

EARNEST Report on Technical Issues

Editor: Kevin Meynell (TERENA)

Contributors: Luca Deri (University of Pisa), Sergi Figuerola (i2CAT), Licia Florio (TERENA), Alexander Gall (SWITCH), Gigi Karmous-Edwards (MCNC), Simon Leinen (SWITCH), Athanassios Liakopoulos (GRNET), Diego Lopez (RedIRIS), Catalin Meirosu (TERENA), Kevin Meynell (TERENA), Milan Sova (CESNET), Stig Venaas (Uninett), and Klaas Wierenga (SURFnet)

Contents

<u>Executive Summary</u>	4
<u>Introduction</u>	7
<u>Basis for Study</u>	7
<u>Methodology</u>	8
<u>1. Transmission Technologies</u>	10
1.1 <u>Fibre Types and Capabilities</u>	11
<u>DWDM and CWDM</u>	12
<u>Fibre Performance</u>	12
<u>Fibre Types</u>	13
<u>Fibre Reach</u>	15
1.2 <u>Transmission Technologies</u>	16
<u>SDH/SONET</u>	16
<u>G.709 OTN</u>	19
<u>Ethernet</u>	21
<u>IP over DWDM</u>	26
1.3 <u>Transmission Equipment</u>	27
<u>Terminal Multiplexers and OADMs</u>	27
<u>ROADMs</u>	28
<u>Optical Switches</u>	28
<u>Tunable Optical Interfaces</u>	29
<u>PONs</u>	30
<u>Modulation Formats</u>	30
<u>Planning and Interoperability</u>	30
<u>Routers</u>	31
<u>Power Consumption</u>	31
1.4 <u>Conclusions</u>	32
<u>2. Control Plane and Routing Technologies</u>	35
2.1 <u>Routing Scalability</u>	35
<u>G. Huston, More ROAP: Routing and Addressing at IETF68, IETF Journal, Vol. 3 Issue 1 (May 2007)</u>	36
2.2 <u>IPv6 trends and developments</u>	39
2.3 <u>IP Multicasting</u>	41
<u>Source-Specific Multicast</u>	42
<u>Bi-directional PIM</u>	42
<u>IPv6 Multicast</u>	43
<u>Automatic Multicast Tunnelling</u>	44
<u>Management Tools</u>	44
2.4 <u>MPLS</u>	45

2.5 T-MPLS.....	49
2.6 ASON and GMPLS.....	49
<i>ASON</i>	50
<i>GMPLS</i>	50
2.7 Control Planes for Grids.....	52
<i>Management Plane</i>	52
<i>Grid Middleware</i>	52
2.8 Conclusions.....	53
3. Network Virtualisation.....	56
3.1 UCLP.....	58
3.2 Overview of Network Virtualisation Projects.....	60
<i>PlanetLab</i>	60
<i>GENI</i>	61
<i>OneLab</i>	61
<i>FEDERICA</i>	61
<i>MANTICORE</i>	61
3.3 Conclusions.....	62
4. Operations and Performance.....	64
4.1 Quality of Service, Overprovisioning, or something in between?.....	64
<i>Environments that enable adequate provisioning</i>	65
<i>Traffic Engineering</i>	66
<i>QoS as a DoS protection mechanism</i>	66
<i>Application-based differentiation, DPI, and network neutrality</i>	67
<i>Lightpath Services</i>	67
4.2 IP Management Issues.....	67
<i>Network Monitoring</i>	68
<i>Network Configuration</i>	68
<i>Management by Box</i>	69
4.3 IP Monitoring.....	71
4.4 Lower-Layer Monitoring.....	73
<i>Optical Transmission</i>	74
<i>OPUk Paths and FEC</i>	75
<i>SDH/SONET</i>	75
<i>GFP</i>	75
<i>GE and 10 GE</i>	75
4.5 PERT.....	76
4.6 Conclusions.....	80
5. Middleware.....	83
5.1 Identity Management.....	84
5.2 Identity Federations.....	85
<i>Basic Elements of Identity Federations</i>	87
<i>Identity Federation Trust Framework</i>	88
<i>Transitive Trust</i>	89
5.3 User-centric Identity Management.....	90
5.4 Abstract Identity Framework.....	91
5.5 Federated Network Access.....	92
<i>eduroam</i>	92
<i>Network Endpoint Assessment</i>	94
5.6 Middleware for Applications.....	95
5.7 Grid Middleware.....	95

<u>5.8 Middleware Diagnostics.....</u>	<u>96</u>
<u>5.9 Conclusions.....</u>	<u>97</u>
<u>Abbreviations.....</u>	<u>99</u>

Executive Summary

The EARNEST Technical Sub-Study focuses on the technologies that are currently used to build research and education networks, identifies the problems associated with them, and considers how technological developments may change how these networks are provisioned in the next five years and beyond. The study was split into four main areas of investigation; namely transmission technologies, control plane technologies, operation and performance issues, and middleware. In particular, it examines technologies that are likely to be suitable for National Research and Education Networks (NRENs), although it also takes into account research and education networking at the regional, metropolitan and campus level.

The findings of this study are based on a series of individual meetings with leading equipment suppliers and research institutions, as well as information obtained from other sources such as technological briefings and research papers. These organisations were identified and interviewed with the assistance of a panel of technical experts drawn from the research and education networking community.

Research and education networks increasingly have access to dark fibre, and in some cases are now managing all aspects of transmission themselves. This means that issues such as fibre quality and the capabilities of transmission equipment have become much more relevant than before. Nevertheless, most research and education networks will likely be limited to existing fibre installations for the foreseeable future.

The fastest transmission equipment is currently able to support 40 Gbps using SDH/SONET (OC-768) over a limited number of wavelengths, but this is prohibitively expensive for most research and education networks. Although prices for SDH/SONET interfaces are soon expected to drop significantly, it would seem that next-generation Ethernet will become the transmission technology of choice for those networks without legacy telecoms issues.

Most vendors appear to be focusing on 40 and 100 Gigabit Ethernet (GE) for next-generation transmission systems, and are adding carrier-class features such as OAM&P and virtual circuit functionality. It is anticipated that the first implementations of 100 GE will arrive around 2010 or 2011, with 40 GE perhaps providing an alternative to OC-768 by 2009. For both 40 and 100 GE, improved modulation techniques are expected to eventually allow these technologies to support long-haul links.

Most vendors appear to see a limited requirement for provisioning more than ~80 wavelengths over a single fibre. 50 GHz channel spacing appears to provide a good trade-off between faster line rates and longer reaches, although this is likely to be of limited concern to research and education networks who currently only utilise a fraction of the potential capacity of their fibres. However, the availability of ROADMs and wavelength-selectable interfaces on switches and routers, will make WDM systems simpler and cheaper to provision and operate.

Research and education networks will continue to heavily focus on IP services, although the trend towards hybrid networks is likely to continue as networks increasingly move towards operating the underlying optical infrastructure. At the

present time, the IP and optical domains largely have to be managed separately, but the introduction of GMPLS and virtualisation frameworks such as UCLP promise a more integrated approach. It is possible that production IP services may eventually become just one of many other services provisioned over a dynamic lightpath infrastructure.

With respect to IP services, there are some concerns over the future scalability of the routing system. Although these problems are not immediately imminent, the IAB and IETF recently started to investigate whether addressing and routing could be made more efficient so that it becomes less reliant on hardware developments. In addition, there are revised predictions that IPv4 address space may be exhausted within the next 5 years¹, which is somewhat earlier than expected. IPv6 has already been widely adopted by NRENs who should be well-placed to support the transition from IPv4, but campuses should start developing migration strategies if they have not already done so.

It is likely that most NRENs and international backbones will continue to rely on overprovisioning to ensure reliable performance. There is currently limited demand for premium services, and there is little value in devising complex bandwidth allocation models where additional links can be established over dark fibre at marginal cost. Where customers or users have very demanding requirements, lightpaths can be used to provision dedicated private networks, and this should become a more dynamic process as better control plane and virtualisation frameworks are developed. However, recognising that some edge networks may still have bandwidth limitations, QoS transparency should be supported in core IP networks in order to allow QoS and other traffic engineering mechanisms to be deployed over them.

Unfortunately, hybrid networks complicate network management as the IP and optical layers have evolved somewhat separately, and therefore have different management protocols, tools, and operational procedures. There are a number of initiatives in the research and education community to develop tools for monitoring and managing optical networks, but these are at quite an early stage of development and offer limited integration with the IP layer. In addition, more comprehensive network monitoring is needed for understanding and managing networks, although it remains hampered by a lack of standards and is challenging to undertake at line rates above 5 Gbps.

Another important issue for network management is that operational experience shows that around 90% of all reported problems in modern networks are due to issues at end-sites, many of which are attributable to so-called middle boxes such as firewalls, NATs, rate-shapers and intrusion detection devices. These devices that are supposed to help manage and secure the network, can often end-up making things more complicated and less secure, as well as hampering performance and delaying network upgrades. Consideration therefore needs to be given to better placement and management of such devices.

The GN2 PERT has successfully traced many of the problems that afflict users of GÉANT and other connected networks; in most cases demonstrating the problems are

¹ <http://www.potaroo.net/tools/ipv4/index.html>

at the end-sites. Unfortunately, it currently has a limited scope, and if the concept of the PERT is to evolve, it needs to be extended to NRENs, and possibly to regional and/or campus levels as well. This should be accompanied by the establishment of standard operating procedures, and common informational and tracking systems.

In the middleware area, identity federations are becoming increasingly important for handling and supporting user access to remote services. The majority of the NRENs in Europe either already have a federation or are in the process of establishing one. Those without a federation should plan to have one in place within the next couple of years.

As NRENs are the natural candidates for providing technical and organisational coordination for research and educational communities, they should support multiple trust infrastructures in order to be able to handle different AAI for different purposes. In addition, as identity federation implementations mature, they should consider how to support new features such as integrating Shibboleth with Grid applications, and facilitating inter-federation peering.

It is expected that SAML 2.0 will become the standard mechanism for exchanging identity assertions for web-based applications. Other services will probably continue to use X.509 personal certificates for now, at least until alternative approaches to integrate these are more fully developed. Authorisation decisions will likely be supported by schemas such as eduPerson and SCHAC, although there is still no well-established standard for communicating identity data to applications. This is an area in which NRENs might therefore be proactive.

Other AAI models such as user-centric identity management or the abstract identity framework may also yet prevail beyond the academic community, and NRENs should continue to monitor these developments. However, users might still be able to use their federated identities with services that support these models.

Introduction

This report forms part of the EARNEST study (an activity of the GN2 project) into the evolution of research and education networking in Europe in the next five years. It aims to provide policy inputs on initiatives that could help keep European research and education networking at the forefront of worldwide development, whilst also helping plan the development of research and education infrastructure and services at European, intercontinental, national and local level.

The Technical Sub-Study focuses on those technologies that are currently used to build research and education networks, identifies the problems associated with them, and considers how technological developments may change how these networks are provisioned in the future. It considers technologies that are just becoming available, but are largely untried and untested, as well as emerging technologies that will only become available in a longer timeframe.

Basis for Study

This study was split into four main areas of investigation; namely transmission technologies, control plane technologies, operation and performance issues, and middleware. This makes the study somewhat broader than the previous SERENATE technical study which largely focused on transmission, routing and switching issues, but it has become clear in the interim that operations/performance and middleware have become increasingly important considerations for running networks.

The investigation into transmission technologies considers the development of networks beyond 10 Gbps, how they will be provisioned, and whether they are likely to be affordable for research and education networks. Although 40 Gbps networks are already available, they are currently very expensive to provision, which has limited uptake to all but the most demanding commercial customers. At the same time, research and education networks have increasing access to dark fibre which provides them with more flexibility in how they build their networks.

The investigation into routing and control plane technologies considers the intelligence that establishes, maintains, and tears-down the different connections within networks. This study considers both optical switching mechanisms and the more traditional IP routing mechanisms used by research and education networks.

The investigation into operations and performance issues considers how to manage and monitor networks in order to better improve usage. It is not always possible to use existing networks to their fullest extent due to inefficiencies within certain protocols, and unidentified problems at certain locations in networks. Equally, network operators need to ensure they are delivering requested service levels, which becomes more complex when they have optical layers to manage as well.

The investigation into middleware considers the services that will help applications take advantage of network resources. As network use has increased, authentication and authorisation have become important issues, complicated by the fact that users have also become increasingly mobile and wish to use resources as they move around.

Allied to this is the increasing sophistication of hackers, which makes ever more secure methods of authentication necessary.

Last, but not least, network virtualisation is discussed in a separate section as it brings together aspects of the other areas of study. It considers the increasing interest in designating specific parts of physical networks as virtual resources, in order to allow institutions and users to build their own networks for specific tasks such as disruptive testing.

It should be noted that the primary target of this study is to investigate technologies that are applicable to National Research and Education Networks (NRENs), although it attempts to address research and education networks at other levels as well. This includes regional, metropolitan and campus networks, as well as international backbones. Reference is also made to technologies and developments in the telecommunications and ISP sectors where these are of relevance and importance to research and education networking, although it should be borne in mind those sectors often have differing requirements.

Methodology

This study was undertaken by TERENA with assistance from a panel of technical experts drawn from the research and education networking community, and chosen for their specific expertise in the areas of investigation. Their role was to identify the important emerging technologies, advise on which vendors and other organisations should be invited to participate in the study, and to assist with vendor interviews.

The technical panel membership was comprised as follows:

Transmission Technologies: Lars Fischer (NORDUnet) John Graham (Indiana University), Otto Kreiter (DANTE)

Control Plane Technologies: Alexander Gall (SWITCH), Gigi Karmous-Edwards (MCNC), Stig Venaas (Uninett)

Operations & Performance Issues: Luca Deri (University of Pisa), Simon Leinen (SWITCH), Dimitra Simeonidou (University of Essex)

Middleware: Diego Lopez (RedIRIS), Milan Sova (CESNET), Klaas Wierenga (SURFnet)

The findings of this study are based on a series of individual meetings with leading equipment suppliers and research institutions, as well as information obtained from other sources such as technological briefings and research papers. The following organisations participated in these meetings, which were held between January and June 2007:

Alcatel-Lucent

Calient

Cisco

Danish Technical University – COM

DANTE

Extreme Networks

Force10

IBM
Juniper
Liberty Alliance
MERLIN Radio Astronomy Project
Nortel
Sun Microsystems
SxIP

Key personnel from each of these organisations were interviewed about the technologies they were currently working-on, and to provide an opportunity for them to present their visions about future developments in networking. The technical panel formulated some key questions that were sent in advance of each meeting to provide guidance for the discussions, although it should be noted that not all questions were relevant to all suppliers.

Non-disclosure agreements were also signed with some of the companies, and as a consequence, certain technical details are described in a non-attributable fashion.

1. Transmission Technologies

The past five years have seen a paradigm shift in the way research and education networks are provided. Such networks have traditionally run a best-effort IP service over circuits leased from telecommunications carriers (more recently based on SDH/SONET), which has limited the flexibility as to how networks were built, and has made optimisation between the different transmission layers difficult. However, it has become increasingly possible for research and education networks to lease or even install their own dark fibre which allows them to implement their own transmission systems; in particular, equipment that takes advantage of WDM techniques. Among other things, this offers the flexibility to configure network topologies on demand, and upgrade capacity as necessary without having to renegotiate prices.

At the same time though, access to dark fibre and/or wavelengths means that research and education networks become responsible for optical switching and often multiplexing and amplification as well. Therefore, issues such as fibre type and reach, location of amplification and regeneration sites, chassis densities, interface counts, and underlying transmission protocols are becoming far more critical than before. More importantly, transmission systems are complex to engineer and manage, which means that new developments aimed at simplifying and reducing the cost of such systems are of great interest to research and education networks.

Having access to the transmission layer provides flexibility in the choice of underlying transmission protocol. SDH/SONET (and to a certain extent PDH) is generally employed by telecommunications carriers, but it was designed to support voice traffic and is relatively inefficient for packet-switched data (e.g. IP), not to mention complex and expensive compared with Ethernet. By contrast, Ethernet was not especially designed with resilience or fault management in mind, although given the fact that data networks are usually based on a mesh rather than a ring topology, this may be less of a concern. Better fault management features are currently being developed for Ethernet, but equally there are SDH/SONET developments to improve its handling of packet-switched data. Nevertheless, as research and education networks primarily run IP-based services and generally have fewer legacy issues (namely not having to support traditional telephony), they should have a freer hand to choose the most cost-effective solution,

There is also the requirement for ever-more bandwidth, as statistics show that Internet traffic is currently roughly doubling every year². There are commercially-available DWDM systems that can support up to 160 channels over a single fibre, but most offer something in the order of 40-80 channels, and this is dependent on other factors such as actual line rates, the type of fibre being utilised, and how often amplification and regeneration is undertaken. In addition, CDWM systems offering up to 18 channels have become increasingly popular, as these are simpler and offer significant cost advantages for shorter-haul applications,

However, provisioning additional links is not always the complete answer as some demanding users such as those in the high-energy physics, radio astronomy and digital cinema communities are capable of generating individual traffic flows that

² <http://www.ripe.net/ripe/meetings/ripe-54/presentations/100GE.pdf>

exceed 10 Gbps. Although it is possible to aggregate several separate links to create a larger virtual channel, this adds complexity and cost in terms of the number of interfaces required, and requires additional processing power in the routers/switches that handle the necessary load balancing, thus limiting how many channels can be aggregated (typically 4 to 8). Furthermore, it can add latency and reduce the efficiency of large flows.

As a result, there is a requirement for line rates to increase, as well as the number of channels that can be transmitted over a single fibre. The fastest commercially-available transmission equipment can operate at line rates of 40 Gbps, although interfaces are still prohibitively expensive and fewer WDM channels can be supported compared with 10 Gbps solutions. Nevertheless, vendors are currently working on cheaper 40 Gbps solutions, as well as looking further ahead to even faster line rates that can operate over extended ranges.

1.1 Fibre Types and Capabilities

Fibre optic cable provides the basis of modern communication systems, by efficiently guiding light signals from one location to another through a transparent core of very pure glass (although plastic can also be used for low bit rate, short reach connections). A signal is inserted into the cable using an LED or semi-conducting lasers, which can then be processed by a fibre optic receiver at the other end of the cable. LEDs can be used for low bandwidth connections (up to 622 Mbps) over short distances (up to 2 km), but lasers are required for transmitting over longer distances at higher bit rates. Lasers are also able to emit light pulses at particular wavelengths (also known as lambdas), which allows multiple signals to be simultaneously transmitted over the same cable (known as wavelength-division multiplexing or WDM).

The capacity of fibres is very high as current state-of-the-art transmission equipment can support 160 lambdas at 10 Gbps, which corresponds to 1.6 Tbps. However, laboratory tests have yielded maximum bandwidths in the order of 10 Tbps, so in principle the limitations are currently imposed by the capabilities of the available transmission equipment. In any case, since much of the cost of installation is attributable to groundwork, it is usual for cables to contain between 32 and 224 fibres each, and it is often simpler and cheaper to spread capacity over more fibres rather than exploit the full capacity of a single fibre.

The laser transmitters used in fibre systems produce light in the near infrared range, utilising spectral bands that are optimum for optical transmission. These have been defined by the ITU-T as follows:

Band	Range (nanometres)
O (Original)	1260 to 1360
E (Extended)	1360 to 1460
S (Short Wavelength)	1460 to 1530
C (Conventional)	1530 to 1565
L (Long Wavelength)	1565 to 1625
U (Ultra-long Wavelength)	1625 to 1675

DWDM and CWDM

There are two main types of WDM system in use today:

Coarse Wavelength Division Multiplexing (or CDWM) offers 18 wavelengths in the 1270 to 1610 nm range using a 20 nm spacing (as specified by G.694.2). This allows the use of laser transmitters that have a high spectral and/or thermal drift, although the drawback is that these wavelengths cannot easily be amplified which limits a CWDM span to around 60 km. In addition, many wavelengths below 1470 nm are considered unusable on older specification fibre due to levels of attenuation in the 1310 to 1470 nm region. Nevertheless, CDWM is well suited for local and metropolitan area networks as relatively cheap transceiver equipment can be used.

Dense Wavelength Division Multiplexing (or DWDM) offers between 40 and 160 wavelengths in the C- and L-bands, depending which wavelength spacing is used (as specified by G.694.1). Using the 1530 to 1625 nm range makes amplification easier, although the narrow channel spacing means that stable wavelengths must be maintained which requires precision temperature control of the laser transmitters.

DWDM systems typically utilise 100 GHz (0.8 nm) spacing which offers 40 channels, although 50 GHz (0.4 nm) spacing offering 80 channels is becoming more common. Some state-of-the-art systems are even using 25 GHz (0.2 nm) spacing offering 160 channels, which is sometimes referred to as Ultra-Dense Wavelength Division Multiplexing (UDWDM),

The ITU-T has also defined a 12.5 GHz (0.1 nm) spacing grid which has the potential to offer 320 channels, although such narrow spacing requires extremely complex transceiver equipment and becomes problematic at high bit rates. Most manufacturers have therefore focused their efforts on providing faster bit rates in the 50 GHz region.

The high capacity and amplification capabilities of DWDM mean that it is usually used for long-haul applications, although it has metropolitan applications as well. It also has the advantage of allowing service providers to incrementally upgrade the transmission capacity of their network using existing installed fibre, albeit that more complex and expensive transceiver equipment is required.

Fibre Performance

There are a number of problems associated with fibre that limit its performance, depending on cable types and transmission rates. The main limitations are *attenuation, chromatic dispersion, polarisation mode dispersion, cross-phase modulation, and four-wave mixing*.

Attenuation is a loss of signal (expressed in terms of decibels per kilometre) caused by absorption and scattering of the light as it travels through the core. This weakens the signal which puts limitations on how far it can propagate before it needs amplification or regeneration. A particular problem with older fibre is water molecules that were inadvertently incorporated into the core material during manufacture, the effect of which is to radically increase attenuation around the 1380 nm wavelength (known as the *water peak effect*).

Erbium-Doped Fibre Amplifiers (or EDFAs) are able to undertake the amplification of an entire optical signal without the need for any optical-to-electrical conversion, and with minimal power consumption. If necessary, Raman amplifiers can also be used in conjunction with EDFAs to boost the signal over longer spans. However, whilst EDFAs can simultaneously amplify multiple wavelengths, they can only do so in the C and L-bands where the energy levels of the photons are close to those of the erbium ions in the amplifier. EDFAs unfortunately also amplify any signal degradation as well, whilst introducing a degree of noise themselves. It is therefore still necessary to regenerate the signal at periodic intervals which depends on cable type and transmission rates.

In principle, Praseodymium-Doped Fibre Amplifiers (or PDFAs) could be used for the O-band, and Thulium-Doped Fibre Amplifiers (or TDFAs) for the S-band, These have been demonstrated, but since DWDM development has largely focused on the C-band to-date, there has been little commercial incentive to exploit these further.

Chromatic Dispersion is caused by the tendency of different wavelengths to propagate at different speeds through a medium (expressed in terms of picoseconds per nanometre per kilometre), which can distort the signal at the receiving end and cause errors. This can be minimised by using wavelengths around the so-called zero dispersion point which is found at 1310 nm on older fibre, but this is obviously unsuitable for the operation of EDFAs. If amplification is required, it is either necessary to accept degraded performance, or to use fibre where the zero dispersion point is shifted to 1550 nm.

Polarisation Mode Dispersion (or PMD) is where imperfections in the core cause different polarisations of the same wavelength to travel at different speeds (expressed in terms of picoseconds per kilometre). Since PMD control was introduced into the fibre manufacture and installation process (from 1992 onwards), it generally only starts to become a problem at higher bit rates (typically 5 Gbps and above), but unlike attenuation and chromatic dispersion, its effects are random and unpredictable. It is therefore necessary to either employ complex PMD compensation techniques, or utilise polarising-maintaining fibre which suffers from higher attenuation and therefore has a shorter reach.

Cross-Phase Modulation (or XPM) affects WDM systems, particularly where long-haul transmission is involved, and is where the carrier frequencies of different channels interfere with each other. It can be mitigated by selecting unequal bit rates on adjacent channels.

Finally, Four-Wave Mixing (or FWM) also affects WDM systems, and usually becomes pronounced as wavelength spacing decreases, where signal levels are high, or where chromatic dispersion is low. It can induce a crosstalk effect between channels, but can be mitigated using uneven channel spacing or by actually *introducing* a limited degree of chromatic dispersion.

Fibre Types

Fibre is categorised as either *multi-mode* or *single-mode*, although there are different types within each of these categories.

Multi-mode fibre has a relatively large diameter core and is really only suitable for local area networks as it is limited to reaches of approximately 2 km at 100 Mbps, 500 m at 1 Gbps, and 300 m at 10 Gbps. It also tends to be more expensive than single-mode fibre these days, although it has the advantage that it is easier to install in buildings, and that much cheaper VCSEL or LED interfaces can be used.

The ITU-T standard for multimode fibre is G.651 which is optimised for use in the 1310 nm region, although it can also operate in the 850 nm region. The ISO 11801 standard provides a further classification, with OM1 and OM2 supporting up to 1 Gbps, and the newer laser-optimised OM3 supporting up to 10 Gbps.

Single-mode fibre utilises a much smaller diameter core that forces light to follow a more linear path that minimises reflection. It is therefore capable of supporting high bit rates over long distances, although it is increasingly used in local area networks as well, especially as higher bandwidths are utilised.

There are several ITU-T standards for single-mode fibre, the earliest of which was G.652 (or Non-Dispersion Shifted). Most single-mode fibre installed between 1985 and 1996 will conform to this standard, and is optimised for the 1310 nm region where chromatic dispersion is minimised. Unfortunately, this is not optimal for DWDM systems that must operate in the C and L-bands, and the water peak effect also restricts the wavelengths available to CDWM systems.

As result, G.652.C and G.652.D (or Low Water Peak Non-Dispersion Shifted) fibre was introduced from 2000 onwards to offer lower attenuation around the 1380 nm region, thus allowing transmission on a wider range of wavelengths. However, chromatic dispersion remains high in the C-band (around 17 ps/nm-km), so whilst such fibre is capable of supporting local and metropolitan area networks that do not require amplification, it is less suitable for DWDM applications.

G.653 (or Dispersion Shifted) fibre was also extensively deployed in some regions of the world such as Latin America the Middle East, as well as in submarine environments. This was originally developed to shift the region of minimal chromatic dispersion to a band between 1500 and 1600 nm where the longer wavelengths would offer improved attenuation. It transpired though, that whilst G.653 fibre worked well with a single 1550 nm wavelength, it was seriously affected by FWM issues when multiple wavelengths were run across it. Although DWDM solutions that support G.653 have recently become available, the difficulties of using this fibre led to the development of the G.655 standard.

G.655 (or Non-Zero Dispersion Shifted) fibre was introduced around 1996, and most commercial carriers will have installed this in their newer routes. It is optimised for use with long-haul DWDM systems as it moves the zero-dispersion wavelength to just before or after the C-band (depending on whether the fibre is NZD+ or NZD-). This introduces a limited amount of chromatic dispersion to the main transmission window (typically 4.5 ps/nm-km at 1550 nm), but this is actually advantageous as it minimises the effects of XPM and FWD.

A further type of fibre was introduced in 2004. G.656 (or Non-Zero Dispersion for Wideband Optical Transport) fibre has low chromatic dispersion (typically 2 to 14 ps/nm-km) in the S, C and L-bands which supports CDWM systems as well as the addition of a further 40 channels in DWDM systems. It is likely to increasingly become available in fibre routes in future.

Finally, there is also a G.654 (Loss Minimised) fibre standard, but this is a specialist fibre for submarine applications. It is optimised for high power operation in the 1500 to 1600 nm region, but is not suitable for WDM. It is unlikely that research and education networks will encounter this type of fibre.

The following table summaries the suitability of the different types of fibre for various applications³:

ITU-T Standard	TDM (1310 nm)	TDM (1550 nm)	CWDM	DWDM
G.652	Good	OK	OK	OK
G.653	OK	Good	Poor	Poor
G.654	N/A	Good	N/A	OK
G.655	N/A	Good	OK	Good
G.656	N/A	OK	Good	OK

Fibre Reach

The type of fibre being used, splices, connectors, and equipment between the receiver and transmitter all contribute to distortion of the signal, which leads to errors and eventually unsuccessful transmission. As already mentioned, the signal can be amplified, but this also amplifies noise and other artefacts which means the signal eventually needs to be regenerated.

