;
- Dynamic provision of LSPs is still under development, but support for static provision is available now; and
- MPLS is in effect a single-domain technique: at the edges, packets emerge from all LSPs and cross domain boundaries in the normal way.

It must be recognised that while MPLS represents a potentially powerful tool for network service engineering, it also represents a substantial increase in the complexity of network operation and support.

3.4.5 Bandwidth Brokering

Within the DiffServ model, provisioning is typically static by comparison with the rate at which flows appear and disappear. Where DiffServ is deployed in support of few classes and a relatively small number of important flows, manual, static provision is possible. Bandwidth brokering is being developed in order to enable more dynamic support for provisioning in the multiple domain context. Brokers for each domain negotiate for resources to enable procuring policy to interact with admission control in response to signalled requests.

3.4.6 Conclusions

Overall, in comparing RSVP/IntServ and DiffServ approaches to the provision of QoS, the following considerations are decisive:

- RSVP has seen some early deployment in both routers and end-systems. However, it has not been widely incorporated into end-user applications. It may eventually be deployed as the end-user signalling system of choice, but it is not a practical option for service deployment at the current time.
- DiffServ has to date been the major focus of deployment effort, on the basis that
 - (a) it is directed at handling aggregate flows, which in turn are currently seen as the more promising way forward because the alternative, non-aggregated, individual “micro-flow” model with its associated (reservation and path) state in every router is perceived not to be scalable in terms of the amount of state required in core routers; and
 - (b) it can generally be deployed without user-application or end-system modification.

Behaviour under failure of network components is an important area of on-going study and development, particularly in the case of MPLS. It is important to realise that there is in effect a ‘feature interaction’ which occurs between use of alternative routes for resilience purposes and for traffic engineering for support of different qualities of service. In the extreme case, it is possible in effect to destroy the resilience associated with all qualities of service, including best effort, unless care is taken in defining fail-over configuration for all service qualities.

References for section 3

- [1] R. Braden, D. Clark, S. Shenker, “Integrated Services in the Internet Architecture: an Overview”, IETF, IntServ Working Group, RFC 1633, June 1994.
- [2] R. Braden et al., “Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification”, IETF, RSVP Working Group, RFC 2205, September 1997.
- [3] L. Zhang et al., “RSVP: A New Resource ReSerVation Protocol”, IEEE Network, vol. 7, September 1993, pp. 8-18.

- [4] D. Black et al., “An Architecture for Differentiated Services”, IETF, DiffServ Working Group, RFC 2475, December 1998.
- [5] R. Braden, D. Hoffman, “RAPI - An RSVP Application Programming Interface – Version 5”, Internet Engineering Task Force (IETF), RSVP Working Group, work in progress, <draft-ietf-rsvp-rapi-05.txt>, August 1998.
- [6] B. Lindell, “SCRAPI – A Simple ‘Bare Bones’ API for RSVP”, Internet Engineering Task Force (IETF), RSVP Working Group, work in progress, <draft-lindell-rsvp-scrapi-02.txt>, February 1999.
- [7] E. Rosen, A. Viswanathan, R. Callon, “Multiprotocol Label Switching Architecture”, IETF, MPLS Working Group, RFC 3031, January 2001.
- [8] Y. Rekhter, E. Rosen, “Carrying Label Information in BGP-4”, IETF, MPLS Working Group, RFC 3107, May 2001.
- [9] L. Andersson, P. Doolan, N. Feldman, A. Fredette, B. Thomas, “LDP Specification”, IETF, MPLS Working Group, RFC 3036, January 2001.
- [10] A. Banerjee, J. Drake, J. Lang, B. Turner, K. Kompella, Y. Rekhter, “Generalized Multiprotocol Label Switching: An Overview of Routing and Management Enhancements”, IEEE Communications Magazine, January 2001, pp 144 – 151.
- [11] B. Jamoussi “Constraint-Based LSP Setup using LDP”, IETF, MPLS Working Group, work in progress, draft-ietf-mpls-cr-ldp-05.txt, February 2001.
- [12] S. Hares, “The Internet2 Qbone Bandwidth Broker Advisory Council”, <http://www.internet2.edu/qos/qbone/QBBAC-0107.shtml>.
- [13] IEEE 802.1D: “Media Access Control (MAC) Bridges (Incorporating IEEE P802.1p: Traffic Class Expediting and Dynamic Multicast Filtering)”, 1998.
- [14] *For an introduction see, for example, G. Armitage, Quality of Service in IP Networks: Foundations for a Multi-Service Internet*, MTP, 2000 (ISBN 1-57870-189-9).
- [15] R. Yavatkar, D. Hoffman, Y. Bernet, F. Baker, and M. Speer, “SBM (Subnet Bandwidth Manager): A Protocol for RSVP-based Admission Control over IEEE 802-style networks”, IETF, ISSL Working Group, RFC2814, May 2000.

4 Policy

4.1 Requirement

The need for policy in relation to quality of service deployment in a network arises directly from the non-zero probability of transient queues forming in routers. Such queues arise as a consequence of the bursty nature of packet flows: if even small bursts on several input ports of a router contend for the same output port, a transient queue will form. (Non-transient queues are a property of an overloaded network, and can only be dealt with by the addition of capacity.)

In general, any such queues in a delivery path will add delay and jitter to packet transport time. The need to control this according to the type of traffic to which a packet belongs implies that packets are not treated identically. Preferential treatment for some may also be viewed as degraded service for others. If admission control is in operation, then some traffic (packets) may not be admitted, or only with reduced quality of service. Policy is required to determine which traffic receives which type of treatment, i.e., what quality of service.

Since in a multiservice network different traffic may no longer receive the same quality of service, it becomes necessary to authorise traffic to receive the ‘better’ classes of service. Authorisation in turn implies a need for authentication. And both imply a need for monitoring. All are necessary components for policy implementation. Although in a commercial environment much of this mechanism may be cast almost directly into financial terms, it should be noted that the mechanisms are required in some form to ensure the operation of the policy rules independent of financial charging at point of use.

4.2 Inter-domain issues

A number of methods exist for giving priority to IP traffic. These may perform well and be quite manageable in a constrained environment, in particular where only one administrative QoS domain is present. However, the reality of provision of end-to-end QoS is that the QoS mechanism will need to be implemented across multiple administrative domains that may be running router equipment from a variety of vendors and using quite different transport methods. There are thus technical (including interoperability) and, more importantly, political issues to be resolved in offering an end-to-end QoS service.

4.2.1 Requirements

A typical scenario may involve a requirement for a guarantee of at least 10Mbps bandwidth between users at two UK university sites. IP traffic will traverse the department (LAN), campus, regional (MAN) and SuperJANET backbone networks, a total of seven administrative domains. SuperJANET4 (SJ4) is a transit network for its regional MANs, but must be instrumental in supplying end-to-end QoS.

The problem becomes more complex when the bandwidth is required on demand on a short-term basis, or when a pre-booked allocation is requested (e.g. for a videoconference one month in advance). The network must be provisioned for the required service, and if the provision cannot be met, the user or service must be notified. In the videoconferencing case, one might expect a certain level of priority traffic to be provisioned at a site’s egress point to a MAN, and for that provision to allow a certain number of conferences to be operated concurrently, and thus booked in advance. Prior allocation is a much simpler task than dynamic provision.

In a scenario where the endpoint for a QoS agreement is not on JANET, inter-domain agreements would need to be negotiated with the appropriate ISPs. For commercial sites external to JANET this may be problematic, but in the academic environment one hopes that UKERNA can liaise with at least GÉANT and Internet 2 to develop common QoS methods. The Internet 2 QoS Working Group is already collaborating with the GÉANT TF-NGN group to this end for definition of Premium IP services (based on DiffServ EF PHB). (As an idea of the overall scale of the problem, there are at least 8,000 autonomous networks in the Internet, and well over 100,000 IPv4 network prefixes advertised on the backbone.)

In the TEN-155 era, DANTE (and UKERNA) offered a managed bandwidth service (MBS) that enabled ATM permanent virtual circuits (PVCs) to be established between end sites across Europe. The MBS service allowed dedicated bandwidth (effectively a “virtual leased line”) between end sites, and also enabled network level tests (e.g. DiffServ, IP multicast and IPv6). DANTE acted as co-ordinator for international requests, while the JANET service was managed by UKERNA (though it never reached full production status).

The new European GÉANT and SJ4 core networks no longer provide the mechanism to offer a virtual leased line service, that is to say there is no provision for dedicated, distinct paths through the network between sites, and alternative approaches will be required.

4.2.2 Provisioning

The application of QoS methods to offer a “better than best effort” service implies the requirement to allocate a certain proportion of the network capacity to that service. In the case of a “virtual leased line” service, that allocation will be at the expense of the regular IP traffic (unless additional capacity is provided explicitly for that purpose). It is important that the Premium IP traffic does not adversely affect the availability of a good, regular service. The operators will thus have to decide to what level to allocate bandwidth to Premium (or other “better than best effort”) services (e.g. 10%, 20%, or more?), and by what method to make that allocation.

The simplest method is to use manual provision of Premium IP bandwidth. In such instances the required allocations can be calculated in advance. The alternative is to use some form of signalling to set up QoS channels or paths on demand, e.g. as planned by Abilene’s Premium Service bandwidth broker [<http://qbone.internet2.edu/bb/>] for automatic provision.

In either case there will always be the problem of worst-case provisioning. If four sites on a network each have an agreement to handle a 10Mbps Premium IP path between them, then it is possible that at any instant 30Mbps of priority reserved bandwidth will be required into one or more of the sites. If worst-case requirements are not provided for, then service level agreements (SLAs) may not be able to be met 100% of the time. Each SLA will typically consist of a service-level specification (SLS) in each direction between the agreeing networks. It is quite possible that the SLS is not symmetric, and quite preferable that standard SLSs are agreed for certain traffic types or profiles to ease service provision.

Destination aware vs. destination unaware

A QoS service may be offered destination aware (i.e. the admission control method for the EF PHB is based upon a given IP destination, and in effect a “virtual leased line” is enabled) or destination unaware (the sender has agreed EF PHB priority based on source address or similar source flow properties, and the service is implemented as an aggregate Premium IP capacity).

In the case of destination aware provisioning, a path through the domain can be configured, e.g. by MPLS methods as demonstrated on the Abilene Premium Service or by DiffServ EF PHB on all routers within the domain. In the case of destination unaware provision, a bandwidth limit needs to be set, e.g. if the sum of all exit points is 1Gbps, and one chooses to accept 50% of the capacity as Premium IP, one would accept in up to 500Mbps. The latter is more flexible and requires less direct management.

Capacity

The availability of DWDM and 10 Gigabit Ethernet solutions will make it possible for the core JANET network to be over-provisioned at a relatively attractive cost; it is envisioned that SJ4 will migrate to a 10Gbps core early in its lifecycle. It is thus entirely possible, especially when migrating from the 155Mbps SuperJANET3 predecessor, to reserve a significant percentage of the newly available bandwidth for QoS purposes, without adversely affecting regular IP traffic. Were 20% of the SJ4 core allocated to Premium IP (“virtual leased lines”), then 0.5Gbps would be available, leaving 2Gbps for regular use. However, it is worth noting that a single GRID application requiring 1Petabyte/year (whether between two sites or for many sites feeding data to a single site) would immediately consume 0.3Gbit/s, or two-thirds of that reserved for Premium IP traffic. This has two immediate consequences; UKERNA needs to

1. Assess the requirement for Premium IP bandwidth over the next 3 years to gauge whether its planned expansion of SJ4 is sufficient.
2. Determine charging methods where significant proportions of the core bandwidth are reserved for single projects.

There are clear advantages to carrying all academic traffic on a single network, but ample provisioning and funding for it is a key requirement.

It should be noted that there are certain limits on external traffic requirements to/from SJ4. These currently are 2x622Mbit/s to Abilene/ESNet (USA), 622Mbit/s to TEN-155 (soon to rise to 2.5Gbit/s), and approximately 1Gbit/s at the LINX (UK commercial exchange). There is clearly the potential for combined

inbound traffic volumes to exceed that available on the core, though the inbound traffic would be spread across the core. Application of QoS mechanisms on international peering points (e.g. Abilene) will be desirable (including, for example, common adoption of LBE marking, if agreed with Internet 2); however, the routers at the boundary of the SJ4 domain must have sufficient capability to provide the QoS functionality that will be required.