Although some research and education networks may be in the advantageous situation of being able to install their own fibre using the latest low-loss cable, this is likely to be only in the local or metropolitan area. In most cases, particularly for long-haul networks, they are far more likely to lease or rent dark fibre from telecommunications or fibre providers who will often have a large installed base of older fibre such as G.652. In some cases, G.655 might be obtainable as well, but in reality a route may comprise a variety of different fibre types. For these reasons, it may prove difficult to take advantage of recent advances in transmission distances, particularly where amplification and regeneration is required.

For example, the latest low-loss conventional fibre (0.15-0.16 dB/km) promises the ability to transport 80 x 40 Gbps channels up to 140 kms before amplification, and up to 2,000 kms before regeneration. Other trials have transported 256 x 10 Gbps channels up to 500 kms before amplification, and up to 11,000 kms before regeneration, or even 1 x 40 Gbps over 1 million kms. The recently introduced photonic crystal fibre (also known as hollow core fibre) can also offer very low attenuation (>0.1 dB/km) whilst minimising non-linear effects such as XPM and FWM, although this is currently extremely expensive and has a limited installed base.

³ <http://www.nordu.net/development/fiber-workshop2007/NNW03012007.pdf>

However, the problems in the telecommunication sector between 2001 and 2004 have moved focus away from extreme performance towards cheaper implementation in order to leverage investments in existing infrastructure. As a result, networks must expect to have to design their topologies around legacy infrastructures for the foreseeable future.

It is possible using commercially available transmission equipment in conjunction with standard G.652 fibre installations, to transmit at 80 x 10 Gbps, or 4 x 40 Gbps channels up to 80 kms before amplification, and up to 4,000 kms before regeneration. Using G.655 fibre it is possible to extend the distances between amplifiers to nearly 400 and 200 kms respectively.

There are likely to be incremental improvements in these performance figures as transmission equipment evolves, particularly for 40 Gbps bit rates as new techniques to counter PMD and non-linear effects are developed. Nevertheless, most new transmission systems are being developed to work with existing fibre installations.

1.2 Transmission Technologies

SDH/SONET

Synchronous Digital Hierarchy (or SDH) is an ITU-T standard (as specified in G.707 and G.708) for transmitting information over fibre optic cables. It was introduced in the mid-1980s to support both telephony and data communications, and due to its flexible time-division multiplexing abilities, it has become the predominant digital transmission standard.

Synchronous Optical Networking (or SONET) is a similar standard that is primarily used in North America, whereas SDH is used in the rest of the world. There are differences in the framing and mapping options, but both define standard *optical carrier (OC)* interfaces that permit interoperability.

The most common SDH line rates are as follows:

Optical Carrier Level	SDH Level	Line Rate (kbps)
OC-3	STM-1	155,520 (155 Mbps)
OC-12	STM-4	622,080 (622 Mbps)
OC-48	STM-16	2,488,320 (2.5 Gbps)
OC-192	STM-64	9,953,280 (10 Gbps)
OC-768	STM-256	39,813,120 (40 Gbps)

Other SDH line rates have also been defined, but these are generally not commercially available.

SDH/SONET was designed to permit interoperability between telecommunications equipment from different vendors, and is able to support high bandwidth links whilst greatly simplifying the multiplexing of channels at a variety of line rates (based on multiples of 64 kbps, the capacity of a voice channel). It is also possible to concatenate channels to create a larger payload (indicated by appending a 'c' to the

optical carrier level), although this requires all the equipment in the network to support this.

These channels can in turn encapsulate higher-level protocols such as IP (using Packet-over-SONET), Ethernet and ATM, as well as support standard telephony signals. In addition, SDH/SONET can be used in a variety of network architectures such point-to-point, point-to-multipoint and star topologies, although a common architecture is as a ring topology (in a LAPS, UPSR or BLSR configuration) that is able to provide resilience with quick failover (typically 50 ms) capabilities. This is made possible by the provision of overhead information for so-called operations, administration, maintenance and provisioning (OAM&P) that also allows for centralised management and troubleshooting of the network. In fact, one of the great strengths of SDH/SONET is that faults can be very precisely isolated,

Telecommunications providers have invested a great deal of money in SDH/SONET equipment over the past twenty years, which generally falls into the categories of *add-drop multiplexors (ADMs)*, *digital cross-connects (DXCs)*, and *regenerators* (although regenerators have largely been made obsolete by optical amplifiers as these do not need to be replaced when line-rates increase). The telco-provided leased lines that have traditionally been used to build IP-based networks, have therefore been available with SDH/SONET interfaces for the past ten years or more. As a result, most router and optical switch vendors have offered SDH/SONET at various line rates as their primary WAN interface, which means there is a large installed base of SDH/SONET-compatible equipment in both core and edge networks.

Unfortunately, there are a number of drawbacks with SDH/SONET. Although it can provision multiple channels at different line rates, in practice it has relatively coarse granularity and is not optimised for handling bursty traffic such as IP. A complete channel has to be allocated to handle data traffic even if it is not being used to full capacity, and neither is it easy to dynamically reallocate bandwidth.

This is a particular problem when trying to transport Ethernet over an SDH/SONET network, as the standard line rates (10 Mbps, 100 Mbps, 1 Gbps and 10 Gbps) do not easily map into the SDH/SONET hierarchy which this was designed with TDM voice data in mind. This either requires inefficient concatenation of SDH/SONET frames (where a 1 GE connection requires an STM-48, thus wasting 58% of the available bandwidth), or the use of virtual concatenation that adds complexity and needs to be supported by the ingress and egress nodes on the network. More specifically, whilst virtual concatenation can work where the total bandwidth available on a SDH/SONET link exceeds the Ethernet line rate, it is problematic when the Ethernet line rate that needs to be transported is greater (for example, transporting a 10 GE connection over an STM-192c). For this reason, 10GBASE WAN-PHY was developed as a variant of 10 GE that maps into an OC-192 payload using a convergence layer.

Another drawback is that the much vaunted restoration time of SDH/SONET comes at the cost of 50% of available bandwidth. In a ring configuration using two or four fibres, a fibre or pair of fibres must either lie idle until called into backup service, or be used to transport identical copies of the data in case of deterioration or loss on the other link(s). Furthermore, the separate fibre paths are often only physically separated by a short distance (less than a few metres), which is usually enough to ensure

resilience in the event of accidental damage, but insufficient when cataclysmic events occur (e.g. the collapse of the World Trade Center in New York).

In addition, whilst SDH/SONET is a proven transport mechanism, it is quite complicated to engineer and manage. It has traditionally been taken care of by telecommunications providers, but with the increasing availability of dark fibre, transmission equipment is likely to become the responsibility of research and education networks. These organisations have relatively little experience of SDH/SONET, and generally fewer staff to deal with the associated management issues. They should be aware there is less automation than for IP networks, and the tools that are available are generally expensive and complicated to use. However, modern SDH/SONET equipment can often support IP-based management, and there are ongoing initiatives such as the TL1 Toolkit that supports easier configuration of the TL1 protocol which is used to manage transmission equipment.

Finally, SDH/SONET equipment is very expensive. The need for complex timing to maintain synchronicity adds complexity and therefore cost, and there is also the requirement for add/drop multiplexers and cross-connects to establish the necessary circuits for handling bandwidth demand. For example, when comparing the current cost of a router interface for STM-64 with one for 10 GE, it is more than ten times as expensive (USD 100K compared with 10K), and this difference moves closer to fifteen times when comparing interface costs on optical switches. More pertinently though, both 1 and 10 GE have maintained the traditional cost curve where the cost per megabit is halved with each increase in line rate (usually expressed as a quadrupling of bandwidth for twice the price), and these costs continue to fall. By contrast, the cost per megabit for SDH/SONET has remained fairly static for OC-48, OC-192 and subsequently OC-768 (at around USD 12-14 per megabit).

At the present time, OC-768 is the only option for provisioning 40 Gbps links, although the interfaces are extremely expensive (approximately USD 70-100K per port). However, with vendors working on 40 and 100 GE solutions which promise even greater cost benefits, it will become increasingly difficult to justify the continued deployment of SDH/SONET.

In fact, the future of SDH/SONET is looking increasingly uncertain with seemingly all major vendors focusing on 40 and 100 GE development. Whilst faster standards such as OC-1536 (80 Gbps) and OC-3072 (160 Gbps) have been postulated, there appear to be no immediate plans by vendors to implement these. Due to the large installed base of equipment, SDH/SONET is likely to be supported for another ten years or so. However, networks are likely to progressively move towards Ethernet as the underlying native transmission protocol, especially as recent SDH/SONET developments have largely focused on better supporting Ethernet.

The main developments are the introduction of *Virtual Concatenation (VCAT)* and the *Link Capacity Adjustment Scheme (LCAS)* that can utilise the *Generic Framing Protocol (GFP)* in order to make more flexible and efficient use of SDH/SONET channels.

Virtual Concatenation (as specified in G.707) allows for the more arbitrary assembly of smaller multiplexing channels into larger channels that can accommodate payloads

of different sizes. These also need not be contiguous which means payloads can be fragmented and can utilise whatever slots happen to be available.

The Link Capacity Adjustment Scheme (as specified in G.7042) allows bandwidth to be changed on demand using dynamic virtual concatenation. It also has some limited recovery features, although it has no provision for users to signal their requirements so parameters must be pre-configured.

The Generic Framing Protocol (as specified in G.7041) provides a standard mechanism for encapsulating other protocols (such as IP and Ethernet) for transport across a SDH/SONET network. It supports two mapping mechanisms: *transparent (GFP-T)* that maps the traffic as received (including control and idle data) into GFP frames and therefore requires a SDH/SONET payload that supports the incoming line rate; or *frame-based (GFP-F)* that only maps the actual user data and allows smaller payloads to be used.

The advantage of virtual concatenation is that it only needs to be supported by the path-terminating equipment. As it uses standard SDH/SONET containers, it is transparent to intermediate nodes which means these do not require upgrading. Nevertheless, whilst this helps leverage the existing investments of telecommunications providers, it adds further complexity at the termination points. For those organisations that have access to dark fibre and fewer legacy considerations, the choice of SDH/SONET as the underlying transmission mechanism is increasingly less clear.

G.709 OTN

Optical Transport Network (or OTN) is an ITU-T standard (as specified in G.709) that can be used to seamlessly transmit both TDM (e.g. SDH/SONET) and packet-based protocols (e.g. Ethernet and IP) over a DWDM link. It is a *digital wrapper* technology that encapsulates a client payload and overhead into own payload, and then adds its own management overhead in order to transport the signal across the network. This not only allows the client signals to be preserved without the requirement for processing at intermediate nodes, but OTN also offers fault notification, signalling and remote management mechanisms via a general communications channel that can be applied to services such as Ethernet that do not currently provide their own.

OTN is based around the SDH/SONET frame structure, but takes advantage of the higher capacity of a DWDM channel to add OAM&P and *forward error correction* information. It is currently offered in three line rates:

- OTU1 has a line rate of approximately 2.7 Gbps and is designed to transport a STM-16 signal. It can also aggregate sub-2.5 Gbps services (e.g. STM-1) with add/drop capabilities to provide compatibility with legacy transmission equipment
- OTU2 has a line rate of approximately 10.7 Gbps and is designed to transport a STM-64 or 10 GE WAN-PHY signal. However, it can also be modified to support an 10 GE LAN-PHY at full line rate (i.e. 10 Gbps) by using over-clocking techniques.

- OTU3 has a line rate of approximately 43 Gbps and is designed to transport a STM-256 signal. Presumably it may also be able to support 40 GE in future.

The ITU-T is currently working to define OTU4 which will have a line rate of approximately 130 Gbps. Given the uncertainty over SDH/SONET standards faster than OC-768, and the fact that an OTU4 could not carry a STM-1024 (160 Gbps) anyway, this is likely being specified with 100/120 GE in mind.

Perhaps one of the most useful features of OTN is the ability to include Forward Error Correction (FEC) codes. This is where redundant data is added to the overhead in order to help recipients detect errors introduced during transmission process, whilst often allowing these errors to be corrected without the need for retransmission of data. As both optical fibres and transmission equipment can introduce errors, FEC therefore effectively extends the distance an optical signal can travel before it requires regeneration, as well as be used to detect deteriorating links.

A common FEC encoding is Reed-Solomon (as specified in G.975) which adds around a 6.5% overhead and the equivalent of a 5-6 dB gain in attenuation (effectively increasing transmission distance by 25-30 kms). However, other so-called *strong* (and often proprietary) FEC encodings can be used to increase effective gain by up to 10 dB (50-65 kms), although at the cost of a bigger overhead, not to mention processing power. Indeed, as transmission speeds become higher (particularly beyond 10 Gbps), more effective FEC encoding is necessary to counter the increased signal degradation that occurs due to non-linear dispersion in the fibre.

Another advantage of OTN is that unlike SDH/SONET it is isosynchronous and so does not require a clock hierarchy. OTN can operate with a lower timing accuracy (20 ppm) which enables a transmitter to take its timing from a receiver, thus reducing management complexity and the possibility of timing loops.

Although the OTN was standardised in 2003, take-up was initially slow as carriers were reluctant to upgrade or replace their legacy transmission equipment to support a new framing mechanism. However, the last couple of years have seen increasing deployment of optical switches and ROADMs with OTN interfaces due to an increasing need for bandwidth, more efficient handling of non-SDH/SONET traffic, and the benefits of improved reach that come from the use of FEC. In addition, OTN can reduce the amount of equipment that is required to support a multi-service network, whilst dramatically simplifying management.

As a result, most large carriers are now mandating G.709-compliant interfaces on their long-haul DWDM equipment, as it allows them to gradually migrate to packet-based services whilst still supporting traditional SDH/SONET applications. In addition, the aggregation capabilities of OTN means that carriers are increasingly implementing it in metropolitan networks as a way of reducing the number of interfaces required.

The choice of OTN is perhaps less clear for research and education networks with their own dark fibre who do not have legacy SDH/SONET equipment to support. In addition, the forthcoming OAM&P enhancements to Ethernet may yet undermine the

rationale for these functions being provided via OTN. Nevertheless, OTN is able to offer these features today, whilst offering improved transmission distances with FEC. In any case, many newer switch and router interfaces (both SDH/SONET and Ethernet) support configurable G.709 framing, which means networks need not be locked into a particular approach.

Ethernet

Ethernet encompasses a large range of IEEE standards (as specified in 802.3) for transmitting packet-based data over both copper and fibre optic cables. It was originally designed for use over a shared medium in local area networks, with the first standard appearing in 1980 (although there were earlier implementations). Since then, and despite competition from a number of other standards such as Token Ring and FDDI, it has emerged to become the ubiquitous technology for LANs. With a huge installed base of Ethernet, sheer market size means equipment can be developed and sold at low prices which has encouraged extension of the standards to support ever-faster line rates, a variety of physical transmission mediums, and the addition of new features. However, compatibility is maintained with the original architecture in terms of the Media Access Control (MAC) protocol, the frame format, and minimum and maximum frame size.

The widespread deployment of Ethernet with its lower equipment and management costs has more recently encouraged its adoption in metropolitan and wide-area networks. By adding support for longer-distance transmission over fibre, it has become possible to run Ethernet end-to-end without the need to convert payloads into other transmission formats (such as SDH/SONET). With most data traffic originating from Ethernet-based LANs, this becomes increasingly attractive in terms of reducing equipment and management complexity, not to mention being able to take advantage of the economies-of-scale. Even where legacy transmission systems must still be used, there are still advantages to encapsulating Ethernet as, being a statistically-multiplexed protocol, it is able to use bandwidth more flexibly than TDM-based protocols.

Since the introduction of Gigabit Ethernet (GE) in 1999, and 10 Gigabit Ethernet (10 GE) in 2002, it has become increasingly practical to utilise Ethernet in the wide area. Furthermore, with access to dark fibre becoming more prevalent, it has become increasingly easier to provision data networks using this technology, especially where there are fewer legacy issues. Research and education networks have been at the forefront of adopting Ethernet for WAN use, and this trend is expected to continue.

The most common Ethernet standards are as follows:

Version	Medium	Reach
<i>Ethernet (10 Mbps)</i>		
10BASE-T	Unshielded Twisted Pair (Cat3-6)	100 m
<i>Fast Ethernet (100 Mbps)</i>		
100BASE-T	Unshielded Twisted Pair (Cat3-6)	100 m
100BASE-FX	Multi-mode fibre	400 m – 2 km
100BASE-SX	Multi-mode fibre	300 m

100BASE-BX	Single-strand, single-mode fibre	10 km
<i>Gigabit Ethernet (1 Gbps)</i>		
1000BASE-T	Unshielded Twisted Pair (Cat5-6)	100 m
1000BASE-SX	Multi-mode fibre	500 m
1000BASE-LX	Single-mode fibre	2-20 km
1000BASE-BX10	Single-strand, single-mode fibre	10 km
1000BASE-ZX	Single-mode fibre (C-band)	70 km
<i>10 Gigabit Ethernet LAN-PHY (10 Gbps)</i>		
10GBASE-T	Unshielded Twisted Pair (Cat6, 6a)	55-100 m
10GBASE-SR	Multi-mode fibre	26-300 m
10GBASE-LR	Single-mode fibre (O-band)	10-25 km
10GBASE-ER	Single-mode fibre (C-band)	40 km
10GBASE-ZR	Single-mode fibre (C-band)	80 km
10GBASE-LX4	4 x 3.125 Gbps CDWM over MMF or SMF	300 m – 10 km
<i>10 Gigabit Ethernet WAN-PHY (9.952 Gbps)</i>		
10GBASE-SW	Multi-mode fibre	26-300 m
10GBASE-LW	Single-mode fibre (O-band)	10-25 km
10GBASE-EW	Single-mode fibre (C-band)	40 km
10GBASE-ZW	Single-mode fibre (C-band)	80 km

10GBASE-ZR and ZW are actually proprietary extensions of the 10GBASE-ER and EW specifications to increase transmission distances. Although supported by several vendors, they are not yet included in the 802.3 standards. 10GBASE-DWDM is also a proprietary extension available from one vendor that supports 10 GE over different wavelengths up to 80 km.

There are also other standards such as 10GBASE-LRM (for older multi-mode fibre), as well as 10GBASE-CX4 (for Infiniband-type wiring) and 10GBASE-KR/KX4 (for system backplanes) which are aimed at data centre applications. Finally, 10BASE-5 (for RG-8 coaxial) and 10BASE-2 (for RG-58 coaxial) may still be found in places, although these are now largely obsolete.

10 GE is a departure from previous Ethernet standards in that it offers two main variants. LAN-PHY offers a full 10 Gbps line rate and can run directly over a wavelength on a WDM system, whilst WAN-PHY is specifically designed to run over existing SDH/SONET infrastructures and therefore offers a reduced line rate of 9.952 Gbps.

WAN-PHY has not really found favour with network operators and only accounts for about 5% of all 10 GE interfaces sold. Apart from the fact that the lack of volume makes the interfaces expensive, there is the potential for rate mismatches when using it in conjunction with other Ethernet variants, resulting in unexpected packet drops. This may have been an acceptable trade-off when SDH/SONET was needed to provide OAM&P capabilities that could not offered by Ethernet itself, but even this advantage has been somewhat negated by the fact that OTN OTU2 can offer similar capabilities and is also able to support LAN-PHY. The consensus among vendors is that having different variants of 10 GE has not been beneficial to the market, and that future generations of Ethernet should just implement a single framing structure.

There has been a great deal of discussion at the IEEE about what should come next after 10 GE. A majority of vendors supported a 100 Gbps standard on the grounds that some networks were already utilising 40 Gbps connections (using SDH/SONET), and these would likely start to become insufficient for some users by the time a 100 GE standard was completed. Certain other vendors supported a 40 Gbps standard on the grounds that something cheaper than OC-768 was required, and that the technology to develop 40 GE was already available, whereas it was not for 100 GE. In addition, it is unlikely that a serial 100 GE standard will be possible for a number of years, which will mean using multiple channels (possibly 4 x 25 Gbps) thereby increasing cost and complexity.

The IEEE Higher Speed Study Group (HSSG) therefore agreed in July 2007 to work on both 40 and 100 GE standards⁴. It appears the 40 GE standard is intended to be a short-range technology for backplane and server interconnect applications, whereas 100 GE is aimed at network backbones and exchange points.

The outline specifications for 40 GE calls for full-duplex operation on reaches of at least 10 metres over copper cable, 100 metres over OM3 multi-mode fibre, and 1 metre over a backplane; with a bit error ratio better than 10^{-12} . It was also agreed that 40 GE should support a transmission rate compatible with OTN ODU3 as well. Nevertheless, one of the vendors interviewed by this study appeared to be looking at 40 GE in the context of wide-area networking and expected to be able to support reaches of 80 km or more over single-mode fibre. They anticipated that implementations should be available by 2009, although these may initially be pre-standard.

The outline specifications for 100 GE are to support full-duplex operation on reaches of at least 10 metres over copper cable, 100 metres over OM3 multi-mode fibre, and 10-40 km over single-mode fibre; with a bit error ratio better than 10^{-12} . It is anticipated this standard will be completed by 2010, with the first implementations available from 2011 onwards.

Most vendors interviewed by this study were working towards 100 GE, although there are still serious technical challenges to overcome. It is therefore expected that early 100 GE implementations will actually operate over four or more separate channels, with true serial implementations not being available until at least 2012. It is also likely that initial implementations will be utilised for interconnect applications in data centres and exchange points.

Unfortunately, whilst Ethernet is capable of offering high bandwidth and is cost-effective, it also offers few of the operations, administration, maintenance and provisioning (OAM&P) features that are necessary in a carrier environment. In particular, it is difficult to isolate faults in an Ethernet network, and restoration times can often be measured in the order of minutes rather than milliseconds. Furthermore, it originated as a shared transmission medium for a limited number of hosts, and was not designed with scalability in mind. Having said this, work is currently being undertaken to address these limitations and add carrier-grade features similar to those provided by SDH/SONET.

⁴ http://grouper.ieee.org/groups/802/3/hssg/public/july07/minutes_01_0707_unapproved.pdf

Although Ethernet has gradually evolved from only being a LAN technology, it has retained some of the inherent characteristics of this environment. A port in a bridge therefore needed to maintain a list of the MAC addresses of all the devices connected to a network. Bridges also learnt about new devices by snooping on the source address in the Ethernet frame header, whilst frames addressed to unknown destinations were forwarded to all active bridge ports (flooding). However, these solutions were clearly not scalable as the size of networks increased. Figure 1.1 depicts the evolution of Ethernet addressing:

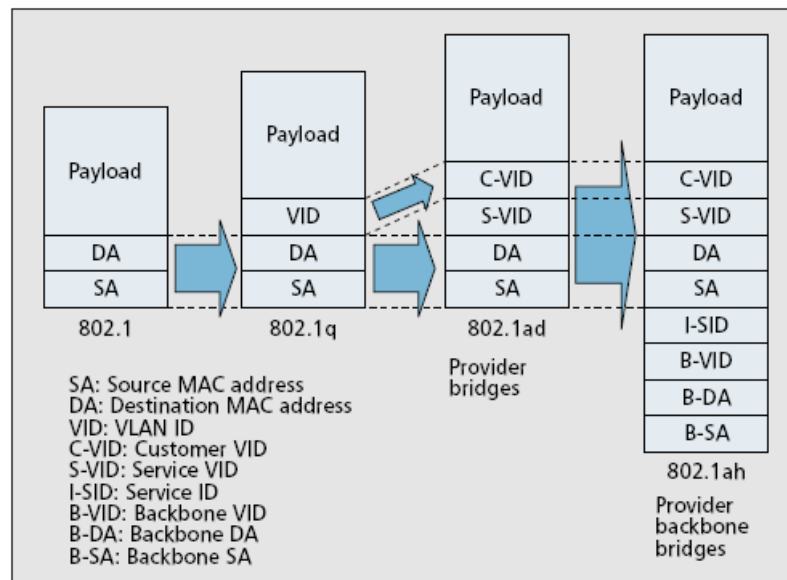


Figure 1.1 - Evolution of Ethernet addressing © IEEE

The introduction of the 802.1Q-1998 standard allowed partitioning of a network into 4096 virtual spaces by introducing an additional field (VLAN ID or VID) in the Ethernet frame header. This meant that a port only needed to maintain a list of addresses for the VLANs to which it belonged, providing a degree of scalability. However, the 12 bits allocated for the VID severely limited the number of customer networks that could be supported, which meant customer VLAN tags could not be maintained throughout a carrier network. The 802.1ad-2005 standard (also known as Q-in-Q, and currently referred to as Provider Bridges) therefore added an additional VID which was reserved for use inside a carrier infrastructure (known as Service-VID or S-VID). This allowed customer VIDs to be transported over a carrier network, although it did not address any of the other limitations.

The P802.1ah standard (formerly known as MAC-in-MAC, and currently referred to as Provider Backbone Bridges or PBB) that is expected to be finalised by May 2008⁵, aims to significantly increase the number of S-VIDs that can be provisioned from 2^{12} to 2^{20} . It will also define the protocols necessary to interconnect multiple Provider-Bridged Networks, as well as an appropriate management framework. However, historic Ethernet features like the automatic learning, flooding and hence the necessity of loop-free topologies (per-Backbone VLAN, Service VLAN and per-Customer VLAN) will apparently still be maintained.

⁵ <http://www.ieee802.org/1/files/public/docs2007/ah-editor-report-0507.pdf>

In parallel with PBB, another group is developing a series of extensions to the standard that would address the provisioning problem introduced by maintaining the historic Ethernet features. This work is being undertaken by 802.1Qay Working Group (also known as Provider Backbone Bridge Traffic Engineering or PBBTE) and is based on developments originating from Nortel and British Telecom⁶ (termed Provider Backbone Transport or PBT).

PBBTE will effectively create a connection-oriented network based on Ethernet framing and control protocols. Point-to-point circuits could be provisioned by allowing a network operator to disable unknown destination forwarding and source address learning for administratively selected VLAN Identifiers, whilst allowing other network control protocols to dynamically determine active topologies for other services. These interoperability capabilities will be supported by SNMP MIB management of individual bridges; by extensions to other control protocols specified in the standard; through the use of CFM with addresses and VLAN Identifiers specifying traffic engineered connections, and through 1:1 path protection switching capable of load-sharing.

It should be noted that the PBBTE specification is only concerned with intra-domain operation though. In particular, the control plane associated with populating the forwarding tables is excluded from the standard. Static entries to configure the point-to-point circuits may be configured directly by the bridge management software or automatically through some established control plane (GMPLS has been already proposed for this task⁷).

For inter-domain operation, Connection Fault Management (CFM)⁸ has been developed by the 802.1ag Working Group to allow hop-by-hop detection, as well as isolation and troubleshooting of connectivity problems. This was developed in conjunction with the ITU-T group working on the Y.1731 recommendation that adds performance management functionality.

As well as PBB, PBBTE and CFM, other work is under way to address the limitations of the historic Ethernet protocols. 802.1aq (Shortest Path Bridging) aims to provide an alternative to the Spanning Tree protocol that provides loop-free forwarding in Ethernets. 802.1Qaw (Management of data-driven and data-dependent connectivity faults) will also provide a set of protocols and managed objects for data-sensitive connectivity verification based on CFM. This work will address the problem of troubleshooting connectivity problems in networks that are increasingly deploying firewalls, access lists, traffic policing, or automated behaviour such as link aggregation or multicast group configuration through IGMP snooping.

As most of these standards are still under development, it is difficult to estimate when fully-compliant equipment will become commercially available. However, pre-standard PBBTE implementations are already available from several manufacturers,

⁶ D. Allan et al, *Ethernet as Carrier Transport Infrastructure*, IEEE Communications Magazine, Feb 2006, pp.134-140

⁷ <http://www.ietf.org/internet-drafts/draft-fedyk-gmpls-ethernet-pbb-te-01.txt>

⁸ <http://www.ieee802.org/1/pages/802.1ay.html>

and research and education networks are already investigating the possibilities of these⁹.