At the other end of the spectrum, it should not be forgotten that many JANET sites, especially new FE colleges coming online, have only 2Mbit/s connections, and that a number of sites that might hope to benefit from GRID applications may “only” have 8Mbit/s links. In such cases, the ability to offer high bandwidth QoS is limited, and the remaining regular best effort IP availability is also a more precious commodity.

4.2.3 Other projects

There are a number of major projects already looking at end-to-end QoS in academic environments. These are typified by the activities of the Internet 2 QBone [<http://qbone.internet2.edu>] and Web100 [<http://www.web100.org>] in the US, and by the TF-NGN work within GÉANT in Europe.

GÉANT

Under GÉANT, the TF-NGN [<http://www.dante.org.uk/tf-ngn/>] group has formulated a proposal for the provision of two services on the GÉANT core [cf. GÉANT D9.1 and SEQUIN report]:

1. **A Premium IP service.** This uses the Expedited Forwarding Per Hop Behaviour (EF PHB) of the DiffServ model [RFC 2474, 2475] to offer a virtual leased line service.
2. **An IP+ service.** This is a combination of a prioritised bandwidth (“better than best effort”) and a guaranteed bandwidth service and uses various levels of Assured Forwarding Per Hop Behaviour (AF PHB) [RFC 2597].

The Premium IP service is the focus of the GÉANT D9.1 deliverable. It is defined by four metrics: bandwidth, IP packet loss, delay, and delay variation, while IP+ is defined by bandwidth with looser requirements of the other parameters. The GÉANT Premium IP service offers QoS across the core network between NRNs; provision of QoS to the end sites then becomes the responsibility of each NRN.

The EF PHB is aimed at low loss, low delay, low jitter networking, using DiffServ codepoint 101110. The default PHB codepoint is 000000; any unrecognised codepoints are mapped to the default (i.e. best effort IP). The LBE codepoint, as suggested by the Internet 2 Scavenger project, can be integrated into this scheme.

The GÉANT Premium IP document defines Premium IP as mapping to EF PHB, though this may in reality be implemented in any conformant way, e.g. by priority queuing or an ATM CBR link; the choice is up to the participant network. The document describes the interface specification between domains, the traversal of which must be compliant to EF PHB.

The router requirements are primarily that the DSCPs can be recognised, and that the router can measure the arrival rate and compare it to the defined profile (as, for example, can be done using the Committed Access Rate (CAR) method).

Timely forwarding can be achieved in a number of ways, e.g. strict priority queuing, weighted fair queuing (WFQ), or various round robin algorithms. Each will perform differently, and thus affect the handling of best effort traffic differently.

Experience has been drawn from a number of European research projects, including TEQUILA, AQUILA and SEQUIN.

The TF-NGN meeting and mailing list discussions have raised a number of interesting issues and points regarding QoS and high-speed connectivity, for example:

1. The regular best effort IP traffic must not be starved, either by over-allocation of core bandwidth to Premium IP or other “better than best effort” services, or by poor choices of queuing methods for servicing IP traffic at participating routers.
2. There is no consistent definition of what is meant by an over-provisioned network. Is it that there is no congestion, or that average utilisation is some fraction of the provisioned bandwidth?
3. There are operational expenses associated with deploying mechanisms in the network that enable traffic engineering, both in terms of hardware and staff skills. Is it cheaper to overprovision than

make more efficient use of what you have? The money used to manage networks could be used to buy more capacity.

4. MPLS is popular with telcos who want to allocate low bandwidth VPNs to large numbers of customers. Do academic networks require that model of operation? If privacy is an issue, other security measures can be applied (on the network, rather than in it). It has also been argued that MPLS may be deployed most easily if present in a network from day one, rather than being turned on later. (LeNSE is an example of a UK MAN running MPLS.)
5. Latency is a very important network characteristic. It is to be noted that the SLA UUNet offers to its customers lists latency bounds as the first criterion.

QBone

The QBone [<http://bone.internet2.edu>] is a QoS-enabling network being developed as part of Internet 2. Unlike the provision of IPv6 and Multicast IP as tunnelled overlay networks, the provision of QoS requires more fundamental engineering. The initial QBone architecture is described in: [<http://www.internet2.edu/qos/wg/papers/qbArch/1.0/draft-i2-qbone-arch-1.0.html>]

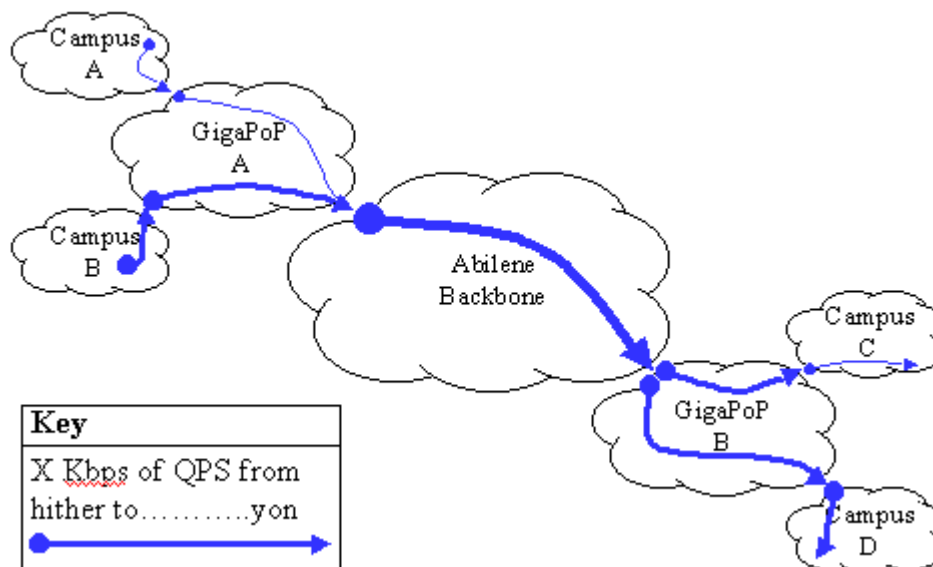
The QBone model breaks the overall network up into DiffServ (DS) domains. The aim is to provide additional inter-domain services beyond regular best effort. The first new service is the Premium, or Virtual Leased Line (VLL), service. For this to be enabled, every QBone DS domain must support the EF PHB and configure its traffic classifiers and conditioners (meters, markers, shapers, and droppers) to provide a VLL service to EF aggregates.