IP over DWDM

A few years ago, the only way to send IP packets over DWDM systems was to connect IP routers to ATM switches that would encapsulate the packets in ATM cells and transport them using SDH/SONET. As IP became increasingly prevalent and the need for multi-protocol networks diminished, Packet-over-SONET (POS) started to become popular as this allowed IP to be run directly over SDH/SONET (although technically it runs over PPP/HDLC over SONET). This not only allowed expensive ATM equipment to be dispensed with, but was more efficient as the use of ATM imposed a significant overhead. As result, most research and education networks currently run POS on their wide area connections.

Nevertheless, the continued need for SDH/SONET equipment when running DWDM systems is still not ideal as it adds complexity (e.g. protocol overhead, optical-electrical conversion, and cost) and there is significant functional overlap with IP in terms of multiplexing, protection and restoration, switching/routing and network management. Where different networking layers are unaware of each other, it becomes more difficult for the IP layer to optimise routing over the actual physical infrastructure, manage resources, and respond to performance problems or failures in the network. In addition, different management systems are required for each layer, and whereas IP networks were designed with automation in mind, provisioning SDH/SONET and DWDM systems requires much more manual intervention.

GMPLS (see Section 2.6) was therefore developed by the IETF to allow IP routers to communicate with WDM and SDH/SONET equipment. This allows discovery and establishment of paths through any underlying infrastructure, and enables optimal routing of IP packets using particular fibres, wavelengths and/or timeslots. In other words, it provides a mechanism to make an IP router aware of the capabilities of the underlying DWDM system.

IP-over-DWDM has been discussed for a few years, although products have only recently been marketed under this banner in order to exploit the fact that IP has become the ubiquitous protocol for transporting all types of data traffic. As well as traditional Internet applications, it is now very common for telephony, video and audio to be provisioned over IP, and in fact the volume of IP traffic has long since overtaken that of voice and other traffic on carrier networks. Proponents argue this negates much of the rationale for using SDH/SONET, especially when much of its value-added functionality could actually be provided at the IP level.

This said, the name *IP-over-DWDM* is a bit of misnomer as IP does not directly run over wavelengths, but over a link layer based on either simplified SONET or Ethernet framing. Indeed, one current commercial solution is actually based on 10 GE running over G.709 OTN which provides the Forward Error Correction and OAM&P functionality. However, it supports an enhanced version of GMPLS that allows IP and optical domains to be managed separately, ensuring that the propagation of topology and control information can be restricted across administrative boundaries.

⁹ <http://staff.science.uva.nl/~delaat/sne-2006-2007/p34/report.pdf>

As such, IP-over-DWDM is not a protocol in itself, but refers to initiatives to simplify transmission systems, whilst more closely integrating IP routing and optical control systems. It is perhaps better to describe it as umbrella term that brings together a variety of other initiatives aimed at providing better management and control of DWDM systems.

1.3 Transmission Equipment

Terminal Multiplexers and OADMs

A traditional WDM system comprises several discrete components, although these are increasingly being consolidated into *multi-service provisioning platforms*. The optical interface of a client system (routers or other equipment) typically transmits a signal on a standard wavelength of 850, 1310 or 1550 nm, depending on the connected fibre-type and reach. This signal known as *grey light* is then fed to an *optical line terminal multiplexer* that contains a transponder for each wavelength it can support, and which electrically converts the input signal into a particular wavelength. Multiple input signals (depending on the capabilities of the WDM system) can therefore be combined as different wavelengths and re-transmitted onto an outgoing fibre.

At the other end of a link, a terminal de-multiplexer splits a multi-wavelength signal back into individual signals, usually electrically converting them to an appropriate wavelength, before transmitting them onwards to client systems. In modern systems, the same unit can usually be configured to function as either a multiplexer or de-multiplexer.

Such equipment is sufficient for a simple point-to-point WDM system, but where multi-directional nodes are required, there would need to be duplication of terminal equipment, either in a back-to-back configuration or with switching equipment in-between. Alternatively, *optical add-drop multiplexers (OADMs)* can be used to insert or remove particular wavelengths on a trunk connection at a given node. Earlier OADMs required the whole signal on the trunk to be (de)multiplexed in order to add or drop wavelengths, but newer equipment is able to (de)multiplex specific wavelengths whilst allowing the others to pass through.

There are several issues with this approach though. The terminal (de)multiplexers require costly transponders for each wavelength, and are limited to specific channel bands and spacings. In addition, as optical-electric-optical (OEO) conversion is undertaken, they can only work with specific protocols (e.g. SDH/SONET, OTN or GE) at specific line rates. Therefore, in order to upgrade the network (number of wavelengths and/or line rate) or run different protocols, it is often necessary to upgrade the terminal (de)multiplexers as well.

OADMs are also statically configured, are only two-way (i.e. one trunk input, one trunk output, plus the add/drop port(s)), and can usually only add/drop a limited number of wavelengths from a trunk. Any reconfiguration needs to be done on-site, and it can be necessary to upgrade or cascade OADMs when additional wavelengths

are required. This limits the ability to dynamically reconfigure network topologies, especially in the event of link failure.

ROADMs

A relatively recent development is the *Reconfigurable Optical Add-Drop Multiplexer (ROADM)*. These can be remotely configured to add or drop wavelengths from a trunk on demand, whilst also automatically undertaking the necessary power balancing. Early ROADMs required all incoming wavelengths to be de-multiplexed which caused significant signal loss, but the current generation utilise wavelength-blocking techniques which limit signal degradation to more acceptable levels.

ROADMs typically support up to 40 wavelengths, although devices are becoming available that can support up to 80. Although ROADMs tend to be focused on DWDM systems, there are also CWDM versions available, as well as a few implementations that can support hybrid systems.

ROADMs are more expensive than OADMs, but they offer greater flexibility and lower operating costs once installed. As a result, they have started to be implemented in carrier networks, as well as a few research and education networks (e.g. Internet2, CANARIE and CESNET).

Further advances in the ROADM concept are the *Multi-Degree ROADM* or *Wavelength Selective Switch (WSS)*. These utilise MEMS (Micro Electromechanical Systems) or liquid crystal array technology that is not only able to add and drop wavelengths onto a trunk, but also direct them onto one of several other paths (typically up to eight). As well as being able to provide a cross-connect capability in a low-power device, this potentially allows the construction of mesh networks rather than the typical ring topologies used with existing ROADMs. However, this technology is not yet mature, and there are other problems associated with all-optical meshed networks.

Optical Switches

Another way of adding, dropping and switching wavelengths is to use either an *Optical Cross-Connect (OXC)* or *Photonic Cross-Connect (PXC)*, typically in conjunction with (de)multiplexers.

OXCs are more common, and switch wavelengths in the electrical domain. This also allows for wavelength conversion (changing from one frequency to another), and for signal regeneration. The downside is that OXC interfaces are protocol and line-rate specific, and the complex electronics required have traditionally made the equipment expensive.

However, the introduction of *photonic integrated circuits (PICs)* has allowed the separate components of a WDM system (transmitting lasers, modulators, multiplexers, attenuators and receivers) to be integrated onto a pair of chips (one for sending, one for receiving). A currently available solution (also marketed as a *Digital ROADM*) supports a 10 x 10 Gbps WDM system on a single PIC, several of which can be combined into a larger system. Not only does the reduced component count offer significant reductions in power usage (up to 75%) and by extension improved

reliability, but it greatly improves the cost-effectiveness of OEO conversion. In addition, it dramatically reduces the real estate required for a WDM system by consolidating the various separate units into a single chassis. Further developments in this area include an announced 10 x 40 Gbps PIC, and laboratory trials of a 40 x 40 Gbps PIC, although the latter is not expected to be commercially available for a few years.

PXCs represent more of a niche market, and are currently being developed by just a handful of vendors. They optically switch signals from one fibre to another using MEMS or piezoelectric beam steering technology, regardless of what is being carried. This allows them to either be used as automated fibre patch panels, or as a wavelength switch when used in conjunction with (de)multiplexers.

The advantage of PXCs is that they are relatively cheap (at around USD 500 per port), have very low power consumption, and can offer high port densities in small form factors (up to 320 ports). In addition, as they switch signals transparently, they are not protocol, line-rate or wavelength-density dependent.

Nevertheless, there are a couple of problems associated with PXCs. As they transparently switch signals, they cannot perform regeneration or wavelength conversion, which either means they must be located where this is unnecessary, or these functions must be undertaken by other devices. In addition, as only one particular wavelength can be used on any given link, signals from different sources using the same wavelength cannot be switched onto the same link. This problem (known as *wavelength blocking*) potentially reduces network capacity and requires careful engineering of topologies. It is especially problematic for meshed networks as more wavelengths become utilised.

Another issue is that any given link in a WDM system will have specific characteristics that require the signal power to be adjusted. Where a wavelength traverses two or more links, this not only makes it difficult to power balance the whole signal, but other factors such as amplification and regeneration must be taken into account.

A newer development is the integration of (de)multiplexers into PXCs or WSSs. This would allow wavelengths to be transparently switched, added/dropped, or converted as necessary. Although no commercial products are currently available, they potentially offer a solution for meshed networks. In reality though, it is likely that future research and education networks will be provisioned with a mixture of ROADMs, OXCs and PXCs, depending on requirements.

Tunable Optical Interfaces

Other recent WDM developments include the introduction of tunable optical interfaces on client systems (e.g. routers) and OXCs. By configuring the lasers to transmit at a particular wavelength, this allows the signal to be fed directly into the WDM system without the intermediate transponders. Several vendors already have tunable interfaces available for their equipment, although these currently have certain limitations (e.g. lack of support for FEC) that reduce their reach. Nevertheless, it is expected that tunable interfaces will become increasingly common on client systems.

PONs

Whilst a multiplexer is still necessary with tunable interfaces, this can in principle be a passive all-optical device that is protocol transparent and requires little or no power. This also allows such devices to be installed in sites where there is limited power availability (e.g. street cabinets), thus making fibre-to-the-home a possibility. Indeed, there are a series of ITU-T and IEEE standards with respect to *Passive Optical Networks (PONs)* such as BPON (G.983), GPON (G.984), EPON (802.3ah) and 10GEPON (802.3av). Such systems have already been trialled in the US, Europe and the Asia-Pacific region, with commercial deployment expected to start in 2007.

Modulation Formats

One of the problems with operating higher bit rates over WDM systems is that non-linear effects such as PMD, FWM and XPM start to become very problematic (See Section 1.1). Whilst newer types of fibre can help reduce this problem, the reality is that transmission systems must generally work over a mixture of different fibre types. For longer reaches, it is therefore currently necessary to employ expensive and complex compensation equipment.

Most current transmission systems utilise a *Non-Return-to-Zero (NRZ)* modulation scheme which works well for bit rates of up to 10 Gbps, but provides poor performance at higher speeds. For 40 Gbps systems, several modulation schemes have been proposed, but currently only *Carrier-Suppressed Return-to-Zero (CS-RZ)* and *Duo-Binary* are used commercially. Unfortunately CS-RZ is unable to work with a 50 GHz or closer channel spacing (which means that DWDM systems are limited to 40 channels or fewer), whilst Duo-Binary modulation still provides inferior performance when compared to NRZ operating at 10 Gbps.

Other possible modulation schemes include *Differential Phase Shift Keying (DPSK)* and *Differential Quadrature Phase Shift Keying (DQPSK)* which offer performance advantages over other techniques, although phase encoding adds significant complexity to the receiving equipment. More recently though, new techniques such as *Dual-Polarisation Quadrature Phase Shift Keying (DP-QPSK)* have become practical due to advances in high-speed DSP technology. When operating at 40 Gbps and beyond, these promise similar reach and channel densities compared with existing 10 Gbps systems, and without the need for compensators.

Planning and Interoperability

The main concern with WDM systems is the need for careful planning, and this does not change even with more easily reconfigurable elements in the network. It is not only necessary to plan which wavelengths should be added and dropped at each node, but the power levels on every link must be balanced in order to minimise signal degradation. All the elements in the transmission system must also be budgeted for, as each introduces a certain degree of power loss.

In addition, although there is a standard ITU-T frequency grid for both DWDM and CWDM that aims to facilitate interoperability between equipment from different vendors, some vendors utilise non-standard frequencies to implement additional wavelengths, or use different wavelength numbering. This added to a lack of

standards with respect to power levels and modulation schemes, means that it can still be quite challenging to provision multi-vendor WDM networks. As a result, most WDM networks are still implemented using equipment from the same vendor, or at least vendors that have close collaborations.

Routers

IP routers are mature products and are likely to remain as the backbone of the Internet for the foreseeable future. There is some debate about whether IP services would be better provisioned by larger or fewer routers in the IP core, or whether smaller edge routers could be utilised in conjunction with more intelligent core switching at Layers 1 and 2. Forthcoming Ethernet (see Section 1.2 ‘Ethernet’) and GMPLS (see Section 2.6 ‘GMPLS’) developments offer this prospect, although there does not currently appear to be much consensus on this issue amongst vendors.

The newest recent router designs are capable of supporting 1-2 Tbps bi-directional aggregate throughput, and up to 100 Gbps per interface card. Vendors claim these systems have been developed with 100 GE in mind, so should be able to support these interfaces when they arrive. Alternatively, they may be used for 40 Gbps (OC-768) or multiple 10 Gbps interfaces (POS or 10 GE).

The fastest available router interfaces are currently 40 Gbps (OC-768) and these are currently extremely expensive. However, advances in chipset and laser technology are expected to significantly reduce these costs in the near future. In addition, at least one vendor is offering an OC-768 solution that utilises inverse-multiplexed 4 x OC-192s in order to provide a workaround to the current transmission issues at 40 Gbps (see Section 1.1).

As it is likely that Ethernet will become the transmission protocol of choice, this should considerably reduce the cost of high-speed interfaces in the next few years. 10 GE interfaces are expected to fall to USD 1-2K per port by 2010 or earlier. 40 GE interfaces may also be available by 2009, at around twice the prevailing cost of a 10 GE interface. Looking further ahead, 100 GE is expected to be available by 2011, although this is initially likely to be provisioned as 4 x 25 Gbps.

Power Consumption

An increasing concern in the past couple years has been the ever increasing amounts of power required by networking equipment. More bandwidth requires more processor cycles, memory, and interface optics that all consume electricity and generate heat. Five years ago, a typical carrier-class router consumed 1 to 3 kW of electricity, but today this has risen to 5 to 7 kW. This has caused planning issues for some data centres, and in some cases limited the amount and types of equipment that can be hosted.

Nevertheless, manufacturers have become increasingly aware of this issue and appear to be designing future equipment to be compatible with existing data centre infrastructures. Although most carrier-class routers and OXCs have been designed with a maximum power consumption of 10 kW in mind, increasingly efficient chipsets, optics and backplane designs mean that 1-2 Tbps routing and switching platforms are likely to stay within the 5 to 7 kW range. In addition, other power

savings can be made through a reduction in the number of WDM and/or SDH/SONET elements required to provision a network, or through the use of more passive components.

The interviewed manufacturers suggested that current chassis designs should be capable of supporting multiple 100 Gbps interfaces in future. Looking beyond this, it is expected that liquid-cooled systems will be required, and at least one manufacturer already has designs for this.

Another solution may be more widespread use of DC power supplies in data centres. Most networking equipment takes an AC supply (although some offer DC options), but this must be converted for use with the electronics and optics. These conversions are relatively inefficient, especially when multiplied over all the equipment in a data centre, so one solution might be to undertake the conversion at the ingress point to a data centre, with DC then being distributed internally. Unfortunately, whilst this potentially offers power savings of 10-20%, it would require existing power distribution infrastructures to be replaced, not to mention that DC distribution cables are larger and more unwieldy than their AC equivalents. Furthermore, DC power supplies tend to be more expensive, so any savings in electricity costs would have to offset against increased capital expenditure.

1.4 Conclusions

Research and education networks increasingly have access to dark fibre, and must therefore provision their own transmission systems. In addition, they are increasingly managing the long-haul links themselves, whereas earlier networks tended to use carrier-managed fibre. This means that issues such as fibre quality, maximum transmission distances, the location of amplification and re-generation sites (if required), and the capabilities of DWDM or CWDM equipment have become more relevant.

New low-loss fibre is being developed that promises the ability to support multiple 40 Gbps channels over distances of 140-150 kms before amplification, and up to 2,000 kms before regeneration. This should also be able to support multiple 10 Gbps channels over distances of up to 4,000 km before regeneration. In addition, the introduction of G.656 fibre provides good support for both CDWM and additional DWDM channels.

Nevertheless, whilst some research and education networks are in the advantageous situation of being able to install or specify the most suitable fibre for their requirements, in reality most providers will lease or rent older fibre such as G.652. Newer fibre such as G.655 or G.656 may be obtainable as well, but it is quite possible that any given route will comprise a variety of different fibre types. It is therefore necessary to tightly specify requirements when procuring fibre, and then undertake extensive acceptance testing upon delivery.

For these reasons, it may also prove difficult to take advantage of recent advances in transmission capabilities, although it may not be desirable to push the limits of long-haul fibre installations anyway. Most equipment manufacturers develop their

specifications with existing fibre installations in mind, and faster line rates such as 40 and 100 Gbps may not support the more extreme ranges in future. Optical amplifiers work in the analogue domain and should not require upgrading for faster line rates or when wavelengths are added, but they are unable to correct the non-linear effects that become more pronounced at higher bit rates and/or with closer channel spacing.

At the present time, the fastest transmission equipment is able to support 40 Gbps using SDH/SONET (OC-768) over a limited number of wavelengths. Unfortunately, the cost of the interfaces is prohibitively expensive for most research and education networks, although prices are soon expected to drop significantly as more manufacturers come into the market with OC-768 solutions. However, the future of SDH/SONET is unclear as most vendors appear to be focusing on Ethernet for next-generation transmission systems.

The larger market and relatively simpler technical requirements of Ethernet means that it offers significant cost advantages over equivalent SDH/SONET equipment (e.g. an OC-192 interface is approximately 10 times the cost of a 10 GE interface). The addition of features to support wide-area usage means it can now compete with SDH/SONET in carrier-class networks, whilst still supporting a variety of transmission mediums in local and metropolitan area networks. As a result, most manufacturers are developing 40 and 100 GE systems which also have the OAM&P and virtual circuit functionality needed by carriers.

It is anticipated that the first implementations of 100 GE will arrive around 2010, although these are initially likely to require four or more channels and will be restricted to short-haul applications. True serial implementations are not expected before 2012.

In the meantime, even though 40 GE is aimed at data centre applications, implementations are expected to be available by 2009. This may therefore provide a interim solution before 100 GE arrives, with 40 GE interface costs expected to be around 40% of equivalent OC-768 interfaces. In addition, advances in processing capabilities will allow new modulation techniques that minimise the effects of PMD, XPM and FWM to be used, thus supporting spans of 80 km without amplification, and up to 2,000 km before regeneration over existing DWDM systems.

Another interim solution might be to concatenate multiple 10 GE channels to create bigger channels. This is already being utilised at network exchange points, and with average prices of 10 GE interfaces expected to fall to around USD 1-2K by 2010, this may be a cost-effective approach. However, this approach is limited by the number of ports that can be physically accommodated on each switch, and the processing power needed to handle the load balancing. Furthermore, splitting traffic over multiple physical connections may cause problems for some applications.

With respect to DWDM systems, many vendors would appear to see limited requirement for increasing the number of wavelengths that can be supported on each fibre. Although some vendors offer UDWDM solutions (≤ 25 GHz spacing), most appear for now to have settled on ~ 80 -channel systems, as 50 GHz spacing has been found to provide a good tradeoff between faster line rates and longer reaches. This is perhaps of limited concern to research and education networks who generally do not

use more than a handful of wavelengths on each route, and should have sufficient spare capacity in this respect for the foreseeable future.

A recent development has been wavelength-selectable interface cards for optical switches and routers, and ROADMs that allow individual wavelengths to be added or dropped from a fibre on-demand without the need to convert all the channels into the electrical domain. This simplifies WDM systems as it eliminates the need for separate transponder and OADM equipment that needs to be physically configured, and opens the possibility of automatically configurable WDM systems.

Another interesting development is the introduction of PICs that consolidate the separate components required for WDM systems onto a couple of chips, and have the potential to make OXCs more cost effective and easier to manage. At the same time, more advanced PXC's that can switch at the wavelength level are likely to appear from 2008 onwards, and these may have interesting applications where cheap, dense, and/or low-power switching is required.

Finally, power consumption and heat dissipation of transmission systems has become an increasing concern. However, this is perhaps more of a long-term problem as vendors realise they need to design their equipment to be compatible with existing data centre infrastructures. Increasingly efficient chipsets and optics mean that 1-2 Tbps routing and switching platforms can still fall within the 5-10 kW range, not to mention that power savings can be made through a reduction in WDM and SDH/SONET equipment counts. It is anticipated that conventional chassis designs should even be able to support multiple 100 Gbps interfaces, although beyond that liquid cooled systems will be required (for which designs are already available).

2. Control Plane and Routing Technologies

Control planes are the intelligence that establish, maintain, and tear-down the different connections within a network. They should also have the ability to calculate optimal paths between end points, distribute traffic loads, and re-route connections in the event of failure. Although the term control plane commonly refers to the switching mechanisms used in optical networking, for this purpose of this study encompasses IP routing mechanisms as well.

The research and education networking community has a lot of experience with IP routing, and it is generally well understood by network administrators. IP routing in its various guises also largely works automatically with minimal manual intervention, although this may change as the Internet grows in size and becomes less hierarchical in terms of the relationship between service providers and/or big users. Routing protocols are required to handle ever greater amounts of information, but they also have minimal security mechanisms to protect against abuses. It is therefore perhaps necessary to enhance, supplement or replace some of the existing protocols in order to improve resilience, security, and peering and policy management, and possibly in conjunction with certain middleware developments as well.

By contrast, research and education networks have little experience of managing optical networks, and the levels of automation within optical switching are currently very limited. End-sites/users typically have to request bandwidth at fixed rates for predefined periods of time, with (re)configuration of the network being done manually. This makes it difficult to dynamically provision bandwidth (e.g. some end-sites/users may require more bandwidth at certain times), and there is a lack of resilience in the event of link failure.

In addition, as research and education networks move from IP-only services towards running the underlying optical infrastructure as well, they need to consider whether to adopt an overlay or peer approach. The overlay approach keeps the optical and IP domains functionally separate, which is simpler and easier to understand. However, in the peer approach where switches and routers share topological information, there are opportunities for connection and management efficiencies. In reality, it is likely that NRENs will have to adopt both approaches – for example, adopting a peer model for their internal networks, whilst supporting overlay networks for their end sites/users.

2.1 Routing Scalability

In the past few years, there has been general recognition that the Internet routing and addressing system is facing serious scaling problems¹⁰. This is not only due to a huge increase in the number of hosts, but also due to a rise in multi-homing, provider-independent addresses, traffic engineering, and policy routing. At the same time, there are sub-optimal address allocations that exist for both historical reasons, and because of more recent mergers and acquisitions between service providers. In addition, there seems to be a general trend towards more direct peerings between larger organisations, that is somewhat deprecating the traditional hierarchical structure of service providers (i.e. Tiers 1-3).

¹⁰ <http://www.ietf.org/internet-drafts/draft-iab-raws-report-02.txt>

Routing is of course a critical component of the Internet as it provides the information that allows IP packets to be sent across the Internet to their final destination. Routing protocols also attempt to compute the most optimal route for these packets by exchanging data with other routers, subject to any policy and transit restrictions that may be applicable. From this, a local routing table (more formally known as Routing Information Base or RIB) is generated that in principle allows a router to know how to reach any destination on the Internet, and from this derive the forwarding table (Forwarding Information Base or FIB) that identifies which interface on a router that packets need to be sent in order to reach a specific destination.

The routing protocols currently used within the Internet are very mature technologies, having been extensively deployed at least fifteen years. They are well understood, and have generally supported the scaling of the Internet from a few thousand hosts up to the many millions that exist today. In particular, BGP4 is well-established as the interdomain routing protocol of choice, and is used by most (every?) service provider in order to establish how to route traffic between each other, and increasingly, for routing traffic internally as well.

Unfortunately, one of the largest problems faced by BGP is the growth of the routing table, especially for those routers in the so-called default-free zone (DFZ) that must maintain a comprehensive routing table advertising the whole Internet. However, whilst the DFZ is somewhat analogous with the Internet 'core' that is comprised of higher-tier service providers, it is also possible for multi-homed end-site routers to have routing tables of comparable size.

The primary concern with the DFZ routing table is its current growth at greater than linear rates. As of July 2007, it had reached approximately 230,000 entries, and is projected to increase to 370,000 entries within five years¹¹. Looking ahead, some projections suggest that it could reach 2 million entries within 15 years¹², and when considering these figures, it should also be remembered that many routers will also have a variety of other routes configured for internal, traffic engineering, and VPN purposes.

Another concern is the increasingly number of dynamic updates that are required to maintain the routing tables as more networks join the Internet. Routes regularly appear and disappear because of engineering and policy decisions, temporary outages, and misconfigurations, and as a result a router may observe up to 400,000 BGP updates per day¹³. This problem is exacerbated as routing prefixes become increasingly de-aggregated, which exposes the core routers to the generally less stable edge networks.

Both the size of the routing tables and the number of routing updates required, obviously has implications for processing and memory requirements. The availability, cost, and to a certain extent thermal characteristics of these technologies, will effectively determine the ability to develop the high-performance routers of the future.

¹¹ http://www.cidr-report.org/as2.0/#General_Status

¹² <http://bgp.potaroo.net/>

¹³ G. Huston, *More ROAP: Routing and Addressing at IETF68*, IETF Journal, Vol. 3 Issue 1 (May 2007)

In fact, none of these problems are especially new. Back in the early-1990s, the size of the routing tables was rapidly outstripping the capabilities of the contemporary routers, but the solution then was to introduce CIDR (Classless Interdomain Routing, RFC 1518) that drastically dampened the rate of growth in the size of the routing table and provided operators with a great deal of breathing space. Since then, manufacturers have largely relied on Moore's Law (which essentially observes that chip densities double every two years for the same cost) to ensure they can build scalable routers at cost-effective prices.

Unfortunately, the Moore's Law paradigm appears to be starting to have some limitations, especially with respect to memory. Most commodity memory is based on DRAM technology, but whilst its physical capacity has more than doubled every two years, its speed has only increased by about half that rate. As routing and forwarding becomes more demanding, this becomes more problematic as routing tables are usually stored in DRAM, as are forwarding tables on lower-end routers. One solution might be to utilise lower-latency SRAM which has traditionally been used in small quantities for processor caching and consumer device applications, and which is also used for forwarding tables on high-performance routers. Unfortunately though, the router market has not proved sufficiently mainstream to leverage low unit costs, especially with respect to the large memory sizes that are required.

Another problem is that high-performance routers are no longer able to use commodity CPUs (such as those from Intel or Motorola), and thus take advantage of economies-of-scale. In order to achieve the necessary performance, specialised ASICs have to be used, which means they are subject to higher costs due to the need for customised fabrication, low production volumes and shorter technology life spans. Whilst Moore's Law also applies in principle to the capability of ASICs, the doubling of capability every two years is being accompanied by a 1.5 times increase in cost. This essentially means that to achieve the expected levels of improved routing performance, a 33% increase in expenditure will be necessary every couple of years; and this without taking into account the additional costs associated with powering and cooling higher performance hardware.

Over the past seven years, routing tables and updates have been observed to have grown by a factor of between 1.3 and 2. Whilst this implies that technology should just about be physically capable of keeping pace with this level of growth for the foreseeable future, it also implies the need for increasing expenditure in real terms which may be prohibitive. Furthermore, these assumptions are based on traditional usage trends, when in fact new developments may significantly exacerbate the problems.

The most obvious concern is increasing IPv6 deployment which could potentially have a huge impact on routing. Not only are IPv6 addresses four times longer than IPv4 addresses, but the sheer number available vastly increases the potential number of routes. Although IPv6 was designed with route aggregation principles in mind, subsequent changes in business structures and increased use of provider independent addressing have affected IPv6 in same way as IPv4. At the moment, there has not been a huge uptake of IPv6, but this may yet change as IPv4 addresses become increasingly scarcer.

Another potential development is a large increase in the number of mobile networks, such as the in-flight services pioneered by the airline industry. Although these have yet to be a commercial success, there may potentially be many thousands of networks on planes or ships dynamically appearing and disappearing, and having to send frequent routing updates as they move around the world.

Similarly, many millions of mobile end-user devices are likely to appear in the next few years. Their direct impact on routing issues is not clear as Mobile IP (RFC 3775) offers a solution whereby mobility is handled through redirection via home agents. However, they will clearly bring an increase in the number of hosts, which in itself has implications for routing.

Both the IAB and IETF have recognised there is potentially a serious scaling problem with the routing and addressing system, and have therefore started to investigate whether there are any alternative approaches that could reduce reliance on hardware scalability in the future. They concluded that the use of IP addresses for both location and identification is a large part of the problem, particularly at the edge of the network.

It is quite clear that customer networks do not wish to renumber if they change providers, as it is time consuming and expensive. Many networks have therefore settled on using private addresses internally and NAT to the outside world, even with the attendant problems this brings for certain applications. Even if one uses IPv6 with its larger address space and much vaunted autoconfiguration features, there will still be devices and applications that use static configurations that complicate renumbering. At the same time, many customer networks are multihomed for resiliency, geographical or other reasons, which either requires provider independent addresses that are difficult to aggregate, or the assigning of different addresses from each service provider which again complicates routing.

As a result, increasing consideration is being given as to whether IP addresses could be split into separate locator and identifier elements. The idea is that part (or all) of an IP address could be used as a globally unique identifier for an endpoint, with the remainder of the address (or even a prefix) being a locator that could be dynamically updated depending on where one happens to be situated in the Internet. In principle, this should simplify routing by ensuring there are always contiguous destination addresses that can be aggregated into a smaller number of routes in the routing tables.

Similar ideas have been proposed in the past, back when IPng (which became IPv6) was being discussed, and since then with GSE (RFC 1955) and SHIM6 as proposed IPv6 multihoming solutions. There are also certain analogies with IP mobility, where it is unnecessary for a sender to know exactly where the recipient is located, so long as it knows where to send packets where they will be forwarded onwards. More recently, there has been work in both IRTF and IETF on the Host Identity Protocol (HIP) that introduces a new name space for unique identification of hosts based on public keys, although this uses IP addresses merely as locators and requires applications to use separate host identifiers. This has implications for backwards compatibility, even though mechanisms have been defined to ease the transition.

Unfortunately, whilst the principle of splitting addresses into locator and identifier elements is gaining ground, the best way in which this might be implemented is still at an early stage of discussion within the IAB and IETF. At the moment, the discussion is largely concerned with defining the nature of problem that needs to be solved, and what are the most practical ways of solving it. However, even if a practical solution is found to the locator/identifier issue, it is still unclear whether the rate of BGP updates will eventually exceed the capacity of the routers to process them, particularly considering the increasingly meshy and dynamic nature of the Internet. Further investigation is needed to determine whether BGP can be modified or adapted to cope with future requirements, or indeed whether a completely new interdomain routing protocol will be required.

The implication of all of this for research and education networks is unlikely to be apparent for at least the next few years. The deployment of any sort of new or modified routing technology is clearly some way off, and would take longer to refine and have a significant impact on the routing system. In addition, whilst some (although not all) core NREN routers are exposed to DFZ conditions, carrier-class routers are today able to handle up to 1 million routes, with most enterprise-class routers being able to handle up to 500,000 routes.

The impact of BGP updates is perhaps more of an unknown quantity, particularly as there is also a trend to integrate more functionality in routers such as traffic shaping, policing, filtering and deep packet inspection which all consume processing power. However, the current generation of routers are likely to be able to cope with the expected growth of routing tables and associated updates during their service lifetimes, although NRENs running mid-sized routers in the DFZ should be aware of the possible need for incremental upgrades.

This said, the routing and addressing issue represents a fertile area of investigation to which the research and education community might usefully contribute in terms of operational experience and standards development.

2.2 IPv6 trends and developments

IP Version 6 is intended to be the successor protocol to IP Version 4, which is in general use in the Internet. The main improvements are a significantly larger address space (2^{128} compared to 2^{32}), autoconfiguration features, and better support for QoS management, multicast, and mobile IP. In principle, IPv6 also allows the end-to-end communication model to be restored by eliminating the need for network address translation.

The core IPv6 specifications and most of the related protocols were completed by the IETF some years ago, although some design issues remain unresolved such as IPv6 multihoming. Most router vendors support IPv6, and many NRENs are already running IPv6 alongside IPv4 in so-called dual-stack systems. Unfortunately though, usage remains minimal.

Some academic institutes have deployed IPv6 throughout their campuses, but most only support IPv6 for limited groups of users. This said, IPv6 support in current

operating systems is quite good and is often enabled by default, whilst many common applications are increasingly supporting IPv6 and favouring it over IPv4. Hence the use of IPv6 should increase quickly as IPv6 is deployed on the edge networks and in hosts/clients.

The main reason for the slow uptake of IPv6 is that there has been no urgent need for it. Most agreed that IPv6 would be needed at some point, but this would be many years ahead and was not something to worry about now. However, the original predictions that the IPv4 address space would be exhausted in 10 to 20 years time, are now being revised to 3 to 5 years hence¹⁴. In other words, IANA may soon have allocated its last /8 to a Regional Internet Registry (RIR), with the RIR pool being exhausted a couple of years after that¹⁵. As a result, it is being speculated that people may attempt to acquire even more addresses before it is too late, and that there will be trading of addresses at high prices. There will undoubtedly be increased use of NAT which will mean that a smaller percentage of hosts on the Internet will have globally reachable addresses.

The main question is what extent users will require globally reachable addresses. We are likely to see more networks with few global IPv4 addresses using private IP addresses and NAT, but with IPv6 in addition. There will also be networks going IPv6 only, but using translation or encapsulation (using gateways) in order to communicate with IPv4-only networks as well. Academic institutes may have more IPv4 address space available to them than many commercial enterprises, to the extent that the move to IPv6 may be somewhat less urgent. Nevertheless, it will become increasingly difficult for institutes to obtain additional IPv4 address space, and even if they have sufficient addresses, and they may still wish to deploy IPv6 to avoid the use of translation when communicating with others who are forced to move to IPv6. New network deployments and services are increasingly likely to be IPv6 only as well.

Deploying IPv6 is in principle straightforward, although there are several potential hurdles. For instance management or security-related systems such as firewalls or intrusion detection systems may need to be updated. In addition, some institutes use systems to track the use of IP addresses that would need to be modified, and make use of DHCP snooping to ensure that users can only use assigned addresses or limit the port on which DHCP responses arrive. Unfortunately, such features are not yet widely available for DHCPv6.

With IPv6, one may also use Stateless Address Auto Configuration (SLAAC), where host addresses are based on the link layer address of the interface, which may affect maintenance of DNS and access lists. In addition, when either DHCPv6 or SLAAC is used, a host makes use of Router Advertisements to learn about default routers. This may be a security risk if a host sends rogue advertisements, and whilst some switches can block DHCP responses from servers on non-server ports, there appear to be no switches that can block Router Advertisements on non-router ports.

Most difficulties are related to deployments on edge networks, but there are some issues related to effective deployment in the core as well. Although many NRENs deployed IPv6 some years ago, issues were discovered with some routers that were

¹⁴ <http://www.arin.net/statistics/statistics.pdf>

¹⁵ <http://www.potaroo.net/tools/ipv4/>

unable to undertake efficient forwarding of IPv6 packets at wire speeds, even though the situation is improving. There are also issues related to network management as few routers currently support NetFlow on IPv6 flows, and whilst IPv6 MIB support is improving, the variety of tools available remains limited. The IETF recently decided to adopt new IP version independent MIBs, and whilst one could adapt existing management tools to support IPv6-specific MIBs, it probably makes sense to use the protocol independent versions although it will take a while for these to be fully supported.

2.3 IP Multicasting

IP multicast technology is now about twenty years old, and whilst there has been a lot of interest from academic communities who have developed and deployed many interesting applications, it never has become the success that many hoped for. More to the point, whilst it is heavily used in certain closed networks, interest in using it in the wider Internet has been fading in recent years.

It is difficult to exactly say why multicast hasn't taken off. One issue is that a multicast service can be hard to manage, many of the protocols are fairly complicated, and there are not that many tools available. It is also unclear how the general Internet service provider can benefit, unless that service provider works together with content providers, or is themselves a content provider. As a result, the lack of multicast means few applications support multicast, whilst others (e.g. peer-to-peer) do the equivalent of multicast at the application layer instead.

As previously mentioned, multicast is used heavily in certain closed networks. One trend in recent years is increased interest in distributing television using IP multicast. In some cases, this is only used internally in television provider networks, but there are also cases where IP multicast is used to the end-user (e.g. with so-called *triple play*).

Increased use of multicast in closed networks should make it easier to deploy multicast on the Internet, but there are some particular issues to be considered when undertaking inter-domain multicast (i.e. crossing administrative boundaries). This said, there are a number of technologies and tools that might make it easier to deploy multicast.

The most common multicast routing protocol used today is PIM-SM (PIM Sparse Mode, RFC 4601). There may be some networks still using older protocols, but it is believed that most will have now moved to PIM-SM. Whilst PIM is used for signalling between routers in order to perform multicast routing, another protocol named IGMP (Internet Group Management Protocol) is used by hosts to signal to the routers which groups they wish to receive.

Most networks today are probably using IGMPv2 at the edges. There is increasing deployment of IGMPv3, which is needed for IPv4 SSM (Source-Specific Multicast), but there are still some routers and hosts that do not support it.

When doing multicast in large switched LANs, it is useful to have the switches constrain the multicast flooding to the switch ports where there is interest. The most common way to do this is IGMP snooping, where a switch listens in to IGMP reports from the hosts, in order to learn which groups should be forwarded to which ports. The issue here though, is that if you deploy IGMPv3, which is needed for SSM, you will need IGMPv3 snooping support in the switches. There are other techniques like the Cisco-specific CGMP which appears to be gradually disappearing, and the IEEE GARP GMRP which has seen little deployment.

For PIM-SM, one needs to have one or more routers acting as so-called Rendezvous-Points (RPs) for traditional multicast (Any-Source Multicast, ASM, as opposed to SSM) to work. If there is a need for multiple RPs for the same multicast group (typically for inter-domain multicast), but this is also wanted a single domain as well, one needs to use MSDP (Multicast Source Discovery Protocol, RFC 3618). In general, MSDP is needed unless one only has a small domain with no external multicast connectivity. However, there are serious issues with the scaling of MSDP, at least in the wider Internet where there are likely to be thousands of RPs.

In summary, current multicast deployments use a combination of PIM-SM, IGMP and MSDP, which all have problems regarding complexity, management and scalability. Many academic networks, have deployed them and they mostly work, but it takes some effort to understand, manage and debug the whole system.

Some new multicast technologies have recently been introduced that might allow for wider deployment and use of multicast.

Source-Specific Multicast

Source-Specific Multicast (SSM) is a relatively big change in the multicast service model. It requires that receivers specify which sources they will receive from, which means that source discovery has to be taken care of at the application level.

This approach brings two big benefits. The first is that multicast routing is greatly simplified as PIM-SM is still used for multicasting, but there is no need for RPs, and it is possible to immediately construct and use the so-called *shortest path tree*. This not reduces network requirements, results in quicker convergence, and makes for easier management. The other benefit is added security as only multicast packets from sources that are explicitly joined will be forwarded.

SSM is not exactly a new technology, but IGMPv3 for IPv4 (and MLDv2 for IPv6) is supported in the most popular operating systems, and many switches support IGMPv3 (and to some degree MLDv2) snooping. There is still a lack of applications supporting SSM, but this will hopefully change now that the prerequisites are in place.

Bi-directional PIM

Bi-directional Protocol Independent Multicast (BIDIR-PIM) is a relatively new variant of PIM, although some of the major router vendors have had implementations for a few years. BIDIR-PIM makes use of Rendezvous-Points and shared trees like PIM-SM, but the difference is that bi-directional shared trees are always built when

the routers learn the RP addresses, This means that there is absolutely no source state, hence less state in the network and more source mobility. There are also no PIM register messages since packets from a source are sent towards the RP using the pre build shared tree (the tree is bi-directional).

This has several benefits. There are no data-driven events and a router is not required to function as an RP, making multicast much easier for the routers. It may seem like a contradiction that no router needs to function as an RP, but all that is actually needed is an RP address that need not be the address of an actual router. The RP address just needs to be routed towards a link in the network, so that some link is the root of the shared tree.

The fact that a pre-built shared tree is always used, means that the tree is already in place when a source starts sending, and the initial packet can easily be delivered. This is also useful for some applications that may have a gap of several minutes between individual packets; a situation where PIM-SM does not work so well. Having a single forwarding path also makes things easier to manage, although the downside is that packets may often not take the shortest path through the network, making the location of the RP significant.

One problem with BIDIR-PIM is that it cannot be used for interdomain multicast, since there can only be a single common RP for any given multicast group, through which all traffic would flow. For PIM-SM, one can MSDP to bridge domains, but this cannot easily be done with BIDIR-PIM. In principle, it should be possible to use a BIDIR/SM border router based on MSDP that can join an external source and forward data internally. Then in other direction, a border router would be required that sees all BIDIR traffic (e.g. a router on the link where the RP address is), and then signals all sources seen using MSDP. However, there do not currently appear to be any implementations adopting this approach.

BIDIR-PIM has recently been specified in RFC 5015.

IPv6 Multicast

In principle, IPv6 multicast is quite similar to IPv4, as one uses PIM-SM and BIDIR PIM in the same way. For IPv6, there is MLDv1 and MLDv2 that are quite similar to IGMPv2 and IGMPv3 respectively, with the only difference being that IPv6 messages make use of ICMP. MLDv2 is needed for SSM in the way that IGMPv3 is required for IPv4 SSM.

One difference is that IPv6 multicast has a well defined address and scope architecture, and has with many different scopes defined. In addition, there is no MSDP for IPv6 (as MDSP does not really scale since any source activity in one domain is announced to all other domains running MDSP), and hence one cannot do IPv6 inter-domain multicasting without adopting a new trick.

This is trick is called Embedded-RP (RFC 3956) which provides a new RP discovery mechanism for IPv6. The idea is very simple, namely to encode the RP address into the multicast group address, which is possible by putting some minor restrictions on the choice of RP addresses. The result is that a router that sees a join request or a data

packet sent to such a group, will immediately know the RP address. It's far from obvious why this is useful for inter-domain multicast, but for IPv4 with MSDP, every domain can have their own RP for the same group, avoiding the need for one domain to have to rely on another domain which may not even be the source. Since there is no MSDP, or other source signalling for IPv6, one will need to have only a single RP for any given global group on the Internet, at least to have global connectivity between all sources and receivers for the group. It would be bad to have to rely on a single RP in the Internet for this, but using Embedded-RP makes it easy for the domain providing the content, to have their own RP and pick a group address that embeds of the RP address. In this way, everyone on the Internet listening or sending to the same group, will make use of the same RP, since all routers on the Internet that see a join or a data packet, will use the same RP address.

With MSDP you would typically have only one or a few RP addresses used by any router, although the same group would have different RP addresses in different domains. By contrast, with Embedded-RP you will have many RP addresses used by a router, but the same group will have the same RP address in all domains. This provides a much more scalable solution than IPv4 with MSDP.

Automatic Multicast Tunnelling

As mentioned in the introduction, there is a chicken-and-egg problem with the deployment of multicast. The lack of multicast in the network means that few applications support it, and with few applications it means there is little interest in deploying it. To solve this problem, the IETF is working on a solution known as Automatic Multicast Tunnelling (AMT). The current specification is in draft-ietf-mboned-auto-multicast-07.txt.

The idea, inspired by mechanisms like 6to4 and Teredo for IPv6, is that a host or home router will find a convenient multicast relay if native multicast is not available, which it will use to send and/or receive multicast through a tunnel terminated at this relay. If relays are deployed, and host systems or home routers support this, then hosts would generally have multicast connectivity irrespective of the network.

This means that application developers could rely on multicast to exist in most places, even if in many places the multicast packets would be encapsulated inside unicast packets. The hope is that if a service provider sees a lot of AMT unicast traffic on their network, they will realise that they can enable native multicast in order to reduce the amount of traffic. There are already many peer-to-peer applications doing multicast at the application layer, and it would also be interesting to have such applications automatically make use of multicast if available.

Management Tools

In general, there are not that many tools available for managing or debugging multicast. One tool that has been around for a long time is *mtrace* which requires router support, but is quite useful in that one can trace multicast trees and pin-point multicast problems in the network. It allows an end-user a way to debug multicast, even if it can be hard to understand mtrace output, but also allows an administrator in one network domain to detect a problem in another. This means that if inter-domain

multicast fails, one can figure out in which domain the problem is, and contact that domain's administrators. Unfortunately, not all vendors implement this, and it is uncertain whether vendors will continue to support it. The IETF has been working to standardise mtrace for several years, but the process has been very slow due to a lack of interest. However, this work is now being revived, and there is at least some hope that a standard may eventually appear.

A new tool that has recently become available is ssm ping (and asmping), which offers something that looks like the common ping (ICMP echo request/reply) command for checking various aspects of multicast connectivity. In addition to verifying connectivity, it can also be used to measure delays, re-ordering, tree establishment, and so on. The protocol is being worked on in the IETF (draft-ietf-mboned-ssmping-00.txt) and an implementation available¹⁶.

2.4 MPLS

Multiprotocol Label Switching (MPLS) is a technology that attempts to optimise IP routing. The classical IP packet forwarding model requires each router along the path from the source node to the destination node to analyse the destination address contained in the network layer header, with all of them doing so independently of each other.

This model has been extremely successful and widely deployed in the last decades, as it has proven robust in network failures and scalable in terms of the number of network nodes and links. However, under specific networking conditions, IP packet forwarding model may be not an optimal solution, and can increase the operational costs of a network. As a result, there was a need for new innovative packet forwarding techniques, which eventually lead to the development of MPLS. This can improve the price versus performance ratio of networks, improves scalability in terms of transmission speed, and provides greater flexibility in delivering network services such as *Quality of Service (QoS)* and *Traffic Engineering (TE)*.

MPLS labels and MPLS forwarding mode

As with the preceding Frame Relay and ATM technologies, MPLS combines the benefits of (Layer 3) packet switching and (Layer 2) circuit switching. As shown in Figure , MPLS assigns *labels* (or *tags*) to datagrams prior to transporting them over a packet- (i.e. IP) or cell- (i.e. ATM) based network. The generic label, a fixed-length 20-bit field, is used by MPLS-enabled routers, known as *Label Switch Routers (LSR)*, to process and switch incoming packets along the path towards the destination.

¹⁶ <http://www.venaas.no/multicast/ssmping/>

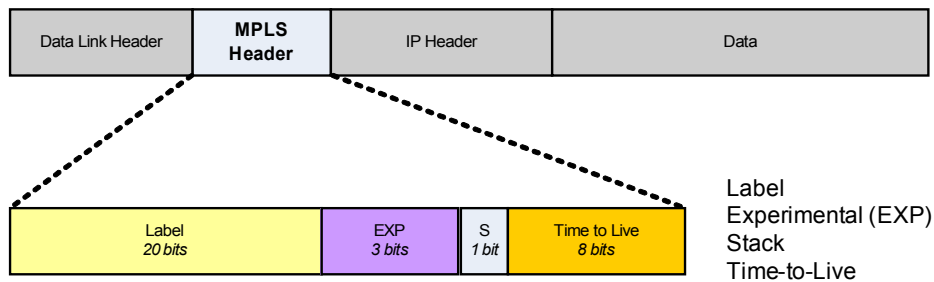


Figure 2.1: Generic MPLS label

The label of the incoming datagram allows the LSR to determine the next-hop along the destination node without performing IP routing. The label of each datagram, prior to being forwarded to the outgoing interface, is then replaced with an appropriate outgoing label, as shown in Figure 2.2.

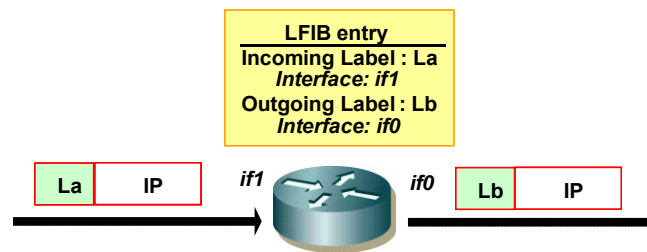


Figure 2.2: Label swapping

A label distribution protocol, such as *Label Distribution Protocol – LDB* (RFC 3096) or *Resource Reservation Protocol: Traffic Extensions – RSVP-TE* (RFC 3209), is needed for building up the *Label Forwarding Information Bases (LFIBs)* of the LSRs. Such protocols allow LSRs to discover and establish relationships with each other for exchanging label-binding information. Labels are distributed by two ways; *downstream unsolicited label distribution* or *downstream on-demand label distribution*, although both methods may be supported in a network.

As IP packet enters an MPLS-based network, the edge router known as an *ingress LSR*, associates the packet with a *Forward Equivalence Class (FEC)*, or in other words a set of packets with common characteristics (although not necessarily with the same network layer header). For example, a FEC may consist of packets that need to be sent towards the same source/destination address, or packets with the same destination prefix address and *Differentiated Services Code Point (DSCP)* bits in their header.

All packets belonging to the same FEC are switched though the MPLS core network using the same path, which is referred to a *Label Switched Path (LSP)*. The last router in an MPLS-enabled network known as the *egress LSR*, may then remove the label and route the packet using the IP header information, although the *penultimate LSR* usually removes the label and the egress LSR only performs IP routing to the incoming packets of a FEC (Figure 2).

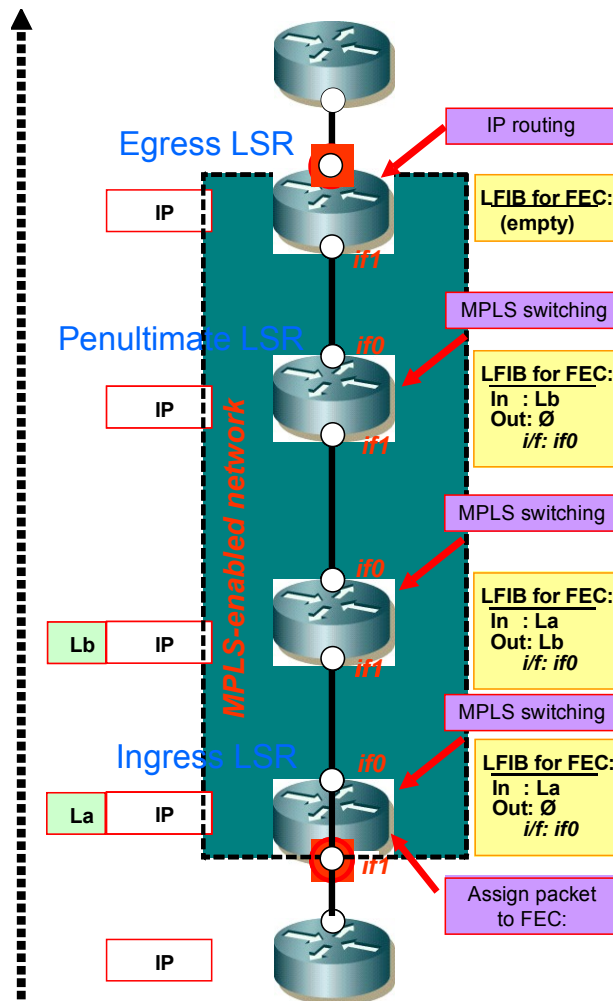


Figure 2.3: MPLS switching

MPLS Advantages

MPLS is able to reduce the complexity of transporting data in high-speed networks. IP routing needs to match the longest-prefix of the destination address (of the IP header) for each incoming packet, with the variable-length entries in the routing forwarding table (FIBs). By contrast, MPLS packet switching is based on fixed-size labels stored in Label FIBs, which are usually much smaller than IP FIBs. This reduces the complexity of hardware in core routers and minimises the use of processing and memory resources.

MPLS also simplifies the deployment of VPN services, at both Layers 2 and 3. Layer 2 VPNs (i.e. point-to-point LSPs or VPLS services) are suitable for customers that intend to interconnect small number of end-sites and are willing to take the responsibility of routing functions (control plane) on their intranet. By contrast, Layer 3 VPNs are particularly suitable for customers with a large number of end-sites who prefer to delegate the routing management to their service provider. In all MPLS-based VPN services, the customer does not need to support MPLS and requires minimal configuration changes.

Traffic Engineering (TE) is the process of mapping traffic demand onto a network. The use of MPLS LSPs and TE enhancements maximise the utilisation of network

links and minimise the service impact after a network failure, especially when pre-establish backup LSPs are available and *Fast Reroute* functions are supported by the MPLS-enabled routers. In addition, the *DSCP* field of the IP header may be used for assigning packets into specific MPLS LSPs, which allows traffic to follow specific paths based upon quality-of-service requirements.

MPLS Disadvantages

The MPLS layer is added between IP and data link layers, as shown in Figure 2.1. Consequently, introducing MPLS to a purely IP network increases management complexity and requires core routers to be upgraded to support the MPLS control plane (i.e. signalling and label forwarding tables, and tag-switching functions). The deployment of MPLS though, enhances the service portfolio of a network and makes feasible the provisioning of services that would be administratively difficult or impossible to deploy on a purely IP network.

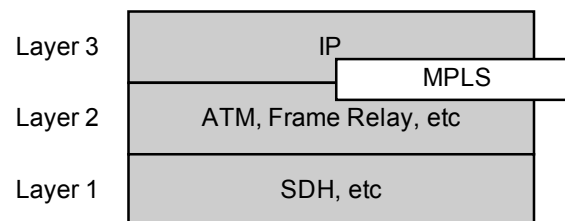


Figure 2.1: MPLS sub-layer

The provision of MPLS-based services in a multi-domain environment is feasible but, in practice, is seldom implemented in research and education networks. For example, an LSP terminating in different domains, or VPN services with end-sites in different domains require manual or semi-automatic procedures to be followed. Inter-provider VPN techniques, such as *Carrier Supporting Carrier (CSC)*, simplify the provision of services in a multi-domain environment, but increase the management complexity of the network.

Supporting multicast services in a MPLS-enabled network can also be challenging. In an MPLS-enabled network, traffic is forwarded in pre-established point-to-point LSPs. As a result, multicast traffic has to be copied to multiple identical packets, each of them forwarded via separate LSPs, which is a counter-argument for the deployment of multicast services in a MPLS-based network. In addition, extra configuration is needed for supporting multicast services, and although vendors support multicast-enabled MPLS VPNs, such services are not widely deployed.

MPLS Deployment

The decision whether or not to deploy MPLS-based services is mainly influenced by the diverse set of services that a service provider wished to offer its customers. As previously mentioned, MPLS can simplify the provision of VPN services, but it also increases management complexity and minimises the advantages of multicast services.

2.5 T-MPLS

Transport Multi-Protocol Label Switching (or T-MPLS) is a recent ITU-T standard (as specified in G.8110.1, G.8112, G.8121, G.8131 and G.8132) developed in conjunction with the IETF. It is designed to provision connection-oriented packet-based services over existing SDH/SONET, OTN and PDH networks in order to support the increasing convergence between IP and voice applications, particularly in the mobile telecommunications sector. The aim is to allow carriers to gradually transition towards packet-based services that can scale to meet future demand (data currently only represents 10% of traffic on mobile networks), whilst being able to emulate legacy services (e.g. voice) with the same QoS and reliability guarantees.

T-MPLS works in a similar fashion to MPLS (see Section 2.4), but focuses strictly on Layer 2 functionality, stripping-out unnecessary features such as ECMP, LSP merging and PHP. It also adds capabilities such as bi-directional LSPs, bandwidth allocation using LSPs, protection and restoration features, and advanced OAM support, in order to support the high availability (99.999%) requirements of carriers.

T-MPLS therefore takes an existing tried-and-tested technology and attempts to make it less complex and easier to manage, whilst supporting existing operational procedures and processes. Unfortunately, unlike MPLS, no signalling protocol has been defined although it is envisaged that GMPLS signalling will be utilised. This means that T-MPLS provisioning will likely have to rely on manual configuration for some time.

Although T-MPLS has been designed as a generic transport for packet-based services, its initial focus appears to be Ethernet. This essentially brings it into competition with PBBTE (see Section 1.2 'Ethernet'), although T-MPLS is further ahead in the standardisation process and can support other services as well as Ethernet (e.g. ATM and Frame Relay). However, there are still a number of interoperability issues to be resolved, and existing management tools and interfaces need to be updated to support T-MPLS.