The QBone architecture is aiming to deploy bandwidth brokers [<http://qbone.internet2.edu/bb>] in each DS domain to enable dynamic QoS provisioning. The SIBBS (Simple Inter-domain Bandwidth Broker Signalling) protocol is being designed for this purpose, but is still some way off full-scale deployment.

Abilene Premium Service (APS)

The Abilene Premium service (APS) [<http://www.internet2.edu/abilene/qos/>] aims to implement IP QoS using the Internet 2 QBone principles.

The end-to-end provisioning has the classic problems of inter-domain policy management as described above. In Abilene, the GigaPoPs can be loosely compared to the MANs on SJ4:



The APS design has been discussed in a number of documents [<http://www.internet2.edu/abilene/qos/>]. The initial trials also show that the MPLS DS-TE traffic engineering methods work on Cisco products [http://www.internet2.edu/abilene/qos/internet2_dste.pdf]. The DiffServ-aware MPLS-TE method is used for edge to edge tunnels.

Clearly the APS work on MPLS in the DS context should be tracked. Another paper of relevance is “On the utilisation of MPLS for interdomain flows”, Olivier Bonaventure, [<http://www.info.fundp.ac.be/~obo/private/COST263-Budapest.pdf>]

The choice of whether to deploy MPLS at an early stage is an important one for UKERNA. Experience from the QBone initiative will be invaluable to inform future developments within SJ4.

4.3 Policy enforcement

There are many aspects to QoS policy enforcement. The subject area is a complex one, but one which must be addressed at some stage in a QoS service roll out. However, it is unlikely to be prudent to attempt to do so from day one. There is also a performance impact for both running DiffServ on a router and on top of that an extra loading for policy monitoring and enforcement (e.g. traffic shaping or packet drops). Admission control rules at QoS-enabling routers may include a variety of properties, e.g. IP source or destination, the ToS field, the ports used, or the protocol. Matching against these requires computational effort.

It should be noted that policing is independent of the general issue of congestion and congestion avoidance for regular traffic, which itself is usually performed via traffic dropping, e.g., using an algorithm such as WRED (Weighted Random Early Detection) and ECN (Explicit Congestion Notification). Whilst strictly beyond the scope of QoS provision, the use of such techniques needs to be understood and where appropriate accommodated within the provisioning of SJ4.

Policing does involve monitoring of the SLA between domains, and in turn of the SLS in place for traffic passing in both directions between those domains. In a scenario such as a set of NRNs connecting to the GÉANT core, there is a choice as to whether to police on ingress (from GÉANT to the NRN). The GÉANT document recommends not doing so in the initial deployment.

There should be a revision of provisioning between domains at regular intervals, for which monitoring statistics will be essential.

4.4 Initial policy model

There are three IP QoS services that could reasonably be implemented in a first-stage QoS deployment on SJ4:

1. Premium IP, using DiffServ EF PHB, equivalent to a virtual leased line where the IP destination is known.
2. IP+, a “better than best effort” service, up to an ingress capacity, using DiffServ AF PHB.
3. LBE, “less than best effort”, a service where LBE traffic is dropped first at any congestion points on a network.

DS codepoints have been suggested/agreed for all these traffic types in other communities and need to be supported within the SJ4 and indeed the ac.uk configuration. The AF PHB has four main classes of traffic; an initial deployment may only require one or two such classes to be used, merely to differentiate “better than best effort” from regular best effort IP.

A router infrastructure should be established to support the three IP services enumerated above. LBE traffic may also be routed down failover links as a means to utilise resilient links without the danger of network implosion if the resilient link is already used for regular service traffic when the major link fails.

For a longer-term roll-out plan, it is interesting to note the four-phase deployment schedule for APS (the Abilene Premium Service), each with rather esoteric names:

1. Sweetwater:
 - Measurement infrastructure: Surveyor [<http://www.advanced.org/surveyor/>] + SNMP + whois + HTTP access
 - Edge policing: CAR + perhaps QPPB (QoS Policy Propagation by BGP)
 - Manual set-up of policies on routers
2. Midland:
 - EF core forwarding: using MDRR (modified deficit round robin)

3. Odessa:

- EF Edge forwarding: MDRR
- Automated set-up: SIBBS bandwidth broker + possibly DS-TE

4. Pecos:

- Shaping: GTS (Generic Traffic Shaping)

For SJ4 it is suggested that a monitoring system be put in place at the earliest opportunity, similar to that in existence on Abilene. Policies could be set manually on routers, accompanied by basic policing. The EF PHB should be implemented early, but automated bandwidth brokering and advances in policing mechanisms need to be tracked and addressed further in a later phase (the Abilene activity, in particular, should be tracked).

The first phase thus relies on over-provisioning (which, currently, SJ4 is – we have yet to see serious GRID requests come in) rather than signalling between brokers in connected DiffServ domains.

QoS provision at a campus is a campus issue. The campus may choose to mark its allowance of Premium IP traffic as it sees fit (if it wants priority Quake traffic, that is a site issue), though the nature of the marking (destination aware or destination unaware) is important with respect to provisioning in the MAN and subsequently the SJ4 core.

A campus may have a sophisticated local brokering system with authenticated admission, or it may simply choose, for example, to mark only fixed videoconference suite IP traffic (as is the case in the Welsh Video Network). The “last mile” connection, to the end-host on the desktop, is also important, for which high-speed switched layer 2 equipment should be used, and mappings between layer 2 priority values [IEEE 802.1D:1998] and the ToS field in layer 3 may be beneficial.

The question for end-to-end provision for all end sites then becomes how much provisioning can be passed into the MAN and the SJ4 core for both virtual leased lines and destination unaware traffic? UKERNA will need to decide what fraction of its core bandwidth to offer for “better than best effort” services.

5 Proposed 'road map'

In producing such a 'road map', the Think Tank recognises that both user needs and technological support for QoS and traffic engineering are undergoing rapid evolution. Any such proposals for prototype service development and testing need to be kept under regular review. Within the recommendations are included specific activities for tracking particular emerging technological developments. However, it is suggested that the overall road map should itself be kept under regular review as part of the overall programme: tentatively, it is proposed that an annual review may be appropriate.

5.1 Technology Recommendations

Section 3 described the options available to support multiple levels of service. Strategically, the DiffServ model architecture offers support for aggregated flows within the core (the scaling significance of which has already been noted), as well as avoiding the immediate requirement for end-user signalling application and system support. It is therefore recommended as the initial basis for the introduction of QoS support into SJ4.