It is probably unlikely that T-MPLS will have a significant role in research and education networks, as it is primarily aimed at carriers which appear to have shown limited enthusiasm for it at the present time. Given that research and education networks are less bound by legacy SDH/SONET concerns, and will likely wish to take advantage of lower Ethernet costs, they will perhaps be willing to wait until PBBTE is available (see Section 1.2 'Ethernet').

2.6 ASON and GMPLS

There is currently a great deal of effort going into the development of a standardised optical control plane. This standardisation has two key goals in mind; namely the automation of heterogeneous networks, and the specification of common protocols supported by all vendors. In the research and education community, a further goal is user control of network resources to accommodate e-science and other computation, instrumentation, and data sharing applications.

Both the IETF and ITU-T are working on control plane standardisation: the ITU-T with Automatic Switched Optical Networks (ASON), and the IETF with Generalised Multiprotocol Label Switching (GMPLS). There is also third body, the Optical Internetworking Forum (OIF), that works closely with both the IETF and ITU-T to accelerate the deployment of the control plane by vendors and carriers via implementation agreements and interoperability demonstrations.

Although the IETF and the ITU have complementary roles in the development of a control plane, in reality there are areas of conflict between the two bodies. The main role of the ITU-T is to develop an architecture specification based on the needs of service providers, whilst the IETF mainly represents vendors and develops protocols to meet industry requirements.

ASON

ASON is a reference architecture targeted at transport networks, and is intended to provide automation which is both scalable and fault tolerant. There are several key ASON standards:

- G.8080: Architecture for Automatically Switched Optical Networks
- G.7713: Distributed Call and Connection Control
- G.7715: Architecture and Requirements for Routing in the Automatic Switched Optical Networks
- G.7714: Generalised Automated Discovery Techniques

The ASON architecture is divided up into domains, which may be defined according to administrative, geographical, or vendor boundaries. Signalling and routing information is exchanged between or within domains at *reference points*, of which there are three types:

- User-Network Interface (UNI) – a reference point between an administrative domain and an end-user.
- External Network-to-Network Interface (ENNI) – a reference point between domains.
- Intra Network-to-Network Interface (I-NNI) – a reference point within a domain.

Information is exchanged in support of resource discovery, routing, connection control and selection functions. Connection routing is only applicable at I-NNI and E-NNI reference points, as no routing functions are associated with the UNI.

GMPLS

GMPLS is a suite of protocols providing control plane mechanisms such as routing and signalling. It is designed to support not only the networking equipment that performs packet switching functions (such as IP routers), but also that performing WDM and/or TDM switching (such as SONET/SDH multiplexers). The development of GMPLS was motivated by the need to have a common suite of protocols across a number of networking layers that have traditionally been managed separately (e.g. optical, SONET/SDH, ATM and IP).

GMPLS evolved from Multi Protocol Label Switching (MPLS), a connection-

oriented switching paradigm where *labels* are added to IP packets to allow for quicker forwarding. It extends this concept with a 32-bit label that also allows particular fibres, wavelengths and/or timeslots to be identified, thus allowing a path through a network to be specified. The aim is to simplify network operation and management, and provide a framework for a unified IP and optical control plane.

GMPLS also provides scalability features such as Label Switched Path (LSP) hierarchy and bundling; and Traffic Engineering (TE) based on the route computation of lightpaths according to available bandwidth, latency, jitter, and link cost. Path computation is based on each node having access to information about topology, link and node characteristics, as well as resource availability within the routing domain.

The GMPLS protocol suite provides the following functionality in a distributed manner:

- Signalling - RSVP-TE (Resource Reservation Protocol with Traffic Engineering)
- Routing - OSPF-TE (Open Shortest Path First TE) based on Link-State protocols (LSA)
- Discovery - LMP (Link Management Protocol)

GMPLS RSVP-TE is the generalised extension of MPLS-TE that separates forwarding information from IP header information, allowing for forwarding through label swapping and various routing options. Signalling is required for the establishment and maintenance of connections between network nodes, whilst traffic engineering provides performance optimisation based on available bandwidth, delay, and packet drops. This requires each node to propagate local resource and topology information to all other nodes in the network domain.

The propagation of such information is undertaken by link-state protocols, such as OSPF via LSA. OSPF is a hierarchical link-state protocol, which provides a means for it to support multiple routing areas. There are also different types of LSAs for broadcasting link and node information, or summarised information to be used for building the link-state topology database that is maintained by each node. The addition of traffic engineering (OSPF-TE) information provides a richer set of information for the topology database, and using this, each node can compute explicit paths to each available destination in order to generate the next-hop forwarding table. When a reservation request is received by a node, it uses the topology database in order to determine whether the request can be fulfilled. In addition, if the path is being computed on a hop-by-hop basis (rather than source routed), the node will then forward the request to the next node.

Another important and very powerful protocol is LMP, which is currently being standardised by the IETF. It is mainly used for neighbour discovery, although it has other features such as link parameter correlation (for checking whether adjacent nodes have consistent properties), link fault localization (for detecting faults), and control channel management (for negotiating parameters), which are extremely useful for optical networks.

More recently, the IETF has been working to address the issues of inter-domain

interoperability in multi-layer network through the specification of a Path Computation Element (PCE) architecture, and specifically the Path Computation Element Communication Protocol (PCEP). This is similar to the concept of a bandwidth broker in that it enables constraint-based path computation where complex traffic engineering is required.

At the present time though, GMPLS has not seen a great deal of deployment by carriers. There is still limited interoperability between equipment from different suppliers, and the inter-domain signalling issues are still ongoing. Although GMPLS has been deployed in some carrier networks in Japan, it has to-date mostly been utilised on NLR and other North American research testbeds. However, it is currently being trialled by Internet2 with a view to provisioning dynamic circuits for backup purposes, so is likely to become more widely adopted in the near future; if initially only in core networks using specific solutions.

2.7 Control Planes for Grids

In Grid networks, initiation of an RSVP-TE signal comes from an end-user or application making a request via Grid middleware. The idea is that once requests are made for Grid resources, the Grid resource manager and resource allocation components process and coordinate the request, effectively behaving as an RSVP-TE client. The RSVP-TE signaling then continues across the network on a hop-by-hop basis, or using Explicit Route Object (ERO) which allows a client to specify a route in order to establish a point-to-point reserved path with the requested quality-of-service.

Management Plane

Optical networks have traditionally been managed in a centralised fashion using so-called FCAPS functionality, an acronym for Fault, Configuration, Accounting, Performance and Security management. Management plane mechanisms rely on client/server model, usually involving one or more management applications (structured hierarchically) communicating to each network element in its domain via a management protocol, (such as SNMP, TL1 and XML). Information exchange between the network elements and the management application is usually via a Command Line Interface (CLI) or information base (such as MIB, RDF).

Grid Middleware

Layer 1 Grid services (sometimes known as LambdaGrids) are dependent on seamless vertical integration between applications, Grid middleware and optical networks. The Grid middleware acts as an intermediary between applications and the network, and provides APIs so that applications can take advantage of Grid features such as dynamic resource scheduling, monitoring, data management, security, information, and adaptive services. In addition, the Grid middleware has to be aware of the status of the network and successfully provision the necessary on-demand connections with deterministic quality-of-service. To solve these challenges, the optical networking community, in conjunction with the Grid community, has to rethink the role of both the management plane and the optical control planes and their interaction with Grid middleware.

Initiating on-demand connections that link end-stations, instruments, and sensors rather than edge nodes is a critical requirement of the Grid community. Such requirements cannot be met using the traditional management methods (often manual configuration) of optical networks, and control plane signaling is more capable of handling the dynamic setup and teardown of Layer 1 connections. Implementing such functions in a distributed control plane should provide a more scalable solution and speed-up setup times, whilst reducing operational costs. However, certain high-end applications provide unique challenges as they often make active use of end-to-end network connections, transfer datasets (often in the order of several terabytes) across large distances, require low jitter and low latency, have connection requirements in the order of a few microseconds to many hours, and need close to real-time feedback of network performance.

It is therefore recognised by the standards bodies that control plane solutions will need to address application-initiated optical connections, handle the interactions with Grid middleware and higher-layer protocols, as well as support inter-domain operation. The research challenge is to define the roles and mechanisms of the entities involved in operating the network, as there is currently little consensus on how to demarcate the management plane, control plane and Grid middleware functions.

2.8 Conclusions

IP routing is critical to the operation of the Internet and the protocols are mature and well understood. However, the BGP protocol used for inter-domain routing is again facing scaling issues as the Internet continues to grow.

The number of routes that must be maintained by routers located in the DFZ was around 230,000 in July 2007, and this figure is expected to grow by 60% within the next five years. Further projections suggest the number of routes could approach 2 million by 2022. It should also be borne in mind that many service providers typically configure additional routes in their routers for other purposes, so these figures can often be significantly higher.

Of course, routing is designed to operate in a dynamic fashion, which means routes need to be advertised as they come-and-go, This is undertaken with BGP updates, but maintaining 230,000 routes means that routers must currently process up to 400,000 update messages per day. The number of updates is in fact increasing at a faster rate relative to the number of routes, which is due to the increasing exposure of core routers to generally less stable edge networks because of topological changes in the Internet.

Concern has been expressed that this growth may start to outstrip the processing and memory capabilities of routers. The specialised ASICs and fast memory required by modern routers are becoming more expensive to produce, and are not tending to follow Moore's Law which predicts a doubling of capability every two years for the same cost. Whilst it is perhaps not an immediate concern as current enterprise-class routers have sufficient capabilities to process 500,000 to 1 million routes, it is an issue that research and education networks should be aware of.

The IAB and IETF have also started to investigate whether addressing and routing could be made more efficient so that it becomes less reliant on hardware scalability. This process is at an early stage and there are no specific recommendations yet. Nevertheless, current thinking seems to be heading in the direction of splitting IP addresses into separate locator and identifier elements, allowing for more localised routing. These developments are something to which research and education networks might provide some useful input.

One of the aims of IPv6 was actually to improve the aggregation of routing prefixes, as well as vastly expand the address space and offer improved support for other features. Unfortunately, even though the core protocols have been available for some time and are supported by most routers and operating systems, take-up of IPv6 has been limited.

NRENs have been amongst the foremost adopters of IPv6, and most (if not all) run dual-stack systems. In addition, some campuses support IPv6, although it is far from universally deployed even within institutions. However, even though IPv6 is widely available in the academic community, usage still remains quite low.

The main problem with IPv6 is that there has not been any urgent need for it, and no specific application actually requires it. In addition, some routers still have features that are not IPv6-enabled, whilst certain user equipment (e.g. VoIP and videoconferencing devices) does not support it at all.

This said, recent revised predictions suggest that the IPv4 address space may be exhausted within 3 to 5 years, rather than 10 to 20 years as previously predicted. This obviously makes the transition to IPv6 much more urgent than hitherto thought, and means that IPv4 address space will become harder to obtain in the coming years.

Research and education networks are perhaps well positioned in this respect, having run dual-stack networks for a number of years, and gained a lot of experience of working with IPv6. However, those campuses not yet running IPv6, should start planning how to transition their networks, including middleboxes (e.g. firewalls, intrusion detection systems), network management systems and other equipment. Although mechanisms exist to support legacy IPv4 systems, this will increase management complexity and it may simply be more painless in the long run to transition as many network systems as possible.

Traffic engineering is generally not utilised by research and education networks, although it may become increasingly important as customers and demanding users request VPN or VPLS services. MPLS was originally designed for fast packet forwarding, but it has subsequently been widely adopted by carriers for provisioning virtual circuits over their IP networks. Unfortunately though, whilst MPLS is supported by most enterprise-class routers today, inter-domain operations either require manual configuration or additional protocols to be run, multicasting is problematic, and it makes management of networks more complex. Furthermore, it is now possible to establish virtual circuits using L2TPv3 (Layer 2 Tunnelling Protocol - Version 3) or even lightpaths established over optical wavelengths.

The decision whether to deploy MPLS is therefore dependent on the services a network needs to provision. A few research and education networks currently utilise it to support specific customer requirements, although there is an increasing trend towards provisioning private networks over lightpaths.

The flexibility to offer lightpath services has been made possible as research and education networks increasingly move from providing IP-only services towards running the underlying optical infrastructure as well. Automated management of optical equipment still remains very limited though, which makes it difficult to dynamically configure circuits at the optical level.

Whilst GMPLS promises the ability to integrate IP routing and WDM (and SDH/SONET) control planes, its development has been a long time coming and it has only recently started to be deployed. However, there are still significant signalling and interoperability issues to resolve, stemming from the differing requirements of the Internet and telecommunications sectors. Another issue to resolve is whether peer (where all layers in the network are managed as one domain) or overlay (where the IP and optical layers are treated as separate domains) configurations should be used. The peer model offers a more integrated approach to network provisioning and may be more appropriate for research and education networks where all elements of a network are under their control. Nevertheless, the overlay model is simpler and may be appropriate where only limited reconfiguration of optical circuits is necessary.

GMPLS is starting to be deployed in Japan and North America, although largely using vendor-specific solutions. It is expected that this trend will continue, although deployment will likely initially be limited to specific uses on certain networks.

3. Network Virtualisation

Virtualisation in one form or another exists in many areas of computing and networking, where it commonly refers to the abstraction of a resource (in a broad sense) by detaching its properties from any particular physical representation. In platform virtualisation, this is generally translated to an abstraction layer (implemented in software) being introduced between the actual hardware, and the operating systems and applications running on top of it. In the case of resource virtualisation, the abstraction layer may federate various types or system resources (storage, networking, processors, etc.) that may not be located within the same physical machine.

From a systems design perspective, one of the benefits of virtualisation is that higher-level entities are provided with a stable and unique interface to the virtual resource. The mapping of its properties to various physical representations can be isolated and need only be implemented once.

From an operational point of view, the cost of adding an additional virtual instance of a resource is often marginal compared to expanding the underlying physical infrastructure. A typical example in computing is a set of virtual machines that share the resources of a single physical machine and appear to be independent systems to the applications that use it.

In networking, the physical resources are essentially links, switches and routers, and current systems typically support virtualisation from Layer 2 upwards. For example:

- The VCAT and LCAS specified by ITU-T for the next-generation SONET/SDH standards essentially follow a virtualisation paradigm by allowing a larger bandwidth circuit to be partitioned onto multiple lower-order circuits with capacities that can be adjusted on-the-fly by the network manager.
- The Link Aggregation standard (IEEE 802.3ad) that allows multiple Ethernet links to operate together, thus providing a higher bandwidth pipe.
- At Layer 2, a *Virtual Leased Line* (VLL) provides a similar service as a physical cable between two points of the network. Conceptually, it can be thought of as a tunnel that transports Layer 2 frames within frames of a higher layer (e.g. MPLS or IP in the case of a procedure like L2TP).
- At Layer 3, a *Virtual Private Network* (VPN) connects independent IP networks across a network shared with other traffic. The packets belonging to distinct virtual networks are marked with some kind of tag (e.g. a MPLS label), and the routers in the shared network have to be aware of the virtual networks to be able to route them properly. This is done by running separate instances of the routing protocol for each virtual network, known as *Virtual Routing and Forwarding* (VRF).
- At higher layers, peer-to-peer networks apply de-facto virtualisation technologies by building overlays on top of the current Internet and providing a unified image to the end-user.

VLL and VPN are standard services provided by most ISPs today. Tunnels that use encapsulation in IP packets make it possible to build *overlay* networks on top of the existing networks between arbitrary locations. This technique is used extensively in projects dedicated to network research such as the PlanetLab platform¹⁷.

Following the influence of the US-based GENI initiative¹⁸, a new type of infrastructure for network research is emerging that takes the concept of virtualisation to a new level by extending it down to the physical layer, allowing specific parts of the physical substrate to be allocated to a virtual resource. For example, a virtual machine may obtain exclusive control over a CPU from a pool of available CPUs on a physical host, rather than being multiplexed onto a CPU with other virtual machines. A virtual router should also be able to allocate enough physical resources (route-processors, interface cards etc.) so that it can perform its function at wire-speed, and it should equally be possible to assign entire physical links (including a particular wavelength within an optical transport system) to a particular network slice.

While giving users more control of the infrastructure that makes up their virtual environment, this new paradigm also largely removes the isolation between the virtual and physical layers, because at least some of the physical properties have to be exposed to the manager of the virtual resource. One of the big challenges is that, in general, all devices that eventually make up a network slice either need to support virtualisation themselves, or provide means to provision them accordingly. This is vastly different from pure overlay networks like PlanetLab, where only the nodes themselves are virtualised. However, multi-core server-grade CPUs currently support advanced virtualisation features, and network interface cards that include virtualisation primitives emerged on the market at the end of 2006¹⁹. High-end commercial routing platforms also allow multiple router instances running in parallel to be defined on the same management module.

It is envisaged that virtual network slices may be created on-demand and may have very short lifetimes, which implies there needs to be a management framework which can control all devices that are part of the physical substrate. Some parts of this problem space could be covered by existing or emerging technologies (e.g. UCLP or GMPLS), but a framework that covers all levels of virtualisation and all types of devices does not yet exist.

Another concept that is crucial for an infrastructure supporting network research is programmability. For example, a project that investigates new routing and forwarding algorithms must be able to deploy code in the virtual routers of their network slice. It may even require programmable hardware components if performance is of concern. Software routers in traditional virtual machines allow the former, but not the latter, and current off-the shelf commercial routers allow neither.

These requirements are far from trivial. Some of the features, like light-paths or any other kind of virtual circuit, overlap with the demands from certain communities. Other features like programmability of all sorts of devices appear to be

¹⁷ <http://www.planet-lab.org/>

¹⁸ <http://www.geni.net/>

¹⁹ http://www.netxen.com/news_events/press2006/pr060918.html

mainly of interest to the research community and will most likely not be part of what will be the future Internet. This is the reason why, for example the GENI project, includes designs for devices that offer a variety of functionality.

3.1 UCLP

User Controlled LightPaths (UCLP) is a network virtualization framework upon which communities of users can build their own services or applications. It allows them to do so without having to deal with the complexities of the underlying network technologies, yet still maintains the functionality that the network provides. The system is based on Service Oriented Architecture (SOA) where Web Services and Web Service Workflows are the basic building blocks.

The UCLP software (currently at version 2) provides an alternate network provisioning process that allows users to control their own packet- or switched-based network, including topology, routing, virtual routers, switches, virtual machines and protocols. The UCLP concept consists of many separate, concurrent and independently managed Articulated Private Networks (APNs) operating on top of one or more network substrates across different ownership domains. An APN can be considered as a physically isolated or *underlay* network where a user can create a customised multi-domain network topology by binding together network links at layers 1 through 3, computers, time slices, and virtual or physical routing and/or switching nodes. This capability is realized by representing all network elements, devices and links as web services, and by using web services workflow to allow the user to bind together their various web services to create a long-lived APN.

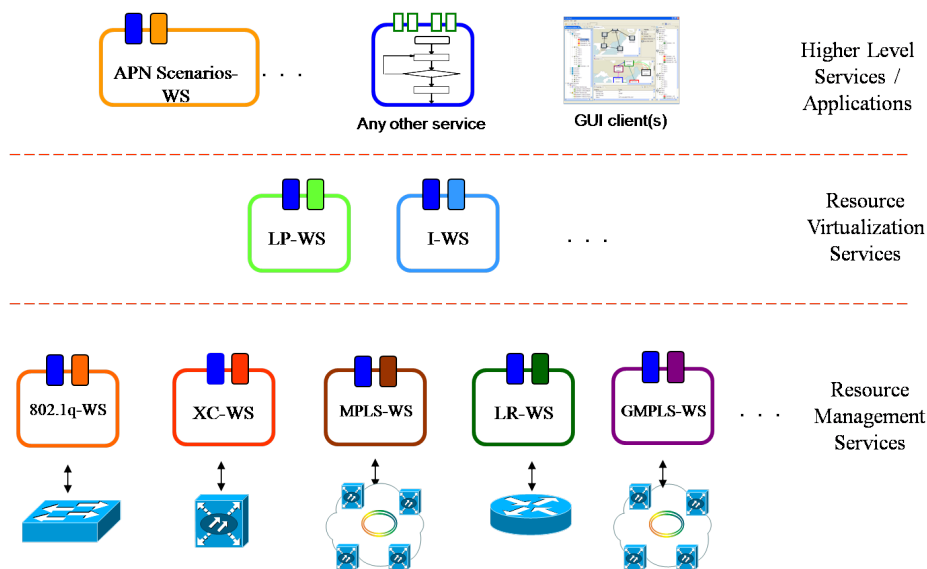


Figure 3.1: UCLP Service Oriented Architecture

The UCLP service oriented architecture is depicted in Figure 3.1, where each box represents a different service. The Resource Management Services or Network Element web services, is a group of services that manage and control the resources on a physical device; each single web service dealing with a different technology. The

Cross Connect Web Service (XC-WS) controls devices that perform cross connections (TDM, wavelength, fibre); the GMPLS Web Service (GMPLS-WS) deals with GMPLS networks; the 802.1q Web Service (802.1q-WS) manages VLAN-enabled equipment; the MPLS Web Service (MPLS-WS) controls MPLS networks and the Logical Router Web Service (LR-WS) controls logical as well as physical routing platforms. Resource Virtualisation Services provide a layer of virtualisation so that the technology of the physical devices is abstracted, but none of the features that these devices offer are lost. A LightPath Web Service (LP-WS) is the partition of a physical link, whereas an Interface WS (I-WS) is the partition of a physical endpoint. Finally, higher-level services or applications exploit the virtualisation capability just described to build complete end-user solutions or other services without having to deal with the underlying network complexities. Any of the service groups presented in the architecture can be extended to support new technologies, or to provide new capabilities in the virtualisation layer.

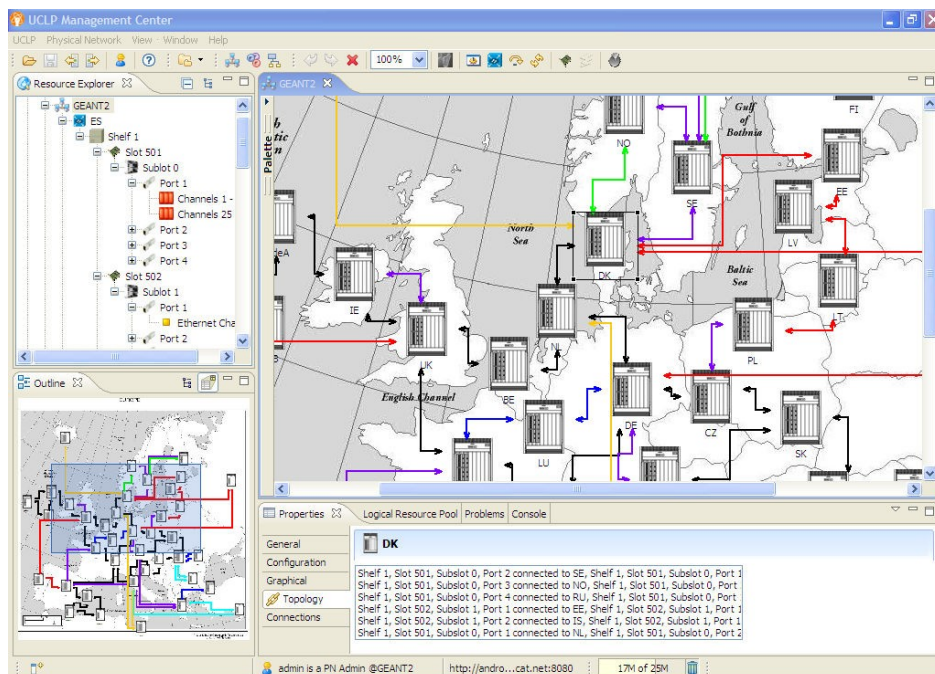


Figure 3.2: UCLP Management Center

With UCLP, advanced users can become administrators of *underlay networks* or APNs built from resources obtained from a number of Physical Network administrators (PN admins or network operators) or other advanced users. This is why they are known as Articulated Private Network Administrators or APN Admins.

APN Admins can talk to different PN Admins or other APN Admins to obtain all the resources (LP-WSs and I-WSs) they need via resource lists, and these together, they can create and reconfigure different topologies over one or more physical networks. APN Admins can also lease part or all of their resources to other APN Admins for a period of time, and can acquire new resources from other providers at any time. Finally, APN Admins can create and run new (web) services that work on top of their resources without the provider even noticing (such as the APN Scenarios service).

A initial implementation of the UCLP is being deployed on CANet 5, the Canadian NREN, and it is currently being used by Environment Canada, a federal government department, to enable an APN linking its research facilities across the country. Planned future work includes adding support for SDH, VLAN-WS (Ethernet with VLANs), VCAT (Virtual conCATenation) connections and extending the implementation to other technologies such as Reconfigurable Optical Add and Drop Multiplexers (ROADMs), virtual or logical routing, and GMPLS interoperability. The UCLP developers, Inocybe Technologies Inc., CRC and i2CAT have established Argia²⁰ to develop a commercial product based on UCLP technology. However, a UCLP Community Edition will be made available to research and education institutes as open source software, and the suite will continue to be extended through open-source plug-ins. There are also plans to offer free Argia licences to research projects wishing to extend or experiment with the software.

3.2 Overview of Network Virtualisation Projects

A comprehensive Layer 1 to 7 approach to the virtualisation of network resources has yet to be introduced. Network services often require the use of an out-of-band control plane for provisioning, management and monitoring. In addition, the validation of new protocol designs is much better undertaken in an environment that can provide real or emulated traffic, rather than simply testing on a software-only simulator. The demand for such functionality was acknowledged by the research community and is being addressed by several projects throughout the world.

PlanetLab

PlanetLab²¹ was probably the first such initiative chronologically. It introduced the concept of *slices* to define a particular set of virtual resources a researcher is allowed to control, although these are ultimately constrained to CPU cycles and memory, and bandwidth is not covered by the virtualisation engine (other than implementing capping techniques to reduce abuse). Slices are allowed to communicate between each other to allow various categories of higher layer services to be offered within the virtual infrastructure. Examples of service categories include²² content distribution, storage and large file transfers, routing and mobile access, multicast, e-mail, measurements, and management, etc.. Each of these categories are illustrated by various implementations of particular services for supporting different research activities

Whilst the overlay-based PlanetLab infrastructure requires no special support from the underlying network, some limitations have surfaced that appear to hinder certain types of network research within such systems. For example, artefacts from the underlying network can distort the results of the experiments (e.g. unrealistic error rates, fluctuating throughput, or timing issues); novel link technologies cannot be adequately tested at this level of virtualisation; and virtual resources are usually implemented as a layer of software which can incur performance penalties.

²⁰ <http://www.inocybe.ca/argia.html>

²¹ <http://www.planet-lab.org/>

²² <http://www.planet-lab.org/files/presentation-2007-05-01-planetlab.ppt>

GENI

The US-based GENI initiative²³ to be funded through the National Science Foundation is expected to extend the PlanetLab idea by virtualising the entire network stack. In order to achieve this goal, GENI would allow physical resources of the virtualised infrastructure to also be placed in the core of the network, as opposed to the edge-only approach of PlanetLab. As such, GENI will have to build a brand-new wide-scale infrastructure in parallel to the Internet in order to achieve its goals. The entire process of designing the hardware platforms, planning the deployment and developing associated software is a work in progress. Recent results²⁴ seem to suggest that a GENI node would be based on an ATCA chassis that would house different types of boards (including general processing, network-processing and packet forwarding-oriented line cards). A virtualisation layer will be implemented in software in order to provide network researchers with a uniform abstract layer on top of the infrastructure.

OneLab

The European project OneLab²⁵ aims to extend PlanetLab beyond the standard wired Internet environment, federate the European deployments of PlanetLab nodes, and provide increased monitoring capabilities. Plans indicate that a monitoring infrastructure will be placed at the core of the network, and topology monitoring will be developed as a OneLab service. The federated structure (basically, uniting independent PlanetLab deployments in Europe) will allow the project to set policies in accordance to European research priorities. It is expected that SMEs would use OneLab services as a realistic environment in which to experiment with novel protocols and services.

FEDERICA

The FEDERICA project proposal submitted under the European Union's 7th Framework 1st Call for Proposals, plans to expand the services provided by NRENs to support research projects and experiments related to new Internet architectures and protocols. FEDERICA aims to create a versatile and scalable *technology agnostic* network infrastructure on which disruptive experiments can be undertaken. It will develop a tool bench to leverage existing virtualisation and network control mechanisms and extend these to support multi-domain management and monitoring. Although FEDERICA may make use of native virtualisation technologies at Layers 1 and 2 in order to separate traffic, the virtualised infrastructure is mainly expected to support mainly Layer 4 to 7 applications. In addition, whilst based on a slice concept similar to PlanetLab or GENI, it is anticipated that FEDERICA will allow dynamic definition and allocation of both core and edge network resources within a particular slice.

MANTICORE

The MANTICORE (Making APN Network Topologies on Internet COREs) project, led by HEAnet and i2CAT with RedIRIS as a partner, started its second phase in July

²³ <http://www.geni.net/>

²⁴ <http://www.geni.net/docs/gbp-bwg-6-07.pdf>

²⁵ http://one-lab.org/pub/OneLab/OneLabPresentations/Onelab_Pres_Templ_Helsinki_F3_Concert.ppt

2007. This project is working to define guidelines and specifications for a system based on web services that offers APN (Articulated Private Networks), LR-WS (Logical Router Web Service) and peering services to end-users, allowing them to create their own logical IP networks over a physical network. The initial phase of the project will focus on defining the specifications for establishing and controlling logical IP networks using UCLP, which will then be implemented in an experimental architecture.

MANTICORE therefore extends existing UCLP technology by adding the ability to manage Logical IP networks and the equipment supporting them by means of web services. This will allow end-to-end logical IP networks (routers as well as links) to be configured independently of the underlying network topology.

The project aims to define its requirements in an open manner, in order to allow their implementation using any manufacturer's device. The intention is for the LR-WS and Peering-WS to manage the physical devices and represent these as abstract logical routers to the UCLP middleware, with an XML API and NetConf being used to control both the logical devices and the physical host equipment. Another aspect will be the IP Network Web Services that will represent to the end-user the IP services offered by the UCLP middleware. This component will represent a device-independent IP service and its associated configurations, hiding specific details about equipment and topology.

3.3 Conclusions

Virtualisation techniques have existed in the lower networking layers for quite some time, but such techniques have recently made inroads towards the higher layers of the networking stack and percolated from the data to the control plane. Other resources (e.g. CPU and storage) may also be managed through virtualised management frameworks. Network service virtualisation therefore promises innovative solutions that are adaptable to a rapidly changing end-to-end environment, which can be deployed faster, and with more effective management and troubleshooting.

Basic virtualisation implemented in certain modern routers enables in-service upgrades of the equipment (including management software) with no service interruption, whilst providing the ability to fallback to a known stable setup in case of problems.

NRENs pioneered the customer-empowered network concept, and virtualisation takes this concept to the next level. It has the potential to enable multiple networks, defined and managed by customers, to run on the same underlying infrastructure provisioned by the NREN. Deployment of UCLP and similar technologies are a first step on the road towards full network virtualisation.

The testing of new disruptive technologies over a production infrastructure is considered to be the Holy Grail within research and education networking. Full network service virtualisation will make a significant advancement in this direction, allowing the research community to experiment on real, widespread infrastructure without worrying about disrupting other operations.

Network service virtualisation as proposed by the GENI and FEDERICA projects goes beyond the standards for Next-Generation Networks specified by ITU-T. They aim to extend the network beyond the operator domain to the campus, and finally to applications running on end-user computers. As other services become more commoditised, research and education networks will need to observe and incorporate virtualisation in order to maintain a competitive edge over both traditional telecommunications operators, and emerging infrastructure providers based on Web 2.0 technologies.

4. Operations and Performance

Most research and education networks currently offer a variety of network services, the number of which has proliferated over the years. They usually also offer agreed levels of service for specific types of traffic, whether through guaranteed bandwidth provisioning techniques (such as Premium IP or optical lightpaths) or through more general service level agreements.

As a result, it is becoming increasingly necessary to develop techniques to manage these services, as well as to monitor them to ensure that they are being delivered, and that their performance is within acceptable tolerances. Unfortunately, certain services are not always being adequately delivered to campuses and/or end-users, nor does their performance always live-up to expectations despite the apparent availability of bandwidth.

There are several reasons for these problems, but they are often due to bottlenecks in a network that has to be traversed somewhere, and/or because of restrictive policies implemented at a local (usually campus) level. Individual networks can only be held accountable for the edge-to-edge performance of their own network, but there is clearly a need for improved coordination and more integrated monitoring of all the elements in a chain in order to identify where the problems lie. More specifically, there is increasing demand for improvements in the end-to-end experience, which means that common interdomain approaches need to be agreed for provisioning and accessing such services. In some cases, it may be possible to resolve the issues through improved management, but in other cases new protocols need to be developed and adopted.

The widespread introduction of optical networks at NREN operational level also brings new challenges. The NREN community has a lot of experience with IP routing, how to manage multi-domain networks is well understood, and there are plenty of tools for doing so. By contrast, there is much less experience of optical switching, particularly between different management domains, and this issue will become more complex as switching becomes increasingly automated. There are currently limited tools for managing switches, managing other transmission equipment, or monitoring lightpaths, nor much in the way of peering policies between optical domains.

4.1 Quality of Service, Overprovisioning, or something in between?

There is a long-standing debate in the Internet community about the usefulness of quality-of-service differentiation mechanisms in the face of increasing network capacity (but also increasing utilisation). Quality-of-Service (QoS) mechanisms have been a main topic of the joint experimental work between NRENs under the auspices of TERENA and DANTE, where ATM, IntServ/RSVP, and diffserv were investigated.

Almost exactly ten years ago, in August 1997, Van Jacobson presented the basic architecture for Differentiated Services (diffserv) and the original *premium* service at the Munich IETF. The diffserv concept garnered enormous interest and was implemented in most routers over the following few years.

The Internet2 community in the US started diffserv deployment very early with the QBone project. Eventually their effort to deploy a *Premium* service was abandoned in frustration due to router hardware limitations, operational overhead, and the absence of a workable inter-domain deployment model.

In Europe, diffserv deployment started at a slower pace, but benefited from more modern hardware and the experience of others. There is currently a working multi-domain *Premium IP* service on GEANT, although the number of NREN domains fully supporting it is still relatively small. In addition, while there is a multi-domain capable provisioning system with advance reservation functionality, it still requires some manual processing.

Whether this system will experience uptake will depend on whether people see value in it. The current GEANT backbone has sufficient capacity everywhere that congestion is not an issue, although there have been events involving multiple link failures where significant performance benefits could be measured for Premium IP (test) traffic with respect to normal *best-effort* traffic.

It is therefore worth considering a few scenarios where QoS (mainly in the form of diffserv) can make a difference, as well as at alternative (non-differentiating) methods to improve network service.

Environments that enable adequate provisioning

QoS mechanisms are a means of maintaining performance (for a part of the traffic) when the network is congested. An alternative approach would be to ensure that there is never congestion. This would make QoS mechanisms unnecessary and improve performance for all traffic.

An operator who wishes to ensure uncongested operation, will need to provision capacity significantly above normal usage in order to accommodate traffic peaks and other unusual situations such as link failures. This means that average link utilisation will be relatively low. Running backbone links at low utilisation runs counter to established optimisation goals in the telecommunications industry and is frequently called *overprovisioning*, even where it makes economic sense (i.e. where bandwidth is relatively inexpensive and the cost of maintaining QoS mechanisms are high). The following elements make generous provisioning interesting for a network operator:

- On the cost side, upgrading links where necessary, is possible and affordable. This is typically easier for operators who run their own transport network than for those who have to lease capacity from another carrier.
- On the revenue side, user contributions roughly match the costs of maintaining the associated infrastructure. Capacity and usage-based components provide incentives for the provider to keep the infrastructure well-provisioned, and for the user to keep low-value traffic low, particularly in times of high usage.

Traffic Engineering

As with differentiated-QoS mechanisms, Traffic Engineering can be understood as a means to provide good service on a tightly-provisioned network. The method overrides the decisions of normal *shortest-path* routing, in order to better distribute traffic load over the backbone.

Traffic Engineering is widely used on large networks, and can be used to increase utilisation of the infrastructure and to help defer upgrades. Other than its operational cost though, there is a performance cost related to the longer-than-shortest paths selected by traffic engineering. End-to-end performance often depends on round-trip-time, especially for interactive applications, so this should be taken seriously.

An interesting alternative to traditional traffic engineering would be to use different network-wide routing policies for different (diffserv) traffic classes. Some vendors have started to support this in the form of *multi-topology routing*. One set of traffic classes could use routing metrics that optimise delay (using metrics based on link propagation times), others bottleneck bandwidth (using inverse link capacities as metrics), and yet others cost (using metrics that favour underutilised parts of the network).

"Non-elevated" Services

Usually when one considers QoS mechanisms in general, or diffserv in particular, one thinks about giving packets better treatment than the default *best-efforts* service. These *elevated* services such as Premium IP have proven difficult to deploy as technically they require implementation throughout the network, including tight control (policing/re-marking) of incoming traffic on all network borders.

There is a case for *non-elevated* services, in which packets are not (strictly) treated better than normal best-efforts packets. One example of this is *LBE* (lower-than-best-effort) traffic, which can be used for time-insensitive bulk transfers. On relatively well-provisioned networks, LBE traffic typically sees identical performance as normal traffic, but could be *sacrificed* when capacity becomes limited (e.g. due to link failures or capacity planning issues). LBE has been deployed on the GÉANT2 backbone, although it is currently not heavily used, presumably because there is a lack of incentive to do so. Since significant levels of LBE traffic helps capacity planning and make the network more robust, operators may want to encourage its use by offering incentives either through reduced charges or other means such as separate rate limits.

QoS as a DoS protection mechanism

It has been proposed that elevated services such as Premium IP could be used as an insurance against total loss of service in times of heavy denial-of-service attacks. However, most targeted denial-of-service attacks now try to overwhelm the packet processing capacity of routers, firewalls, or end-hosts, rather than filling-up links. In these situations preferential traffic handling has only limited benefits.

On the other hand, the classification and policing functions that have been introduced

in routers to support diffserv have been extended to provide protection for router *control-plane* components. These mechanisms can be used to make the network infrastructure itself much more robust against targeted attacks.

Application-based differentiation, DPI, and network neutrality

Network providers have started to deploy devices that recognise particular applications, often using payload inspection (Deep Packet Inspection, DPI). Applications can then be mapped into service classes and rate-limited or prioritised according to configuration. This is reportedly being used by ISPs to limit the amount of peer-to-peer traffic on their networks, which they feel compelled to do because the growth of this traffic undermines the ISPs capacity-planning projections. Other providers prioritise traffic only for their own value-added services such as IPTV (IP Television) or VoIP (Voice over IP).

These types of service differentiation raise some important issues. Blocking/throttling some applications and preferring others solely based on the provider's interest undermines the network's *neutrality*, which has been an important basis for the innovations that have greatly improved the utility of networks over the past decades.

In order to ensure continued innovation, it seems preferable to have usage-based components in the charging scheme, so that when user demands exceeds the provider's predictions, the provider can upgrade the network rather than having to throttle the new demand. Differentiated services should therefore be offered in an application-agnostic and non-discriminatory manner.

Lightpath Services

Where research and education networks are run over dark fibre, an alternative approach to providing guaranteed bandwidth is to provision it using lightpaths. These are SDH/SONET or Ethernet circuits that are provisioned over dedicated wavelengths on a WDM system, allowing customers or users to run their own services.

Lightpaths can offer very large amounts of dedicated bandwidth (typically 1 to 10 Gbps, depending on available interfaces), so tend to be more suitable for customers and users with demanding requirements such as those in the high-energy physics and radio astronomy communities. In addition, configuration is currently a somewhat manual procedure and therefore really only suitable for customers with relatively static requirements. This said, there are ongoing initiatives (e.g. UCLP) to simplify the process of setting-up lightpaths, and to allow lightpath-based networks to be created on an ad-hoc basis.

4.2 IP Management Issues

The operation and management of IP-based networks has evolved somewhat independently of the management of traditional telecommunications networks. The result is that IP-based networks have their own management protocols, tools, and operational traditions. Some interesting questions with respect to this are whether there will be *convergence* of management practices between IP networks and telecommunications networks, and if so, what will the converged practices resemble?

Network Monitoring

The Simple Network Monitoring Protocol (SNMP) developed by the IETF, has been almost universally adopted for managing, and in particular monitoring, IP networks. The *Management Information Base (or MIB)* of observable objects in network devices has been growing steadily, and many generic and special-purpose tools are available. Many operators rely heavily on open-source tools such as MRTG, Cricket, NMIS, etc.. and adapt those tools to their operational needs.

As research networks implement their own optical transmission capabilities, significant gains can be obtained by integrating management of optical (WDM) systems into existing IP network management systems. This would contrast with the *silo* management approaches often found in telecommunications, where each network (layer) comes with its own management system, typically handled by separate staff.

With a wealth of information easily accessible, there is an issue of information overload, where information about significant events is collected, but fails to be acted upon. This can be addressed by through better visualisation methods, making important symptoms apparent even in a sea of information, In addition, more automatic alerting can be used based on operator-defined thresholds, or by dynamic detection of deviations. Both approaches have been used with some success, but there is still significant potential for further development.

Network Configuration

Management of device configurations is another important aspect of network operations. Although SNMP supports this as well, with many MIB modules defining configuration objects, operators don't really use SNMP for configuration management. Rather, they use remote login protocols such as SSH or Telnet to access the Command-Line Interface (CLI) provided by their device vendors. There is widespread use of simple tools (e.g. RANCID) that periodically retrieve device configurations and store them in a revision management system such as CVS, for easy access, comparison and safeguarding. However, variations in CLI input and output syntax make it difficult to write and maintain tools that require a deeper understanding of configurations.

The IETF has therefore defined the NETCONF (Network Configuration) protocol to address some of the configuration management issues faced by operators. NETCONF standardises operations (show, copy, edit) on complete and partial device configurations. The configurations are assumed to be in Extensible Markup Language (XML) form, which allows vendor-specific configuration schemas to be specified in standard human- and machine-readable forms, This in turn facilitates tools that can operate on configurations in an intelligent manner.

Looking beyond protocols and tools to access configuration, there are many larger issues concerning configuration: How should the desired network-wide configuration be defined? Can it be specified completely at a high level of abstraction, - should it simply be the collection of individual device configurations, or some combination? How can this network-wide configuration be mapped to individual devices, taking into account their respective capabilities and restrictions? How can the network-wide

configuration evolve towards the ideal (a moving target), in the presence of device additions, removals, and upgrades, and in a minimally-disruptive way?

Management by Box

It is always attractive when an operational issue can be addressed by deploying a specific-purpose device. Solving a persistent problem with a one-time investment has appeal to operators and managers alike, and of course even more so to the purveyors of the magic boxes known as *middleboxes*.

This approach has been used successfully for IT services that used to be built on top of generic machines and operating systems. For example, file server *appliances* are highly optimized single-purpose packages that hide the messy details of RAID configuration, setting up network protocols, and tuning the operating system.

In networking, the prime examples of such helpful appliances are the abundant firewalls and NATs (Network Address Translators). More recently, the industry added rate-shapers (WAN optimisers, service control engines) and intrusion detection/prevention devices. Typically, such devices sit on the exterior boundary of a (campus) network, so that all traffic leaving and entering the organisation can be processed by a single device. The devices are often designed to be *transparent*, in that hosts and other network devices aren't supposed to know that they even exists.

Firewalls are ubiquitous among campus networks, but whilst NAT devices are less prevalent, but they do exist in the research and education networking as well. This may seem somewhat surprising, as NRENs usually have the necessary resources to obtain sufficient addresses for their members, and lack incentives to artificially keep them in short supply. However, from a customer's point of view, NATs provide greater liberty for managing addressing inside the (campus) network, and reduced dependence on the provider. In addition, NATs perform some firewall-like functions without configuration.

The performance of a forwarding middlebox such as a firewall, NAT, or rate shaper obviously limits the capacity of the path it is deployed on; which typically means the entire external connection of an organisation. The cost and (un)availability of high-speed middleboxes often causes organisations to forego or delay upgrading their network connection. Even where a middlebox has sufficient capacity for the offered load, transient effects such as periodic bookkeeping jobs can have adverse effects on performance.

Middleboxes frequently interfere with higher-layer protocols in unexpected ways: Firewalls and NATs rewrite some protocol values (such as TCP sequence numbers), but fail to do the same where more recent protocol extensions are used (such as in SACK blocks). Firewalls defragment packets for analysis, sometimes breaking MTU discovery by end-hosts in the process, whilst rate shapers slow down important traffic that happens to use port numbers configured as *evil* by the operator. Such malfunctions are difficult to diagnose, because often the middleboxes work in *transparent* mode and cannot be located using tools like traceroute.

Middleboxes, in particular firewalls, often impact availability, which is the most important performance aspect for many users. If there is only one instance of the box, then this is obviously a single (or additional) point of failure. Redundant configurations often introduce significant disruption during fail-over, because connection state has to be rebuilt for all connections passing through. In addition, many middleboxes need to look at both directions of traffic, which prevents asymmetric routing from working. Reliably preventing routing asymmetries makes configurations more complex, further reducing overall robustness.

A more fundamental problem with middleboxes is that they often hamper evolution in the network. New protocols at the transport or application layer can no longer be deployed just by mutual consent of the end-systems, but have to be authorised by firewalls along the path. Countless person-years have been spent defining *NAT traversal* techniques that allow devices behind NATs to be contacted by other devices (possibly also behind NATs).

For new features at the network layer, the presence of an unsupportive middlebox is a very effective means of preventing deployment. It is one of the most common roadblocks for IPv6 and multicast.

While some types of middleboxes (rate shapers) seem to be fading in popularity, firewalls are certainly here to stay. Some of the problematic aspects can be addressed by moving middleboxes further-out towards the edge of the network, although the challenge here is to allow some level of centralised control with distributed enforcement. An alternative possibility, and one which has gained more credence recently, is that those functions could even be moved back to the hosts. This presupposes that host operating systems are becoming more secure, especially with respect to their default configurations. This is desirable anyway, and there are encouraging signs in that direction.

As long as middleboxes are prevalent and deployed on the uplink of internally diverse organisations, (controlled) circumvention measures will be used. These measures can include (authorised) tunnelling, Virtual Private Network (VPN) services, laboratory or *de-militarised zones* (DMZ) for machines requiring non-standard levels of external access, or the construction (e.g. using lightpaths) of private *walled-garden* networks that may bypass the middleboxes. New applications have to go through great lengths in order to just work reasonably in the presence of different types of middleboxes; witness the many variants of encapsulations used by Skype traffic, or the extensive NAT-traversal work spawned by SIP (Session Invitation Protocol) for VoIP.

In the longer term, it should be possible to domesticate middleboxes through explicit protocol support. This is an area of ongoing work, and some approaches include Microsoft's Universal Plug and Play (uPnP), the IETF MIDCOM proposals, and more recently, the NUTSS architecture from Cornell²⁶.

²⁶ <http://nutss.gforge.cis.cornell.edu/>

4.3 IP Monitoring

The analysis of network traffic is challenging because both service providers and users have not perceived traffic monitoring as a mandatory activity, hence network devices usually offer little or support for it. Those users who do need to monitor network traffic, are then forced to either buy additional monitoring cards to plug into network equipment that are rather costly, available only on high-range equipment, and limited in terms of functionality. In addition, standards for monitoring network traffic such as IPFIX²⁷ have not yet reached RFC status. despite the fact that an IETF working group has been actively working on this for more than five years. Furthermore the SNMP community continues to focus on element management rather than on traffic monitoring, and the only RFC for traffic monitoring remains RMON-2 (RFC2021) that is more than a decade old.

Network administrators who want to analyse network traffic are therefore forced to accept the limitations of monitoring equipment, or to develop custom monitoring applications. Up to a few years ago, when many networks were running at 100 Mbps, it was usually possible to use commodity hardware when running monitoring applications, with rare exceptions such as when precise packet timestamp was required. However, because network speed continues to outpace both CPU and memory speed, it is necessary to use some hardware acceleration in order to analyse traffic at speeds of 1 Gbps and beyond. In particular, commodity PCs with standard operating systems are unable to analyse traffic at faster line rates.

Endace²⁸ is a New Zealand company that pioneered the development of network accelerator cards, and with its new OC-768 (40 Gbps) monitoring it remains the market leader. The first generation of cards focused on delivering 100% of all packets to the monitoring applications, by off-loading the computer CPU by means of the network accelerator card. This model worked for some time, when the network speed was of 1 Gbps, as the cards were efficient enough to deliver the whole packet stream to the applications while leaving free CPU cycles for both filtering and analyzing packets. At current network speeds though, software applications cannot cope with the volume of data and the depth of analysis required, hence in the latest network cards it is possible to have onboard packet filtering so that only those packets that pass the filtering rules are passed to user-space applications.

Other companies such as Napatech²⁹ are competing against Endace in the same market. They have solutions that can monitor networks at speeds of up to 10 Gbps, but they operate on the similar principle of accelerating packet capture by means of a specialised card.

Although the solutions listed above seem to satisfy the market requirements, they tackle only one side of the problem, namely packet capture. In fact the cards do not offer much in terms of network stream analysis (e.g. NetFlow/IPFIX analysis) and deep packet inspection, so that network application developers need to use additional cards (if available) for developing a fully functional high-speed network monitoring application. Furthermore packet-filtering facilities often limitations in terms of

²⁷ <http://www.ietf.org/html.charters/ipfix-charter.html>

²⁸ <http://www.endace.com/>

²⁹ <http://www.napatech.com/>

programmability (often filters cannot be reprogrammed on-the-fly without packet loss), complexity and number of simultaneous filters.

Force10³⁰ with their P-series network appliance, offers a platform that features high-speed packet capture at 10 Gbps with the ability to inspect and monitor packets using a supplied open toolkit. Their processing architecture, allows thousands of rules to be simultaneously applied to each packet, making the product a turnkey solutions for the security market.

The current industry trend is to go beyond the limitations of the current generation of products that accelerate packet capture, but that do not offer much (beside Force10 for security) in terms of high-speed packet processing. Technologies such as ASIC and FPGA that power many (if not all) of the solutions listed above, are designed to efficiently undertake a precise activity (e.g. packet filtering), but do not provide much in terms of packet processing and programmability (hardware skills are required). nor are they able to quickly react to new market requirements (e.g. block Skype).

Network processors (e.g. Intel IXP family) represented an evolution of FPGA in terms of programmability and extensibility, as they enabled software developers to develop generic code that was able juggle packets at high-speeds. Unfortunately the limitations in terms of ease of programmability (assembly skills were required) and inability to accelerate memory-intensive tasks relegated them to a niche market. Intel, one of the main market players, has decided to stop the development of new network processor versions as they have embraced a new trend in the industry known as multicore computing which provides both processing speed and programmability.

Multicore computing has been the industry solution to further improve processor scalability while tackling common problems such as power efficiency and the GHz clock race that affected most CPUs. Beside Intel³¹, Sun³², AMD³³ and IBM³⁴ who manufacture generic processors, companies such as Cavium Networks³⁵ and Tilera³⁶ are betting on the need for specialised multicores for creating the next-generation network appliances. The main advantages of this technology are:

- Programmability using languages such as C/C++ with full operating system support that allows existing applications to be ported with little effort to the platform.
- Scalability by using massive multicores (Intel demonstrated a 80-core CPU in 2006).
- No code-space and memory limitations.
- Native support for networking, with very low-latency packet reception and transmission, as well as deep packet inspection.

The use of multicore finally makes possible the development of both high-speed (10 Gbps and beyond) and memory/computing-intensive applications that overcome the

³⁰ <http://www.force10networks.com/>

³¹ <http://www.intel.com/>

³² <http://www.sun.com/>

³³ <http://www.amd.com/>

³⁴ <http://www.ibm.com/>

³⁵ <http://www.cavium.com/>

³⁶ <http://www.tilera.com/>

limitations of FPGA and network processor-based solutions. The market trend is clearly towards multicore solutions as they are the only ones able to offer developers a solution to network monitoring needs, as well as providing scalability for future high-speed networks. The benefit of network acceleration and multicore is also not limited to specialised processors, but is becoming available to everyone by means of technologies such as Intel I/OAT³⁷ that significantly reduces CPU overhead, thus freeing resources for network monitoring.

4.4 Lower-Layer Monitoring

Research and education networks continue to enhance their service portfolios with a variety of packet and circuit switching technologies. As well as purely packet-switch services such as Premium IP and MPLS VPNs, lightpaths running SDH/SONET or Gigabit Ethernet are increasingly being used for point-to-point (P2P) applications. These services require monitoring in order to detect anomalies, performance degradation, and to localise faults in the circuits.

perfSONAR³⁸ is being developed by the GN2 project, ESnet, Internet2 and RNP as a multi-domain monitoring to assess the quality of IP services. However, whilst these services can be accurately monitored at the IP layer, network operators tend to have limited information about the operational status of the underlying links in hybrid networks; often only whether they are *up* or *down*³⁹.

Performance monitoring (PM) information in hybrid networks may be collected from miscellaneous network elements, such as:

- IP routers supporting 1 GE, 10 GE and PoS interfaces.
- Ethernet switches supporting 1 GE and 10 GE interfaces.
- SDH/SONET switches supporting 1 GE, 10 GE and SDH/SONET client interfaces.
- Optical transmission systems supporting 1GE, 10 GE and SDH/SONET client interfaces, as well as OTN OTUk (k=1,2,3) line interfaces.

Performance monitoring metrics (the representation of information using a specific system of measurements) may be categorised into the following technology-oriented groups:

- Optical domain metrics.
- OTN (and Forward Error Correction) metrics.
- Generic Framing Procedure (GFP) metrics.
- SDH/SONET metrics.
- 1 GE and 10 GE metrics.

³⁷ <http://www.intel.com/go/ioat/>

³⁸ GN2-JRA1: Performance Measurements and Monitoring, <http://www.geant2.net/server/show/nav.754>

³⁹ M. Yampolskiy, et al., *Management of Multidomain End-to-End Links*, 10th IFIP/IEEE Int. Symposium on Integrated Network Management, Germany, May 2007.

The collection of Layer 1 and 2 performance monitoring metrics from optical transmission systems and SDH/SONET switches is a complex task as vendors often support different sets of metrics on different equipment and interfaces. For optical transmission and FEC, these are not standardised, whilst the export of monitoring data is often undertaken with proprietary mechanisms such as northbound NMS/EMS interfaces.

Optical Transmission

Optical transmission systems are able to monitor optical signal parameters in order to assess the levels of noise, attenuation, dispersion and inter-modulation effects. They use non-intrusive measurements to analyse the spectral characteristics of the signal in short-time intervals, usually once-per-second. These are then usually correlated in 15-minute and 24-hour periods, with historical data typically stored for up to two days.

The most common measurements are the input/output power on each wavelength, total receiving/outgoing power per group of wavelengths, and maximum/minimum input/output power per wavelength or per group of wavelengths. Power levels can be measured either before multiplexing/demultiplexing, or at internal amplification stages, although the purpose and usefulness of each measurement obviously depends on the specific characteristics of each optical transmission system.

For day-to-day operations, simply knowing the status of a line (either towards a provider's network, or towards a customer's network interface) is usually sufficient given that the majority of operational problems are usually caused by fibre cuts or equipment failure (e.g. laser transmitters)⁴⁰. However, being able to monitor power levels is useful for longer-term network planning.

Fibre-ageing effects, variations in optical noise due to changes in temperature and environmental conditions (e.g. humidity), and fibre maintenance procedures are some of the factors that may affect signal quality. The transmitting power level of all operational wavelengths is also affected when additional wavelengths are added into an optical span, which can result in the power for one or more wavelengths to fall below receiver sensitivity thresholds. In dynamically-provisioned optical networks, monitoring information on optical power levels will need to be taken into consideration before extra wavelengths are activated or re-routed.

The ITU-T also recommends the measurement of *Optical Signal-to-Noise Ratio* (OSNR) and *Q-factor* on a per wavelength basis⁴¹, although these do not appear to be supported by any currently-available transmission equipment. The OSNR metric can be simply and accurately measured in point-to-point systems when no significant noise shaping is present (e.g. without inline signal boosters), whilst the Q-factor is measured by analysing the pulse shape of the optical signals. Under ideal conditions (e.g. where only Gaussian noise is present etc..) the *bit error rate* (BER) can be derived from Q-factor measurements.