1. For time-sensitive applications, such as videoconferencing and voice over IP, DiffServ Expedited Forwarding (EF) is recommended. In order to enable extension beyond SJ4, it is recommended that this be harmonised with the emerging European GÉANT Premium service and any corresponding service offered by Internet2 / Abilene in the USA.
2. An integral aspect of this recommendation is the necessity to provision support of this Premium service: a policy allocation of a percentage of capacity to this service on the core network has to be made. An assessment of this figure is required, but a figure of 20% is proposed initially.
3. As part of harmonising this service with that to be offered in Europe and elsewhere, it is recommended that service trials of Premium IP on GÉANT be tracked.

The DiffServ architecture recognises that individual network domains may achieve specific per-hop behaviour in different ways. Service Level Specifications and Agreements need to be established throughout the UK domains (campus, MAN, SJ4 backbone) in order to achieve end-to-end operation. The establishment of consistent end-to-end QoS support will be greatly simplified if essentially identical SLSs can be used throughout the UK domains.

4. It is recommended that the possibility of establishing a uniform Premium SLS be investigated.

There are two types of service for which better than best effort service is required, though bounded delay and jitter are not essential to the same extent as for interactive audio or video communication. Such applications include interactive use, as for example shared workspace applications and streaming audio/video.

5. For support of services such as streaming video and interactive applications (particularly synchronous shared), investigation and trial of the DiffServ Assured Forwarding (AF) class of services is recommended (to be harmonised with any similar service deployed in Europe as a result of the IP+ service being defined in the SEQUIN project).

Particularly in applications such as streaming video, the importance of providing scalable network support which also assists in avoiding server overload is recognised.

6. It is recommended that the use of multicast for streaming application support be promoted.

In order to support specific application requirements, such as those arising from GRID, e-Science, or former MBS applications, the use of traffic engineering techniques will be needed. In respect of former MBS usage, it must be recognised that IP-level experimentation can not be supported directly in this way, since the IP service is a production one.

7. It is recommended that traffic engineering support be investigated, including particularly MPLS deployment, in support of especially particular categories of Premium use, GRID, and former MBS traffic. Investigations should include dedicated provision and route-pinning for specific services.
8. It is recommended that the Abilene Premium Service work exploiting MPLS and DiffServ traffic engineering be tracked.

The concept of less-than-best-effort (LBE) traffic is based on the idea of encouraging users or sites to mark suitable traffic low priority. Such traffic would be discarded in favour of conventional best-effort traffic in the face of congestion. One aspect of such traffic is that it is amenable to routing over fail-over links without affecting best-effort traffic under fail-over conditions.

9. It is recommended that LBE deployment on SJ4 be investigated.

In order to support testing, development, and operation of a QoS-enabled network, provision of adequate monitoring facilities will be essential.

10. It is recommended that a short study, followed by provision, of monitoring facilities be made to enable monitoring support of prototype service testing and operation of network QoS facilities. (See Section 5.2.6.)

A possible future model for on-demand Internet QoS support is that RSVP is used for user-edge signalling, and that, within the limits of policed SLAs, aggregated flow support is deployed within the core, whether based on aggregated IntServ specifications or evolution of DiffServ specifications. In support of inter-domain aggregated QoS support, the concept of a Bandwidth Broker is under development to enable dynamic alteration of the level of provisioning for different classes of traffic to be negotiated between domains (within an overall envelope of physical service provision).

11. It is recommended that the development and deployment of Bandwidth Broker technology be tracked.

To enable support of policy allocations for different classes of traffic to be implemented, both through policing of dynamic policy limits (and, by implication, to support eventually such technology as Bandwidth Brokering) the Common Open Policy Service is under development to provide a link between policy servers and policy enforcement points.

12. It is recommended that the development of COPS be tracked.

In respect of DiffServ, it should be recognised that experimentation and experience is required to determine parameters for policing, scheduling, packet discard, and provisioning limits. Establishment, across as wide as possible a set of domains, of a uniform set of DiffServ code-points and scheduling disciplines is to be encouraged in support of Premium and IP+ services.

5.2 *Prototype service testing*

As stated in the Introduction, the overall approach taken by the Think Tank has been application service led. It is proposed that this principle be extended to any prototype service development and rollout of QoS across JANET, in order to ensure these are driven by service needs, and the resulting services match end-user requirements. To be effective, QoS must be implemented on an end-to-end basis, which in the JANET environment means campuses, MANs and the SJ4 backbone. It is therefore essential that all these domains participate in the QoS service development.

Section 2 identified a number of different application areas, all of which have different QoS requirements. It is proposed that an open call for proposals be issued for existing application service development projects to work with UKERNA to develop QoS services. It is envisaged that the application service projects will fall into the following areas, most of which have been identified in Section 2 on Requirements as drivers for QoS.

5.2.1 *Videoconferencing*

Lecture room for group teaching:

Videoconferencing to support large group teaching has been used for a number of years and the technology to support this has been based upon leased line, public ISDN network, or ATM infrastructures which have effectively provided a guaranteed QoS. Recently there has been a move toward videoconferencing services based directly on IP, stimulated in part by the decline of ATM as a strategic wide area network technology, coupled with end-user ambition to avoid leased line or ISDN charges.

Using a multiservice IP network such as JANET for a timing sensitive application such as videoconferencing means that QoS is essential. To enable the delay demands of this application, the use of Differentiated Services Expedited Forwarding is recommended.

Desktop videoconferencing for informal small group meetings or on a one-to-one basis:

In many cases the use of more informal desktop videoconferencing may be more appropriate than the lecture-room style of videoconferencing, as for example, to meet on a one-to-one tutor-to-student basis, or for small informal meetings. For desktop videoconferencing to be effective, even in these informal situations, minimum levels of good audio and video quality still need to be maintained, though perhaps at not quite as high a level of quality as required for lecture-room based videoconferencing. It is anticipated that good-

quality QCIF may typify this style of use. Albeit that the quality (frame rate or resolution) may be lower than lecture- or conference-room videoconferencing, the delay requirements remain, and the use of Differentiated Services Expedited Forwarding is recommended.

Video streaming

Depending upon the mode of use of video streaming, different quality requirements will apply and these are described in section 2. The quality spectrum spans the two following scenarios:

- A. Download a video file and then play — this requires no QoS beyond best-effort.
- B. Play a video file whilst streaming its content over the network — this requires QoS support, although not to the extent that an interactive, interpersonal application such as videoconferencing requires.

It is recommended that for scenario B that Differentiated Services Assured Forwarding be used.

5.2.2 Voice over IP

As an interactive, interpersonal communication application strict quality requirements are essential for support of voice. For any application trials in this area two modes of use of voice are envisaged:

- A. **IP Telephony** This class of services exploits IP network transport service, with voice call management handled centrally and facilitated by a single server.
- B. **Voice over IP** This involves calls being routed via telephone exchanges and gatewayed onto the IP network. It may also include individual station-to-station calls initiated without the assistance of any central telephony support.

Because of the real-time interactive nature of this application, with its consequent requirements on delay and jitter, it is recommended that Differentiated Services Expedited Forwarding be used.

5.2.3 Traffic Engineering for GRID & former MBS applications

The GRID requirements in terms of network provision have not yet been fully elaborated, although in Section 2 a broad range of GRID applications have been identified within the emerging e-Science disciplines. It is likely that some of the GRID applications will involve transporting huge amounts of data across the network — data transfer that may need both to be protected from other traffic and protected from interfering with other traffic sharing the network in ways that QoS mechanisms alone cannot do. Here the use of traffic segregation and engineering techniques are appropriate, both to enable such network usage to operate up to defined limits and to ensure that it does not impinge on other traffic that shares the same physical network. For these GRID applications, it is recommended that use of MPLS be investigated. Other GRID applications will span a range of QoS services, for which it is recommended that Differentiated Services AF and EF are used as appropriate.

It is to be noted in this context that there are available proprietary mechanisms (MPLS guaranteed bandwidth from Cisco, for example) which may offer potentially appropriate solutions. The proprietary nature of these, their potential for standardisation, and their appropriateness for use in any particular network domain needs careful assessment.

5.2.4 Outreach

With the requirement for teaching and learning to take place beyond the traditional classroom and lecture theatre and into the workplace and even home means that the 'reach' of the network must be extended. Lecturers and researchers working from these 'outreach' locations will still need and expect to use the same range of applications as used from the traditional places that the network covers, and at a similar or consonant level of quality. An important aspect of any service development activity will be to trial QoS services across any JANET outreach network infrastructure.

In this respect, an important determining factor for synchronous conferencing applications will be the nature of the QoS-supporting SLAs which UKERNA can achieve with commercial ISPs, for example, in support of ADSL access.

5.2.5 Advance booking: on-demand vs. scheduled use of Premium IP

For scheduled videoconference applications, such as teaching, conference-room meetings, etc., currently deployed systems in effect book network capacity in advance at the same time as the physical rooms are booked. Since the transmission capacity required to support the high-quality video streams is a substantial fraction of link transmission capacity reserved for such use, the necessity for this is self-evident.

Within Recommendation 2 above, it is recommended that a fraction of overall network capacity needs to be allocated for use (when required) by Premium IP applications. In order not to over-commit this Premium IP provisioning, applications scheduled in advance need to be able to book the necessary network resources in advance. It is proposed that a project to investigate provision of such a booking system be set up. An initial version would have essentially static knowledge of the provision made within each network domain, and would book resources against this. In the future, one might speculate that the bandwidth brokering architecture might be extended in support of such future booking.

5.2.6 Monitoring

There is universal agreement on the necessity of monitoring in support of the introduction of QoS and traffic engineering into the network. The subject is a large one and has many perspectives. There are also many communities which already have, are just starting, or are currently proposing 'monitoring' activities, among them UKERNA, RIPE, Terena, and the GRID community (UK and international).

End-user communities may typically be interested in the performance of specific applications and relating this to underlying network performance.

An operator needs to have a view of the state of the network, the flows within it and the associated parameters. In particular, there is a need to monitor usage in connection with policing, provisioning and ultimately charging.

Intrinsically, network-level monitoring is needed within each network domain, both to support the management of the individual network domain and also to enable monitoring and policing of interdomain SLAs. In the context of introducing QoS support into the network, it becomes important to be able to identify performance with respect to each of the traffic classes being supported, as perceived at a variety of points within the network, including network edges.

In addition, for specific value-added services (videoconferencing being one such), it seems likely that service-level monitoring will be needed to support those running that particular service. It should also be noted that since such services are end-to-end, this raises questions of how best to engineer the support and management of such monitoring.

It is proposed that this should be a funded area of further activity. Additional study is needed to agree the overall requirements for monitoring, taking into account existing efforts in this area. Following this, effort and infrastructure needs to be put into place to implement the identified requirements.

5.2.7 Timescales

Some of the applications mentioned above are already being developed with a view to providing services within the JANET community. The proposed development and piloting of QoS services is needed to support such application service developments for which commitment has already been made — examples are: the development of IP based videoconferencing services in Wales and Scotland; the e-Science initiative; and associated trial applications such as Access Grid.

It is recommended that a call for proposals aimed at organisations involved in leading service development activities be issued with a view to initiating QoS service development and application trialling during 4Q01.

5.3 Issues for on-going study

5.3.1 Authentication & authorisation

As policing facilities become available, so the necessity to authorise use within a particular category of service becomes a requirement. At the application level, this requirement is already apparent for such applications as off-site access to e-mail and GRID-based e-Science applications. Since the level of provisioning for and use of particular categories of service is a matter of policy, so it becomes necessary to authenticate users and authorise use against allocations. It should be noted that in principle this is a domain-

based activity, though through established chains of trust there is the potential to hand-off specific aspects of such operations to co-operating domains.

5.3.2 Policy model development

The way in which an initial, statically provisioned service may be extended in the future to support on-demand, signalled services is an area for substantial future study. It is possible that RSVP may become more widely deployed in support of user-network-edge signalling. Within the core of a network, aggregated provision is anticipated to continue, with use of bandwidth brokers providing support for dynamic inter-domain requests for provisioning.

5.3.3 GMPLS

Currently, there is considerable interest in the extension of MPLS to support more general traffic engineering than that which has been addressed in this report. Initially stimulated by interest in traffic engineering for SDH/SONET and wave-division multiplexing transmission schemes (and dubbed MP-Lambda-S in the latter context), there is now substantial work within the IETF on Generalized MPLS (GMPLS), which will be of increasing interest in future as DWDM-based network provision is deployed.