⁴⁰ GN2-JRA4 E2E Monitoring System of Cross-Border Fiber, <http://cnmdev.lrz-muenchen.de/e2e/>
⁴¹ *Optical monitoring for DWDM systems*, ITU-T Recommendation G.697 (6/2004).

OPUk Paths and FEC

The ITU-T recommends that the performance of an OTUk (k=1,2,3) path is expressed by the number of errors identified in blocks (e.g. a consecutive set of bits associated with a path)⁴². Errors in OTUk paths are identified using the *bit interleaved parity-8 (BIP-8)* error correction technique, and are reported as *errored blocks* (blocks with at least one errored bit), *severely errored seconds* (seconds with more than 15% errored blocks or at least one defect), and *background block errors* (errored blocks during non-errored seconds). The quality of the OTUk paths is reported using the *severely errored second ratio (SSER)* and *background error ratio (BBER)*.

Long-haul and certain regional optical transmission systems support *forward error correction (FEC)* in their line interfaces which allows receivers to identify and (partially) correct errors in an incoming optical signal. Although standard FEC algorithms are available, many vendors have developed their own proprietary techniques to improve performance. The performance of an OTUk path is calculated after the FEC technique is applied.

SDH/SONET

The performance of SDH/SONET paths/connections is assessed by the number of errors identified in blocks, as with OTUk paths⁴³. Errors are reported as *errored blocks* (blocks with at least one errored bit), *errored seconds* (seconds with at least one errored block or at least one defect), *severely errored seconds* (seconds with more than 30% errored blocks or at least one defect), and *background block errors* (errored blocks not during severe errored seconds). The quality of SDH/SONET paths/connections is reported using the *errored second ratio (SER)*, *severely errored second ratio (SSER)* and *background error ratio (BBER)*. A network operator defines the performance monitoring points inside a SDH/SONET domain, and data is accumulated in 15-minute and 24-hour intervals.

GFP

The GFP adaptation schema is supported by optical transmission systems where client signals are encapsulated into OTNk frames, or SDH/SONET switches where client signals are encapsulated in STM-x/STS-x frames. Performance monitoring metrics count the number of GFP frames with or without errors, such as the number of received good frames, number of discarded superblocks, etc..

GE and 10 GE

Ethernet switches and IP routers support a complete set of performance metrics (e.g. number of incoming and forwarded frames). Unfortunately, support for these on the Ethernet client interfaces of optical or SDH/SONET transmissions systems is usually limited as they aim to be as transparent as possible for incoming traffic. This enables them to forward incoming Ethernet frames without having to implement the full Ethernet standards, and therefore functionality is minimised. In particular, Ethernet

⁴² *Error performance parameters and objectives for multi-operator international paths within the Optical Transport Network (OTN)*, ITU-T Recommendation G.8201 (09/2003).

⁴³ *End-to-end error performance parameters and objectives for international constant bit-rate digital paths and connections*, ITU-T Recommendation. G.826 (12/2002).

performance metrics are almost non-existent for systems implementing the GFP-T adaptation mode to multiplex GE onto OTU1/OTU2 signals.

4.5 PERT

The idea of a Performance Enhancement and Response Team (PERT) is a relatively new one, and aims to be to network performance what a CERT is to network security. The PERT concept came out of Internet2⁴⁴ in response to a recognition that despite large increases in the amount of bandwidth available, many users were unable to fully exploit this due to problems that existed somewhere in the network or end-hosts. More importantly, these problems were often difficult for the user to identify and resolve by themselves, or even to the extent that degraded performance was accepted as the norm.

As it was clear that similar issues faced European research and education networks, a trial PERT was established in late-2003 (under the GÉANT project), followed by a production PERT in early-2005 (under the GN2 project). However, this is quite limited in scope, currently comprising one part-time person to coordinate the activities, nine part-time persons drawn from NRENs to work on cases, and a number of volunteers who can be called upon for their expertise in specific aspects of networking. In addition, PERTs are also known to have been established by FUNET (Finland), GARR (Italy), RESTENA (Luxembourg), SANET (Slovakia), and SWITCH (Switzerland).

Although PERT functions are often performed as a function of an already established NOC, it is important to distinguish their tasks. A NOC is responsible for investigating and fixing failures in the network, which must usually be actioned urgently. By contrast, a PERT will usually receive reports from end users whose applications are not performing as expected, and who suspect an issue with a network connection. Such issues may take longer to resolve, but conversely they are less urgent. In addition, a PERT can generally only recommend remedial action once a problem has been identified, whereas a NOC is expected to fix problems as soon as possible.

Since the establishment of the GN2 PERT, 90% of requests submitted for investigation relate to users experiencing less than expected throughput. These are generally related to high-speed long-distance TCP connections (typically Trans-Atlantic) where old or buggy TCP implementations are being used, where sub-optimal settings have been configured, or simply that there is a lower-capacity link in the path that the user is unaware of. These problems are relatively easy to resolve, but certain network conditions can also cause low rates of packet loss, which can be enough to seriously degrade TCP throughput over long-distance connections due to the way the protocol reacts to congestion. Although newer versions of 'high-speed' TCP (e.g. FastTCP or HSTCP), hardware-based TCP offload engines, and LSL (Logical Session Layer) can improve the situation, these are not without problems either.

Another frequent cause of problems is Ethernet duplex mismatching, whereby two connected systems do not correctly auto-negotiate full-duplex operation. Although this is less of an issue with newer Ethernet equipment, there is still a lot of legacy

⁴⁴ <http://www.internet2.edu/>

equipment around that requires manual configuration and which can severely degrade network performance if not configured correctly.

Other problems generally relate to so-called *middle boxes* such as firewalls, NATs, transparent proxies and rate-shapers that are usually located somewhere on the customer network. Sometimes this is as a result of these devices being misconfigured or simply underpowered, but in other cases they are deliberately enforcing certain policies unbeknownst to the user.

It is important for a PERT to adopt a standard methodology for diagnosing problems, and a good starting point is the network performance framework defined by the IETF IP Performance Metric (IPPM) Working Group in RFC 2330. This defines several metrics that can affect network performance such one-way delay (OWD), round-trip time (RTT), delay variation (jitter), packet loss, packet reordering, maximum transmission unit size (MTU), and bandwidth delay product (BDP). By being able to analyse these elements and compare actual performance with expected performance, it is usually possible to quickly determine where problems may lie.

There are a number of tools available for monitoring networks, which can be in a proactive or reactive fashion. Proactive measurements require the regular monitoring of network nodes and links in order to detect failures and abnormal situations. This may be done in an active manner by injecting diagnostic traffic into the network and analysing its behaviour using tools such as Smokeping or the RIPE TTM system, or in a passive manner by examining the characteristics of regular traffic using the likes of NetFlow or SNMP accounting. Reactive measurements are undertaken when problems are suspected or reported, but the exact nature of the problem needs to be isolated. Such measurements may again be done actively using tools such as ping, traceroute, or iperf, or passively using the likes tcpdump or Ethereal.

Although some diagnostic tools can be used without requiring any special access to the network, certain measurements require access to intermediate nodes, or for monitoring equipment to be placed in strategic locations (such as PoPs). This therefore requires the cooperation of the relevant network operators, and significant expenditure on equipment if comprehensive coverage is to be achieved.

Nevertheless, there are a number of initiatives within the research and education networking community such as perfSONAR, HADES and LOBSTER which aim to establish a network of monitoring sensors that are able to provide good coverage of the major international backbones as well as a number of NRENs.

Unfortunately, whilst active monitoring sensors can be established using relatively cheap hardware, they are limited by the amount of diagnostic traffic they can insert into the network, and how frequently they are able to do so. Too much traffic could adversely affect network performance itself, but too little means that intermittent problems can be missed. It is therefore necessary to undertake passive monitoring to get an accurate view of network performance, but this starts to become expensive in terms of interface, processing and storage costs, especially at speeds higher than 1 Gbps. For 10 Gbps network connections, expensive specialised NICs (e.g. Endace DAG) are required, and even these have limited monitoring functionality. For

example, it is estimated that it would cost in the order of EUR 2 million to install passive monitoring sensors to provide full coverage of the GÉANT2 network.

Another important consideration for a PERT is establishing efficient communication between itself and its users, and as well as good coordination between different network domains. There was no pre-existing relationship between the GN2 PERT and its users, and this meant cases generally only reached the PERT through lengthy chains of referrals, if at all. As a result, there were fewer cases than expected, even though the majority ended-up being satisfactorily resolved.

The multi-domain nature of most end-to-end performance problems also presents challenges for a PERT with limited scope. Successful resolution of cases often requires access to systems at user premises and/or information from intermediate networks, which raises trust, security and privacy issues. Although it is generally possible to liaise with the relevant parties, this is a time consuming process and relies on the goodwill of those involved. In addition, most NRENs do not have any defined PERT function, which makes it difficult to know with whom to communicate when problems arise within their domain.

If the concept of the PERT is to evolve, it needs to be extended further to a national, regional and even campus network level (the exact structure may vary from country-to-country). That way, users would be able to raise a case with their local PERT (possibly in their own language) who would then be able to open the investigation and escalate it to other PERTs as appropriate. The main issue is whether to adopt a decentralised model where PERTs all operate independently, or to adopt a federated model with a central PERT and a number of associated local PERTs.

In the decentralised model, the PERTs would collectively work to establish guidelines and requirements, but each PERT would only communicate with their users and neighbouring PERTs. Although this is potentially a scalable solution, the downside is that cases would proceed in a sequential and possible ad-hoc manner, and the speed of resolution would depend on the nature of communication between each PERT in the chain.

In the federated model, a central PERT would take responsibility for management and investigation of multi-domain cases forwarded from local PERTs. It would also be responsible for establishing guidelines, maintaining a common knowledge base, and operating a central ticket system. This could be largely build upon the existing GN2 PERT, and in principle could provide service to countries whilst they were still establishing their own PERTs. However, this model would require ongoing central funding, and the willingness of local PERTs to adopt common operating methods whilst exchanging potentially sensitive information.

Whichever model is chosen, it is important to establish a common mechanism such as a ticketing system for logging and tracking network problems. This will ensure that cases reach the appropriate team(s) for investigation, and that progress towards resolution of cases can be tracked. There is already a PERT Ticketing System (PTR) in operation, and in principle this could be further utilised by local PERTs as none have apparently developed their own systems yet. Indeed it could perhaps be made a condition of becoming a recognised PERT that a common system is used, although it

may be the case that local PERTs have their own specific requirements (e.g. local language). In these circumstances, it would likely mean that tickets would have to be logged in all applicable systems with references to each other, but this model might have to be supported anyway if PERT systems were established in other areas of the world and needed to refer cases to each other.

Another important service that PERTs should provide and contribute to, is a Knowledge Base. The same issues are often encountered over-and-over, and it can save immeasurable time if these are documented and made available to users and other PERTs. There is already a PERT Knowledge Base maintained by the GN2 PERT in the form of a Wiki which is a simple and scalable solution for adding and maintaining information. However, local PERTs may also wish create localised versions in order to better adapt the information for their own communities, so the main knowledge base should have a licensing policy that allows for this. In reality though, the most productive localisation activities are likely be the creation of value-added introductory, best practice and *how-to* documents produced from information in the knowledge base.

Complementary to the Knowledge Base, would be the development of Best Practice guides. These should largely be targeted at non-specialist users, and should provide an overview of the issues that may be faced when trying to achieve the best network performance possible. They should also provide guidelines and tips for improving performance, as well as outlining some of the more common problems.

At the present time, the GN2 PERT largely focuses on resolving specific types of problems with IP networks. In future, with more research and education networks likely to be operating hybrid networks (i.e. at optical and IP levels), it may well be the case that the scope of PERTs will need to be extended to include investigation of problems at Layers 0-2 as well. The extent to which such problems affect IP network performance needs to be further investigated, but it would represent a quite significant extension of PERT responsibilities. Whilst there is significant experience of IP networks within the academic community, there is much less with respect to optical transmission issues, and it would undoubtedly require additional resources.

Another area where the role of PERTs may be extended, is with respect to videoconferencing. Whilst high-quality videoconferencing has been available for a number of years, it remains under-used (particularly for ad-hoc events) due to its perceived unreliability with respect to set-up (calls often fail to connect) and intermittent problems with video and particularly audio quality. Many of these problems are easily attributable to end-site or intermediate equipment, but users are often not aware these can be a problem, and as result lose faith in videoconferencing itself.

Finally, consideration should be given as whether a PERT should be an integral part of an established NOC, or whether it should be a separate entity. The advantages of the first approach are it ensures the relevant PERT contacts have direct access to the network where problems may lie, and it makes it easier to establish a PERT, especially where demand is uncertain. The downside though, is that NOCs tend to be focused on resolving problems within their network domain, whereas performance issues often span multiple network domains and require a coordinated approach. In

reality though, it is likely that most NRENs would choose to appoint NOC staff as PERT contacts, at least when initially establishing their PERT.

4.6 Conclusions

It is likely that most NRENs and international backbones will continue to rely on overprovisioning to ensure reliable performance. Although IP has fairly well developed QoS mechanisms these days, not only there is limited demand for premium services currently, but neither are there many incentives (financial or otherwise) for customers to utilise different classes of service. This may become important if demand for bandwidth starts to exceed the ability of the networks to provision it (or more specifically fund it), but until then, a lot of time can be spent devising and configuring bandwidth allocation models when it is often simpler and cheaper just to provision another link. This is particularly the case where research and networks have access to dark fibre, as they have the ability to provision additional wavelengths as required.

Of course, some networks may still have bandwidth issues, especially where additional bandwidth must be purchased from a service provider. In these circumstances, they may have little choice but to run QoS or other traffic engineering mechanisms, so it is important that edge nodes are able to support this. In addition, core networks should support QoS transparency in order to allow regional, metropolitan and campus networks to deploy QoS across the core if they wish.

Where research and education networks have access to dark fibre, lightpaths can be used to provision guaranteed bandwidth for very demanding customers and users (e.g. the high-energy physics and radio astronomy communities). Lightpaths are dedicated SDH/SONET or Ethernet circuits established over separate wavelengths on WDM systems, over which customers may provision their own services. Although lightpath configuration is currently a somewhat manual procedure and therefore really only suitable for customers with relatively static requirements, advances in control plane and network virtualisation software are expected to facilitate a more dynamic approach in future.

Unfortunately, the advent of hybrid optical/IP networks complicates network management as there is limited integration between the different layers. The IP and optical layers have also evolved somewhat separately, and therefore have different management protocols, tools, and operational procedures.

There are a number of initiatives in the research and education community to develop tools for monitoring and managing optical networks, but these are at quite an early stage of development and offer limited integration with the IP layer. Even in the IP world where SNMP has been universally adopted for network management, there is a need for improved tools in order to visualise and interpret the large amounts of diagnostic information that is generated.

Network monitoring is also important for understanding and managing networks. It gives an indication of the amount and type of traffic traversing a network, thus allowing performance issues to be identified, and permitting more effective

bandwidth management. It can also be used to provide alerts in the event of failures or cyberattacks.

Unfortunately, traffic analysis has traditionally received limited attention, so few network devices provide much support for it. Whilst this situation is changing, network monitoring is hampered by a lack of standards, and is also challenging to undertake at high line rates. Whilst network monitoring at speeds of 1 Gbps and below is possible with commodity hardware, expensive custom solutions are required beyond this. Moreover, the amount and type of information that can be collected at the higher-speeds is limited by the capacity and performance limitations of current hard disks. As a result, the devices need to have on-board intelligence in order to process the collected data on-the-fly before it can be transferred to disk.

There are currently a handful of commercial solutions that are capable of monitoring 10 and 40 Gbps networks, but these are extremely expensive. The use of commodity multi-core processors instead of FPGAs or network processors offers the potential for faster and more cost-effective monitoring, although the market is likely to remain specialised and low-volume in the immediate term.

Another important issue for network management is the increasingly prevalence of so-called middle-boxes that usually sit on the exterior boundaries of campus and/or edge networks. These include firewalls, NATs, rate-shapers and intrusion detection devices, which are ostensibly to enforce security and traffic policies.

Unfortunately, operational experience shows that around 90% of all reported problems in modern networks are due to issues at end-sites, and a high-percentage of these are attributable to middle boxes. In some cases, this is simply down to misconfiguration, but often the behaviour is intentional, even though this is not always apparent to the users or even the network operators themselves. Aside from the time spent troubleshooting such problems, these measures often encourage circumvention of policies by encapsulating certain prohibited or restricted traffic within other traffic, or by writing applications to run over allowed protocols (HTTP in particular). In other words, devices that supposed to help manage and secure the network, can often end-up making things more complicated and less secure. In addition, they can often hamper performance or delay necessary network upgrades if devices are not available to support higher line rates.

Whilst it is unrealistic to expect that certain types of middleboxes, particularly firewalls, are likely to disappear any time soon, there is an increasing school of thought that their use should be minimised and ideally be moved closer to end-hosts if possible. A more radical idea, but one that is gaining increasing credence, is that most functions performed by middleboxes could actually be undertaken on the end-hosts on a case-by-case basis. With either of these approaches, the main concern is increased management complexity, but operating systems are becoming more secure anyway, and increasingly provide better support for remote configuration.

A longer-term solution is to allow middleboxes to be dynamically managed by trusted third-parties to enable protocol support as required. The IETF and others are working on a number of solutions to this effect (e.g. MIDCOM and SIMCO), although it is unclear when these implementations will be available.

Since its creation in 2003, the GN2 PERT has proved to be quite adept at tracking-down the issues that prevent users from fully utilising GÉANT and other connected networks. Although, the PERT currently has limited effort, it has been able to resolve most reported network performance problems, the vast majority of which are caused by old or misconfigured TCP implementations, Ethernet duplex mismatching, or errant middleboxes on end-user sites. In fact, few problems have been found to be attributable to the GÉANT or NREN networks.

Unfortunately, the multi-domain nature of most end-to-end performance problems presents a challenge, as successful resolution of cases often requires access to systems at user premises and/or information from intermediate networks. If the concept of the PERT is to evolve, it needs to be extended to NRENs, and possibly to the regional and/or campus level as well. This need initially only be a nominated contact in an existing NOC who would take responsibility for handling and referring cases. In addition, it would be desirable to establish standard operating procedures, a common knowledge base, and a central ticketing system.

5. Middleware

The research and education networking community is moving from offering just a simple IP service to a much richer portfolio of services, among which middleware plays an important role. The term *middleware* refers to the applications that connect the lower networking layers with directories and other multipurpose services, and for this reason is also referred to as the *glue layer*.

When considering middleware, emphasis immediately falls on the need to authenticate and authorise users to access resources. The currently deployed authentication and authorisation infrastructures have very different capabilities, ranging from simple username and password-based authentication to sophisticated systems for granting or denying access to resources.

Middleware has become increasingly important as users have become more mobile and need to access their resources from anywhere on the Internet, as well as wishing to share resources such network devices and storage (of which Grid is perhaps the best example). There is also the need to reduce the management overhead that naturally increases when users move from one place to another, and particularly for temporary users who only use institutional resources for short periods. The usage of ad-hoc middleware technologies can reduce such overhead and the inevitable errors that occur when the user data is duplicated in various places.

The use of mobile devices is becoming more practical as mobile techniques mature (e.g. IEEE 802.1X, upon which the eduroam service is based) and they are better able to access resources from different locations. Especially at campus level, wireless provisioning is becoming increasingly ubiquitous due to the fact that it can be deployed there more rapidly and with greater flexibility than wired Ethernet. However, such developments bring with them an increased need for security mechanisms to ensure that users are indeed who they claim to be, and authentication and authorisation mechanisms play an important role in this.

In many modern applications, authentication (proving that the user is who they say they are) is done by a separate organisation or, more generally, in a separate management domain from authorisation (deciding whether that user is permitted to do what they have requested). One of the big challenges that arise when sharing information among different domains is the way trust is built and delegated among these various parties. There are several technologies, such as PKI, signed SAML tokens, DNSSEC and others, that are used to build trust by transporting statements about authentication, authorisation and identity between management domains. Together with such technologies, policies play an important role in the trust-building process, for example an authorising domain may wish to know the minimum identity checks that were performed before the user was given their authentication credentials. The combination of technologies and policies is known as the trust infrastructure: without it authentication and authorisation can only take place in a single domain.

The SERENATE study only touched upon middleware issues, as middleware did not play such an important role at the time. Nevertheless the increased importance of authentication and authorisation was already foreseen during the SERENATE study, with one of the recommendations being the establishment of a pan-European

authentication and authorisation infrastructure. This recommendation was followed up by the GN2 project, which created a dedicated Joint Research Activity on Roaming and Authorisation.

This section will cover the current state-of-the-art with respect to available middleware technologies, as well as developments expected in the next couple of years. It will specifically focus on tools and standards available for (Federated) Identity Management (FIdM) in Higher Education (HE) and commercial environments, as well as for Grids. In addition, it will examine current and emerging mobile technologies in order to determine whether they will meet user requirements in the future. The aim is to provide an overview of what an Identity Management System is, why it is important to have one, how federations can be built, and how middleware technologies are expected to evolve.

5.1 Identity Management

The computerised organisation of user identities in institutions has always been a challenging task. With increasing demand for online access to network resources, institutions need to consider implementing proper systems to allow this without affecting security.

The mapping of users to electronic identities is a strategic element in any organisation, since it constitutes the basis for the access to institutional data and IT services. This process is known as *Identity Management*.

An *Identity Management System (IdMS)* is a system that combines technologies and policies to allow organisations to store the personal information of users and keep this updated. This basically requires identity-related data to be gathered from the systems used to store it (e.g. directories) and linked together so that all data associated with an individual is listed together. This can then be used to authenticate individuals in order to determine what network services and/or applications they may use.

Managing authentication and authorisation mechanisms is quite a complex task in the academic community as researchers, lecturers and students often move between institutions and departments for various reasons. These changes not only affect the individuals concerned, but also their roles within the institutions which have to be taken into account. It is therefore important to be able to keep track of, and handle the changes that continuously take place.

The pre-identity management situation was that of user data being duplicated in many different databases or directories that were administrated by different departments. The result was that users often had to contact each department separately for access, and have to remember several sets of credentials.

In recent years, the notion of identity management has changed with user identities no longer being managed on a per-application basis, but rather as a part of the network infrastructure that interacts with services. This has several advantages:

- Rather than each application requiring its own individual infrastructure to manage users, roles and control access (typically an LDAP database), there is a single,

centralised IdMS;

- Less human effort required (and therefore cost) to maintain the system;
- Less likelihood of errors when users leave or change role within the same organisation;
- Fewer credentials need to be remembered by users, as authentication can be handled in one place;
- Single Sign-On (SSO) becomes a possibility, allowing users to login once and access the full range of services they are permitted to use.

The move to IdMSs has been greatly driven in the research and education community by the eduroam service and Grid networks. However, there are several factors to take account when choosing an IdMS:

- Current and likely future requirements need to be carefully planned;
- The experience of other NRENs and academic institutions should be considered;
- Standards compatibility is important, especially when choosing a commercial product. In addition, is the vendor committed to supporting future standards?

Once an institution has developed an IdMS for its own users' identities, it is possible to consider federating this system with other IdMSs to provide shared access to services across institutions. In the European academic community, the desire to federate identity for access to networks (eduroam) and shared services has contributed to the deployment of proper Identity Management Systems at institution level.

5.2 Identity Federations

Authentication and Authorisation Infrastructures (AAI) and the mechanisms that allow for their integration through federations, are currently an essential component of IT infrastructures within academic institutions. They are also important for realising a common infrastructure that enables the free movement of students and researchers in Europe and the rest of the world.

AAIs are key elements when deploying services like corporate e-mail, e-learning systems, and a wide range of applications that can be autonomously requested by users (access to records, reservation of common spaces, access to library resources, software updates, etc...). Extending such services to the whole academic community without the support provided by an AAI has a huge cost, especially in terms of human resources.

The integral services offered by academic institutions, both in terms of learning and research, can become much more relevant if AAIs are built with federation facilities, allowing the establishment of trust links among the participant institutions. These links are directly related to the goals of the Bologna Process, especially in the cases of collaborative research projects and student/researcher mobility.

Federated identity management, also commonly referred to as *federations*, indicates the collection of all processes, standards and technology that allow a controlled exchange of users' identity data across organisational boundaries. This does not necessarily only mean between institutions, but can also mean between sections or

departments within institutions. The aim is to provide authentication, authorisation and personalisation across a distributed services landscape.

For example, a student belonging to a certain (home) institution might have to be able to access the electronic learning environment of another (remote) institution, perhaps in context of a jointly-developed course. The remote institution will want to grant the student access to its electronic learning environment on the basis of their credentials at the home institution (authentication). It may then decide what sort of access should be given to the learning environment on the basis of the type of course (authorisation). Finally, the electronic learning environment may be tailored for the student based on the course requirements and/or their own personal preferences (personalisation). In these circumstances, identity data is exchanged across organisational borders, without redundant data administration being carried out. It also avoids the use of multiple user names and passwords, or separate authentication questions.

Federations are based upon the principle that a user's authentication is undertaken by their home organisation, and that a resource trusts what the home organisation states about that user. In many cases, the same user can have a different role in different organisations, and for this reason the authorisation is normally performed by the resource that the user wishes to access.

Collaborative tools are an important resource in collective research tasks, and their access and administration becomes much simpler when a federated AAI is available. These mechanisms are fundamental to a new collaboration paradigm known as e-science *virtual organisations*. These virtual organisations allow not only the use of certain resources by their members, but they also make it easier to introduce new members and share resources to with those willing to contribute them. This way, services are centred on the user community, which can employ the middleware infrastructure provided by the federation to build a collaborative environment. Obviously, the virtual organisation paradigm can equally be applied to other fields, such as learning or student interaction.

As interoperation and openness are necessary in this kind of infrastructure, a key aspect is using products that are standard-compliant. It is worth noting that the Authentication and Authorisation framework known as eduGAIN that is being developed by the GN2 Joint Research Activity on Roaming and Authentication (JRA5), builds on top of the national federations provided by the various NRENs. Being part of eduGAIN will mean being able to access various GÉANT2 services without the need for users to supply credentials (e.g. username and password) on an ad-hoc basis.

A temporary solution to allow NRENs that have not yet established a federation to access GÉANT2 services via eduGAIN has been set-up. This uses a centralised server to handle users that do not belong to a federation. However, NRENs should realise that this approach is neither desirable nor scalable as it does not fulfil one of the main criteria for having a federation, which is that each institution is responsible for its own users.

NRENs have been providing communication services to research and education institutions for many years, and are perceived as a neutral, common ground for inter-

organisational collaboration. As identity management federations are built on the basis of trust amongst participatory institutions, the status of NRENs makes the provision of the infrastructure to build these trust links a natural evolution of their services to the user community. The experience to-date in several countries reinforces this, and it is highly recommended that NRENs provide the support for building identity federations within their constituencies.

It also important to stress that there will likely be more than one federation for supporting different needs, to which different service level agreements will apply. Nevertheless, even different federations follow different models, they should still be able to interoperate provided standards are followed. Inter-federation peering issues are currently being investigated by GN2, Internet2, and the TERENA Task Force EMC2.

Basic Elements of Identity Federations

The flow of identity data in a federation goes from the database where it is stored by one of the participant institutions (known as an *Identity Repository*), to the applications that use them at the other participant institutions. This flow is controlled by two kinds of elements:

- The *Identity Provider* (IdP) which has access to the Identity Repository for a certain institution and is able to verify authentication. IdPs can also provide additional identity information (attributes) about their users, which in turn can be used in authorisation decisions.
- The *Service Provider* (SP) which is a site that trusts the IdPs in the federation to authenticate a user on its behalf.