5.4 Related issues

5.4.1 Firewalls

The general need for firewalls needs no comment or explanation here. However, it must be noted that if QoS services are to be provided to users behind a firewall, this places new constraints and requirements on the performance of the firewall, essentially the same as those placed upon (edge) routers in respect of ability to classify, forward and schedule. Performance in these respects will need assessment for their impact on the delay and jitter performance of a firewall.

5.4.2 Privacy & encryption

Privacy and security of inter-personal communication over the Internet is of rapidly growing concern. It is important to note that this applies equally to those newer forms of communication at which the introduction of QoS services is particularly targeted.

For example, the encryption of audio or video communication technically presents no insuperable problems. It does, however, presuppose the existence of a suitable infrastructure, for obtaining and exchanging keys, for example, and as a pre-requisite, user authentication. In particular, this infrastructure needs in principle to be extended to services exploiting network multicast.

Glossary

ADSL	Asymmetric Digital Subscriber Line
AF	Assured Forwarding
API	Application Programming Interface
APS	Abilene Premium Service
ARQ	Automatic Repeat Request
ATM	Asynchronous Transfer Mode
AVO	Astrophysical Virtual Observatory
BGP	Border Gateway Protocol
BT	British Telecom
CAR	Committed Access Rate
CBR	Constant Bit Rate
CCTV	Closed Circuit Television
CERN	European Organisation for Nuclear Research
CODEC	COder–DECoder
COPS	Common Open Policy Service
CoS	Class of Service
CR-LDP	Constraint-based Routing–Label Distribution Protocol
DiffServ	Differentiated Services
DSCP	DiffServ Code Point
DWDM	Dense Wavelength-Division Multiplexing
ECN	Explicit Congestion Notification
EF	Expedited Forwarding
e-mail	Electronic mail
ESnet	Energy Sciences Network
FE	Further Education
FEC	Forwarding Equivalence Class
FTP	File Transfer Protocol
Gbit/s	Gigabits per second
Gbps	Gigabits per second
GMPLS	Generalized MPLS
GTS	Generic Traffic Shaping
HE	Higher Education
HEFC	Higher Education Funding Council
HTTP	HyperText Transfer Protocol
id	identifier
IETF	Internet Engineering Task Force
IN	Intelligent Network
IntServ	Integrated Services
IP	Internet Protocol
IPv6	Internet Protocol Version 6
ISDN	Integrated Services Digital Network
ISP	Internet Service Provider
JANET	Joint Academic Network
kbps	kilobits per second

LAN	Local Area Network
LBE	Less than Best Effort
LDP	Label Distribution Protocol
LHC	Large Hadron Collider
LINX	London InterNet eXchange
LSP	Label-Switched Path
LSR	Label Switch Router
MAN	Metropolitan Area Network
MBS	Managed Bandwidth Service
MC	Monte Carlo
MCU	Multipoint Control Unit
MDRR	Modified Deficit Round Robin
MPLS	Multi-Protocol Label Switching
MPLS DS-TE	MPLS DiffServ Traffic Engineering
ms	milliseconds
NILTA	National Information and Learning Technologies Association
NNW	Network North West
NOSC	Network Operations Service Centre
NRN	National Research Network
NVO	National Virtual Observatory
PHB	Per Hop Behaviour
PVC	Permanent Virtual Path
QCIF	Quarter Common Intermediate Format
QoS	Quality of Service
QoS TT	QoS Think Tank
QPPB	QoS Policy Propagation by BGP
RAPI	RSVP Application Programming Interface
RC	Research Councils
RIPE	Réseaux IP Européens
RSVP	Resource Reservation Protocol
RSVP-TE	Resource Reservation Protocol – Traffic Engineering
SBM	Subnet Bandwidth Manager
SCRAPI	Simplified RSVP API
SDH	Synchronous Digital Hierarchy
SDSL	Symmetric Digital Subscriber Line
SEQUIN	Service Quality across Independently Managed Networks (project name)
SHEFC	Scottish Higher Education Funding Council
SIBBS	Simple Inter-domain Bandwidth Broker Signalling
SJ4	SuperJANET4 (backbone) network
SLA	Service Level Agreement
SLAC	Stanford Linear Accelerator Center
SLD	Service Level Definition
SLS	Service Level Specification
SMVCN	Scottish MANs VideoConferencing Network
SNMP	Simple Network Management Protocol
SONET	Synchronous Optical NETwork
TB/day	Terabyte per day
TCP	Transport Control Protocol
TERENA	Trans-European Research and Education Networking Association
TF-NGN	Task Force – Next Generation Networking (under auspices of TERENA, <i>q.v.</i>)
ToS	Type of Service
TTL	Time To Live

UDP	User Datagram Protocol
UKERNA	United Kingdom Education & Research Networking Association
VC	Videoconference
VCC	Virtual Channel Connection
VIP Demo	Videoconference IP Demonstrator
VLL	Virtual Leased Line
VoD	Video on Demand
VoIP	Voice over IP
VPC	Virtual Path Connection
VPN	Virtual Private Network
WAP	Wireless Application Protocol
WRED	Weighted Random Early Detect
WRQ	Weighted Fair Queuing
WVN	Welsh Video Network
WWW	World Wide Web

Appendix A — Think Tank Membership

<i>Name</i>	<i>Organisation</i>
Jane Butler	Cisco Systems Ltd
Tim Chown	University of Southampton
Chris Cooper	Rutherford Appleton Laboratory, CLRC
Jonathan Couzens	NOSC, ULCC
Chris Edwards	Lancaster University
Paul Jeffreys	JCN / Rutherford Appleton Laboratory, CLRC / Oxford University
Dave Price	University of Wales, Aberystwyth
Rina Samani	UKERNA
Jeremy Sharp	UKERNA
Robin Tasker	Daresbury Laboratory, CLRC

Appendix B — Acknowledgements

The Think Tank wishes to acknowledge with thanks the many helpful contributions it has received from members of the community through consultation, discussion, and comment.

<i>Name</i>	<i>Organisation</i>
Mike Atkinson	Stockton and Billingham College
Hugh Beedie	University of Wales, Cardiff
Bill Byers	EUCS, University of Edinburgh
Colin Cooper	ClydeNet Technical Advisory Group, University of Glasgow
Geoff Constable	University of Wales, Aberystwyth
Steve Cundy	Beverley College
Jon Crowcroft	UCL
Gavin Dykes	NESCOT College
Ronnie Gibb	University of Glasgow
Moira Grainger	FaTMAN, University of St Andrews
Huw Gulliver	University of Wales, Cardiff
George Howat	EaStMAN, EUCS, University of Edinburgh
Peter Kirstein	UCL
John Linn	AbMAN, University of Aberdeen
Paul Matthews	University of Wales, Swansea
David Morgan	University of Wales, Glamorgan
Graham Mort	NILTA and Strodes College
Pat Myers	Network North West
Tony Ollier	University of Wales, Swansea
Nigel Parkinson	Tameside College
John Scott	Dudley College
David Stedham	University of Wales, Bangor
Rob Symberlist	University of Wales, Swansea
Jean Ritchie	SHEFC Communication & Information Technology Programme
Tim Robinson	Network North West
Graham Waite	Regional Support Centre Technical Adviser
Mike Whitehead	University of Dundee
Tom Wiersma	University of Wales, Cardiff
Steve Williams	University of Wales, Swansea

Annex — Terms of Reference

Terms of Reference for a Quality of Service Think Tank

UKERNA
November 2000

Contents

1	Introduction	44
2	Background	44
3	QoS and SuperJANET	45
4	Think Tank Deliverables	45
5	Think Tank Membership	45
6	Timescales	46

1. Introduction

The challenge of large-scale network quality of service support on SuperJANET4 is formidable. The first step towards tackling this challenge is to harness the expertise available within the academic community and from the supplier of the switching equipment used within SuperJANET4 to create a roadmap for the development of network QoS. To this end, this document specifies the Terms of Reference for a network QoS Think Tank whose job will be to produce the development roadmap.

2. Background

The following trends are apparent on today's Internet:

- Many new Internet-enabled applications are emerging, for example pervasive videoconferencing and video streaming services, virtual networked environments etc.
- Convergence of other networks such as telephone, radio, television with the Internet is underway.
- Network traffic is increasing as the number of Internet users and applications increases. Capacity upgrade needs to be complimented by introduction of complimentary network traffic engineering techniques, to improve utilisation, and to tailor services to user needs.
- There is increasing diversity among user application needs, and the network needs to be engineered beyond a single level of service in order to meet this greater range of requirements.

In order to keep pace with these developments increasing Internet bandwidth on its own is not going to be sufficient, but clearly it must happen to avoid traffic congestion. This effectively means making the transition from single service and best effort service networks that exist now into multi-service networks which provides different levels of service.

IP design inherently provides a 'best effort' service which is subject to unpredictable delays and data loss. As stated above, congestion can be avoided by increasing bandwidth, however many new Internet applications are multimedia and not only require a lot of bandwidth, but also have strict timing requirements requiring more than simple best efforts.

Traffic Engineering in this context is a term used to describe how different types of traffic are managed over an IP network. Traffic engineering allows varying levels of network resource to be allocated to different applications based on their performance requirements or by customer allocation i.e. they have paid for a high priority service. There are two fundamental elements of traffic management:

1. The Quality of Service (QoS) algorithms, protocols and policies that enable differentiation of IP packets and the application of appropriate behaviour on a particular traffic flow. For example techniques exist in IP routing technology to mark packets; place packets into queues that are given weighted priority; signal the network to reserve network capacity for end-to-end QoS; match traffic to available network capacity (traffic shaping); and perform network congestion anticipation.
2. Use of service engineering techniques such as caching, multicasting, Multiprotocol Label Switching (MPLS) to reduce, redirect and balance network traffic.

Ultimately traffic engineering should provide a reliable end-to-end Quality of Service to the user. End-to-end however means providing service delivery guarantees that span multiple network management domains and in the context of SuperJANET4 the management domains are campus LANs, MANs, SuperJANET backbone. In terms of implementing end-to-end QoS, it not only means that standard QoS technologies must be deployed, but that service level agreements and operational standards must be in place across the multiple management domains.

3. QoS and SuperJANET

It is important that SuperJANET provides a platform on which a range of existing and future network applications can run to specified service levels and to do this the following areas must be developed and implemented:

- Implementation of QoS technologies to control traffic behaviour
- Manage traffic effectively using techniques such as caching and multicasting
- Implement a policy framework which the QoS technologies carry out, allowing an end-to-end predictable service

Each of these areas must be implemented to current standards and track emerging standards to assist peering across multiple network domains.

4. Think Tank Deliverables

In order to define the foundations on which each of these areas can be developed, a think-tank focusing on requirement for QoS within the JANET network is proposed. The requirement for this think tank is driven as much by a perceived user need for end-to-end (host to host or host to server) QoS as by the increasing demands of existing and developing network applications.

There are three deliverables required from the think-tank and these are:

1. Assess the requirement for QoS on JANET
2. Develop a policy framework on which to implement QoS on JANET
3. Provide initial recommendations for the technical mechanisms by which to implement QoS on JANET

As these deliverables are developed a process of wider consultation will need to take place. This consultation will be with representative experts from the following areas:

- End users from a broad cross-section of the JANET community.
- MAN and institution service managers.
- Experts in technology and JANET policy.

5. Think Tank Membership

The think tank will have a membership of:

- Representatives from the sector actively involved in QoS research (Chris Cooper, RAL, Tim Chown, University of Southampton and Chris Edwards, University of Lancaster)
- A representative from an application that has QoS requirements (e.g. videoconferencing) (David Price, University of Wales, Aberystwyth)
- A representative from the JANET Network Operations Centre (NOSC) (Jonathan Couzens)
- A representative from the research GRID community (Robin Tasker, Daresbury Labs)
- A representative from the of the switching equipment used within SuperJANET4 (Cisco Systems Ltd) (Jane Butler)
- Representatives from UKERNA (Jeremy Sharp and Rina Samani)

The group will be supported by a UKERNA project manager who will be responsible for producing the three deliverables.

6. Timescales

It is expected that the think-tank will exist for a period of not more than 4 months and in that time it will meet three times: once to brainstorm each area of work; a second time to review progress and deliverables; and a third time to approve the deliverables. It is expected that the group will report by the end of April 2001. The meeting timings are as follows:

- Meeting 1: December 2000 / January 2001
- Meeting 2: End February 2001
- Meeting 3: Late April 2001