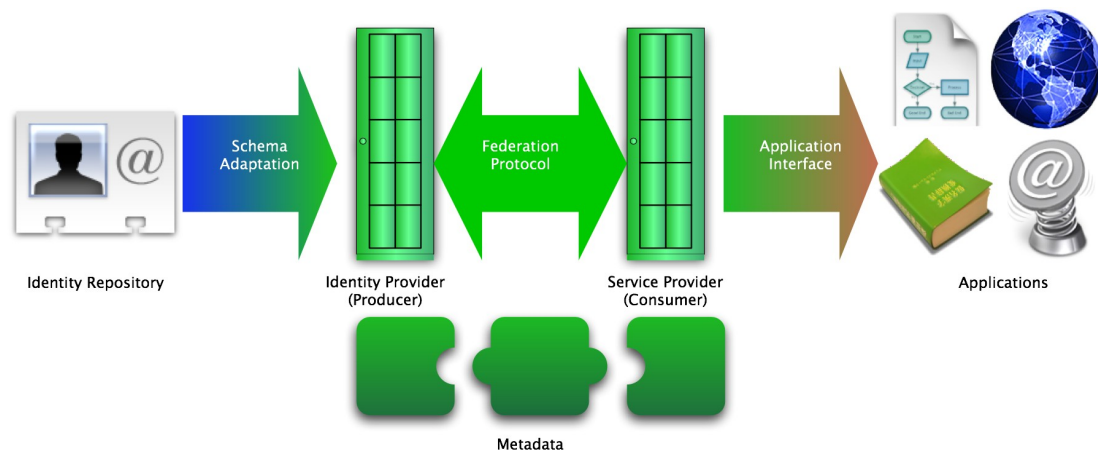


Figure 5.1: Data flow in Federated Identity Systems

Identity data is encapsulated in *Identity Assertions*, so an IdP is an assertion producer and a SP is an assertion consumer. Trust links between a given IdP and a SP are then built by means of *federation metadata* which defines the connection endpoints for each element, the services they provide and/or require, and the material required to properly identify the other party (normally, in the form of public keys or certificates).

Identity data must be exchanged using agreed syntax and semantics. A common schema of data representation defines how the values of attributes are coded and which values are acceptable. A common protocol specifies how the assertions exchange between IdP and SP is going to be performed, whilst a common metadata format is used for the trust assessment procedures. Last, but not least, standard mechanisms for supplying identity data to applications are required as well.

Schemas are application-specific to a large extent, and thus the research and education community must take responsibility for most of the relevant schema standards in their field. This is being already addressed by NRENs worldwide, and initiatives like eduPerson⁴⁵ and SCHAC⁴⁶ are already in widespread use. Evolution of these schema standards, together with their possible future extensions, is already being considered by a number of groups and organisations.

The common protocol used almost everywhere for the exchange of identity data is Security Assertion Markup Language (SAML), a standard promoted by the Organization for the Advancement of Structured Information Standards (OASIS). SAML has evolved from the original specification that is currently used by NRENs (SAML 1.1) into a new version (SAML 2.0) that has been embraced by industry (both vendors and service providers). All the companies contacted during the course of this study expressed their firm belief that SAML 2.0 would be used for the foreseeable future. In this respect, NRENs may suffer from their role of being pioneers in the identity management arena, though clear migration paths are being established, and it is highly recommended to adopt them as soon as possible.

Federation metadata is represented by means of the SAML 2.0 metadata format. Again, there is practically complete consensus in its use, and this includes even those infrastructures currently employing SAML 1.1.

Finally, it is worth noting that there is currently no well-established standard for communicating identity data to applications. This requires a specific adaptation of each application to each SP, seriously limiting pervasiveness of using federated identity management.

Identity Federation Trust Framework

Identity federations provide trust between individual elements of an identity management infrastructure. In most cases the trust fabrics are built using public key technology as it ensures direct trust between peers without the need to pre-configure every possible communication path. On the other hand, systems based on strictly hierarchical architectures, sometimes rely on transitive trust using shared keys pre-arranged between parent and child nodes.

Public Key Technology

Whilst public key cryptography is broadly accepted as useful for building trust fabrics, different communities adopt different technology standards for their infrastructures. The classic X.509 Public Key Infrastructure (PKI) based on ITU-T

⁴⁵ <http://www.educause.edu/eduperson/>

⁴⁶ <http://www.terena.nl/activities/tf-emc2/schac.html>

and PKIX⁴⁷ standards, is used by most Grid middleware, whose Global Grid PKI is maintained by the International Grid Trust Federation (IGTF)⁴⁸. Another X.509 PKI is being deployed for the European GN2 project, whose Certificate Authority (CA) is considering IGTF accreditation even though the project is not directly oriented towards Grids.

Classic PKI is also vitally important for TLS-secured communications. TLS server certificates are the main reason for NRENs to operate their own CAs, although *pop-up free* certificates (i.e. certificates issued by CAs that are implicitly trusted by web browsers) such as those provided by the TERENA Server Certificate Service (SCS)⁴⁹ have tended to replace NREN-provided services more recently.

The utilisation of the Domain Name System (DNS) for PKI has also been considered, using Security Extensions (DNSSEC) developed by the IETF. However, although the DNSSEC standards have been under discussion for more than eight years, the lack of global deployment prevents its broad usage, and forces developers to turn locally-operated PKI.

The complexity of PKIX standards, their imperfect implementation in various software products, and the absence of a truly global PKI, has often led to federation- or project-specific PKIs being implemented with different formats and protocols. In general, systems using XML-based protocols (e.g. SAML) tend to use XML Key Management Specification (XKMS) assertions⁵⁰, rather than using the classic X.509 certificate format. For example, the Shibboleth software uses federation SAML metadata to perform all PKI functions such as binding public keys to entities, key distribution, and key revocation⁵¹.

It should be noted that the choice of using SAML metadata or PKI depends also on the way a federation is implemented. For example, Shibboleth software uses federation SAML metadata to perform all PKI functions such as binding public keys to entities, key distribution, and key revocation. SAML metadata and X.509 PKI should be regarded as equivalent technologies for building trust in a federation.

Transitive Trust

A good example of a transitive trust fabric might be eduroam⁵². The eduroam infrastructure is implemented as a hierarchy of RADIUS servers, with each node only establishing direct trust with its peers – in other words its parent node, and any child nodes – through pre-shared keys (known as shared secrets). The overall trust between two communication end-points (typically the RADIUS servers of the visited institution and the user's IdP) is then derived from the individual trusts established between each segment of the hierarchy.

Transitive trust is not limited to shared key technology though. eduGAIN, the GN2 inter-federation infrastructure uses its own PKI to glue individual member federations

⁴⁷ <http://www.ietf.org/html.charters/pkix-charter.html>

⁴⁸ <http://www.gridpma.org/>

⁴⁹ <http://www.terena.org/activities/scs/>

⁵⁰ <http://www.w3.org/TR/xkms2/>

⁵¹ <http://shibboleth.internet2.edu/>

⁵² <http://www.eduroam.org/>

together. The trust between communication peers (i.e. a resource provider within one federation, and an IdP within another federation) then relies on the PKIs of those two federations, as well as the eduGAIN PKI.

For now, NRENs should support multiple trust infrastructures that support different AAs. While SAML-based infrastructures will often use XKMS statements within federation metadata, classic PKI is required by Grids and for authenticating web servers. Wherever possible, NRENs should strive to minimise the number of trust infrastructures they have to maintain, for example. by reusing existing PKIs.

5.3 User-centric Identity Management

The user-centric model of identity management has been the subject of increasing attention in recent months, as it has been identified with technologies linked to the so-called Web 2.0 (wikis, blogs, and collective services such as Flickr). This generally refers to web-based communities and services that facilitate collaboration and sharing between users. In fact, the user-centric model has been referred to by some of its proponents as *Identity 2.0*.

The user-centric model relies on the principle of giving users almost complete control on how their identity data is delivered to service providers, allowing them to use simple methods such as virtual electronic cards or referring URLs. However, whilst this model has the potential to establish social networks of users in real-time (hence its success in the Web 2.0 environment), it is still unclear whether it will be able to support the trust links required by federations that provide access to sensitive sites or data.

Two main solutions have been proposed with respect to user-centric identity management: CardSpace from Microsoft, and OpenID managed by the Open ID Foundation. Both organisations have announced a strategic collaboration to make both solutions compatible in future.

With CardSpace, users directly manage a set of security tokens in their computer using the visual paradigm of a deck of cards. These tokens are then delivered to the different services they are willing to access. CardSpace itself though, is not bound to a specific token format and it claims to be able to manage different types (SAML assertions, X.509 identity certificates, X.509 attribute certificates, etc..) using the same user interface. In principle, CardSpace could also be easily integrated into federations as an additional authentication method, though its suitability for those applications requiring a rich attribute exchange and/or high level of assurance are still open issues.

With OpenID, users are able to assert control over a certain URL. The current version of the protocol can only be used for authentication (the user demonstrating effective control over the URL), but does not allow any attribute passing, which will be addressed in the next version. The identity federation community is also seeking interoperability with OpenID as an alternative assertion format that can be made from a trusted federation identity provider. This would make it possible to use a federation-

trusted institutional identity for both federation applications and wider OpenID-enabled services.

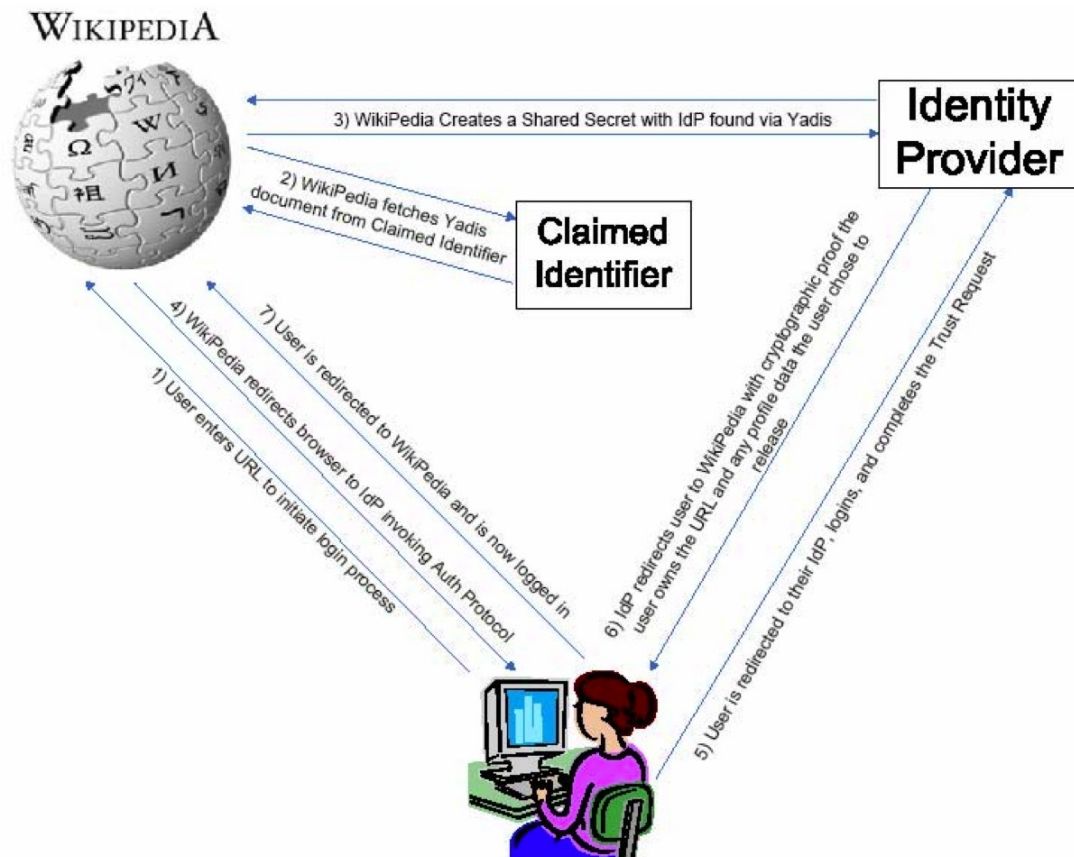


Figure 5.2: OpenId Protocol (courtesy of OpenId)

NRENs should continue to monitor the evolution of the user-centric world (in particular OpenId), and explore the possibilities of interoperability with their AAIs. As more vendors and user communities are interested in this model, there might be market demand to deploy services based on this paradigm.

5.4 Abstract Identity Framework

The lack of well-defined interfaces for applications wishing to utilise identity management services is hampering their wide adoption. Application developers must provide adaptation layers for any product they wish to connect to, and although this is not currently an overly complicated task, maintaining code for all possible scenarios could become cumbersome.

Abstract identity frameworks are therefore intended to provide a common mechanism for applications to interact with identity services. In other words, application developers only need to adapt their applications to the framework, which should provide improved compatibility and obvious savings in terms of effort.

The most promising effort in this area is known as the Higgins, which is supported by IBM and Novell. Termed by its proponents as an *API for identity*, Higgins is a

software development framework that aims to enable the integration of identity, profile and relationship information across multiple systems. Using so-called *service adapters*, new and existing systems can be plugged into the Higgins framework, so any applications written to the standardised API can share information across different systems.

Higgins is particularly oriented towards applications that can be accessed through web browsers and other clients, and is defined in terms of the Service Oriented Architecture (SOA) model⁵³ based on web service frameworks. An implementation that supports Java is in development as an initial reference.

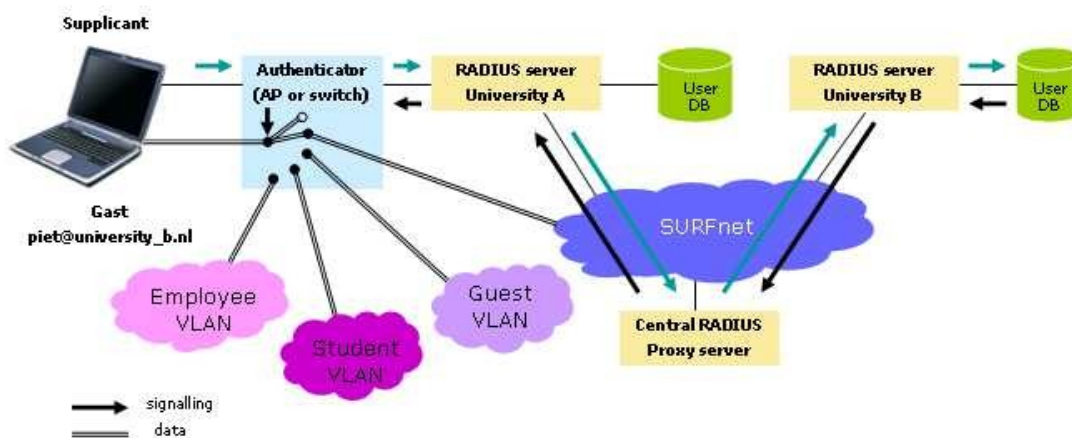
Though Higgins is still at an early stage, it is potentially of interest to the research and education community. It is therefore highly advisable for NRENs to follow the developments, and attempt to influence these where appropriate (perhaps as early adopters).

5.5 Federated Network Access

The term federated network access describes the process of granting network access to (guest) users authenticated by their home institutions. In the area of federated network access, three key developments can be distinguished: the further maturing of the eduroam service, the emerging of technologies for denying access to the network for hosts that are not compliant with some set of requirements (patch levels, up-to-date virus scanners etc.), and end-to-end diagnostics for middleware interactions.

eduroam

The roaming needs of users have led to a number of national and international initiatives to provide network access for roaming users. One of the best-known initiatives is eduroam, established in 2003 by the TERENA Task Force on Mobility as a pilot to provide seamless network access to users visiting eduroam-enabled institutions.



⁵³ http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=soa-rm

Figure 5.2: eduroam structure and operation

The eduroam infrastructure is based upon the IEEE 802.1X secure technology and on a hierarchical system of RADIUS-servers. This utilises redundant European top-level RADIUS servers, to which all participatory European NRENs connect with their national RADIUS server.

The interest in eduroam grew very rapidly, with a considerable number of NRENs joining at the very beginning and consequently a big number of institutions connected to eduroam. This success led to a need for a proper service agreement among the parties involved, hence an eduroam policy.

The GN2 Joint Research Activity on Roaming and Authentication (JRA5) therefore aims to enhance eduroam in order to provide a full service for Europe. The tasks performed by JRA5 in this respect, are the definition and acceptance of the eduroam policy by all participatory NRENs, and improvement of the eduroam infrastructure. In the latter case, to provide a more secure alternative to RADIUS, as well as to provide more dynamic methods of transferring user credentials securely within the infrastructure.

At the national level, each NREN is still requested to implement their national eduroam policy, which is influenced by national regulations and by the rules under which the NREN operates. However, all eduroam national policies mandate the usage of 802.1X and RADIUS (or equivalent) to facilitate authentication from the institution where the user logs in, to the user's home institution.

The eduroam service agreement guarantees the standards of service that member NRENs will receive from each other, meaning that user information will be handled securely and problems that might occur will be investigated promptly. However it does not guarantee the service that will be received by an end-user when they obtain network connectivity at a visited site. A completely different approach to liability is forcing all user traffic to flow through the home institution of that user by only allowing VPN access after eduroam authentication. This approach is taken by a number of institutions in Japan and Australia, with the advantage is that visited institutions are in no way involved in any incidents pertaining to guest users. The downside is that simplicity and routing efficiency is sacrificed, and the implications of this still need to be investigated.

The eduroam agreement also does not specify what connectivity should be provided for visitors, merely that the connection to the Internet should be as open as possible. There is no requirement that visited sites document what connectivity they actually provide. This means that a visitor has no way to find out in advance whether a particular technology (e.g. VPN) will work from a particular eduroam location, nor is there any guarantee that an application that worked from one visited location will also work from another.

Some eduroam participants have addressed this by specifying a minimum connectivity requirement that must be provided by all visited sites. For example the JANET Policy⁵⁴ requires all sites to provide at least sufficient connectivity to run a set

⁵⁴ <http://www.ja.net/roaming/documents/techspec.pdf>

of common applications; namely web browsing, file transfer, e-mail, VPNs and terminal server access. AARNet⁵⁵ have a similar approach, recommending a minimum list of ports and requiring visited sites to publish what connectivity they require. Visited sites must permit these protocols to pass between visiting computers and the Internet (optionally via a proxy, though this fact must be documented to help in providing support), although there is no requirement to provide access to the local area network.

As the popularity of eduroam increases and hence users expecting it to "just work", it is important that the supporting infrastructure is stable and robust and aligned with other federated applications. The deployment of eduroam at national level is a task that many NRENs have already successfully taken up, not only by running the national top-level RADIUS server, but by also providing technical support for the universities. In some cases national funding might be available to help facilitate this effort.

To address the problems related to the different implementations of eduroam from a user perspective, NRENs might consider mandating the same implementation of eduroam throughout the country. Information related to the specific implementation of eduroam in each country could then be maintained on the eduroam website.

Another approach could be to use a *level* or *rating* system where a university indicates which level of eduroam is available, and advertises this on national eduroam websites. However, it should be noted this reduces the universal user experience.

Unfortunately, there is currently little overlap between the SAML-based identity federations for application access and eduroam, which means that two separate infrastructures need to be maintained. The maturing of eduGAIN⁵⁶ though, should improve the possibility for integration between the two.

Network Endpoint Assessment

Network providers have a variety of means for limiting access to their networks, be it an 802.1X/eduroam user identity, or legacy technology such as source IP address or physical point of access. However, the systems that connect to the network can still contain malware like viruses, trojans etc.. This is where technologies like Network Endpoint Assessment (NEA) ⁵⁷ come into play.

Network operators need a way to assess the state of systems connecting to the network and, when deemed necessary, require remediation. NEA technology therefore typically consists of client software that interacts with a server operated by the network operator.

Unfortunately NEA, and similar architectures like Trusted Computing Group's Trusted Network Connect (TNC), Microsoft's Network Access Protection (NAP) and Cisco's Network Admission Control (CNAC) do not take into account the roaming

⁵⁵ http://www.eduroam.edu.au/policy/Australian_eduroam_policy_version_2_18-Mar-2007.pdf

⁵⁶ http://www.geant2.net/upload/pdf/GN2-07-024-DJ5-2-2-2-GEANT2_AAI_Architecture_And_Design.pdf

⁵⁷ <http://www.ietf.org/html.charters/nea-charter.html>

use case. A roaming user's system will typically not easily communicate with the controlling server at the visited network, due to the fact that no trust exists between the visiting user's and the network operator's systems. Currently the existing architectures are also not interoperable, although the major vendors have announced their intentions to undertake this.

As these technologies proliferate, it is important for NRENs to assess the implications on the roaming infrastructure and continue to work with the vendors to ensure compatibility with the eduroam service. NRENs should also closely follow the IETF developments in this field.

5.6 Middleware for Applications

Federated identity management is becoming more-and-more relevant for not only web sites and web-enabled applications, but also for e-mail, instant messaging, real-time collaboration and VoIP. Aside from the obvious appeal of providing a richer, seamless user experience, the use of identities can be applied to tasks such as:

- Simplifying the security administration of certain services, providing scalable ways to fight against phenomena such as spam as its equivalents *spim* in instant messaging, and *spit* in VoIP.
- Enabling a much simpler sharing procedure for specific resources so that virtual user organisations are able to provide a seamless resource-sharing service.
- Controlling access to novel services related to network-scale infrastructures, such as massive distributed storage and peer-to-peer systems.
- Allowing for a seamless integration of network and application layers, making possible the deployment of intelligent applications that take advantage of the advanced network services.
- Deploying *open services* that can be freely combined to build more complex services in a dynamic way.

There is a common consensus that web services (either by means of SOA or REST architectures) and web service security frameworks should be used as building blocks for the above tasks.

5.7 Grid Middleware

For a number of reasons, middleware developments in the Grid and NREN communities have historically followed different paths. One reason is that the Grid has been led by the High Energy Physics community (which is a small subset of the general NREN user community) which gathers expert users with very specific needs, namely secure access to supercomputers in various locations. The needs of this community were clearly different to those of the general NREN user community which is primarily interested in e-mail access, digital libraries, and e-learning etc..

Grid AAI has been implemented following a token model, based on X.509 user certificates where users present their digital certificates to the resource they wish to get access to. NREN AAIs have been implemented following a federated identity

model, where authentication is performed by the users' home organisations and a richer set of attributes can be exchanged to gain access to resources.

More recently, there has been more crossover between the communities, with NRENs looking with more interest at what is happening in the Grid community and vice-versa. Two factors can be observed:

- NREN and Grid communities are keen to find ways to leverage each other's infrastructures, and there is lot of investigation into how identity federations can support Virtual Organisations (VO). Most efforts are focusing on how to integrate Shibboleth with some of the Grid software, such as gLite and Globus.
- Many NRENs have been approached by their national Grid centres to discuss how NRENs could operationally support the Grid. For example, by running the Grid PKI, or providing monitoring tools.

At the end of 2005, TERENA helped facilitate this collaboration by periodically organising *NRENs and Grid Workshops*, in order to provide a forum to exchange experience and facilitate collaboration. This activity has proved to be quite successful, and one of the recent suggestions was to study how attributes are handled in Grid middleware and NREN identity federations. The latest developments show that interoperability between Grid Middleware and identity federations should soon be possible, at least at authentication level.

Given the growth of the Grid user community and the NRENs' expertise in operating infrastructures, it is advisable that NRENs operate and support the identity infrastructure for their national Grid centres.

5.8 Middleware Diagnostics

The increasing complexity of middleware interactions across multiple federations and administrative domains, involving many different entities like RADIUS servers, access points, switches, LDAP directories, and authentication servers means it has become increasingly difficult to diagnose problems. Therefore, the requirement for good diagnostics has become more pressing, and some national initiatives have already been established.

The TERENA task force EMC² is also discussing the requirement for a standard for log files, mainly for authentication and authorisation purposes. Log files generated by applications mostly have different formats, but having a standard would make it easier to exchange information and diagnose problems, especially in a federative context.

For example, SURFnet, has created a system known as usertracking⁵⁸ that can provide a single view of various identity management elements, and an ad-hoc diagnostics tool called Network Detective⁵⁹. In addition, Internet2 has developed an ambitious framework in the End-to-End Diagnostics Discovery (EDDY) project⁶⁰ that supports

⁵⁸ <http://usertracking.surfnet.nl/online.php>

⁵⁹ <http://detective.surfnet.nl>

⁶⁰ <http://www.internet2.edu/pubs/2006AR/eddy.cfm>

the integrated analysis and diagnosis of distributed, layered, but interdependent systems.

In addition, the lack of a consistent inter-federation monitoring system is being recognised within eduroam. National federations usually operate their own diagnostic systems providing operational information at different levels and in different formats. Combining those systems into a tool that provides a health map of the service at the international level, is a task for a new Service Activity within GN2.

Monitoring and diagnostics of any federated AAI usually demands close cooperation between service providers, users, and their IdPs. This makes the task difficult and requires not only technical but also organisational solutions.

5.9 Conclusions

Identity federations (or federations) are considered the solution to handle and support user access to remote services. The majority of the NRENs in Europe either already have a federation or are in the process of establishing one. NRENs that do not yet have a federation, should plan to have one in place within the next couple of years.

Identity federations are based on the concept of allowing users to access resources anywhere on the basis of authentication performed by their home institution. This model has some major implications:

An institution always performs the authentication of its own users, no matter which service the users wish to access within the federation. This is made possible via agreements between the various entities providing resources and handling users within the federation. These agreements are ruled by policies, which are very much federation-specific.

- 1) The authorisation of users is handled mainly at service provider level, which is aware of the role of the user in the context of the resource that the user wishes to use.
- 2) Only a user's home institution handles that user's information. This guarantees the user's privacy (as their data is only known to their home institution) and makes user management easier and less vulnerable to errors (as their data is managed in one place).
- 3) The fact that access to a resource is granted based on authentication performed by a different domain implies that a trust relationship needs to be created between the resource owner and the domain that authenticates the users. This trust is built using technologies such as X.509 certificates or XML Key Management Specification (XKMS) assertions; the technology chosen very much depending on the way the federation is constructed. While SAML-based infrastructures will often use XKMS statements within federation metadata, classic PKI is required by Grids and for authenticating web servers.

It is recommended that NRENs support multiple trust infrastructures in order to be able to handle different AAI. Of course, NRENs should wherever possible minimise the number of trust infrastructures being maintained, for example by reusing existing PKIs.

As identity federation implementations mature, it enables new ways of utilisation to be explored. One of these is focusing on how identity federations can support Virtual Organisations (VO) and thus Grids, and there are several projects that are implementing ways of integrating Shibboleth with some of most commonly used Grid software (e.g. gLite and Globus). NRENs that operate federations should examine these developments with great interest, whilst those without should have further motivation to federate.

NRENs are the natural candidates for providing technical and organisational coordination within their constituencies as well as representing national federations in the international environment.

In the future, there will be more federations with different SLAs that will serve different communities, and it expected that different federations will peer with each other. The model to facilitate inter-federation peering is currently being investigated. NRENs as representatives and coordinators of national federations should take a leading role in defining and deploying federation and inter-federation policies.

In the near future, widespread adoption of SAML 2.0 for exchanging identity assertions can be expected at least for web-based applications. Non-HTTP services will most probably still use X.509 personal certificates or some kind of secure tunnel to pass user credentials to their IdP for authentication, although other approaches to integrate these services within federated AAIs are being tested. Authorisation decisions will be supported by attributes defined by internationally-adopted schemas such as eduPerson and SCHAC.

As previously stated though, there is no well-established standard for communicating identity data to applications. NRENs might therefore be proactive in this area and some preliminary investigations via a task force might be conducted.

The user-centric identity model may yet take off for services outside of national academic federations, but in this event, users might still benefit from the potential ability to re-use federated identities with those services. However, with the expansion of federated services, the requirements for monitoring and diagnostic tools will more demanding.

Abbreviations

10GEPON	10 Gigabit Ethernet Passive Optical Network
AAI	Authentication and Authorisation Infrastructure
AARNet	Australian Academic and Research Network
AC	Alternating Current
AMT	Automatic Multicast Tunnelling
API	Application Programming Interface
APN	Articulated Private Network
ASIC	Application-Specific Integrated Circuit
ASM	Any-Source Multicast
ASON	Automatic Switched Optical Network
ATM	Asynchronous Transfer Mode, a transmission protocol
BDP	Bandwidth Delay Product
BER	Bit Error Ratio
BGP	Boundary Gateway Protocol, used for IP routing
BIDIR-PIM	Bi-directional PIM
BLSR	Bi-directional Line Switched Ring
BPON	Broadband Passive Optical Network
CA	Certificate Authority
CESNET	Czech Research and Education Network
CFM	Connection Fault Management
CIDR	Classless Interdomain Routing
CLI	Command Line Interface
CPU	Central Processing Unit
CRC	Communications Research Centre, a Canadian government agency
CS-RZ	Carrier-Suppressed Return-to-Zero
CSC	Carrier Supporting Carrier
CWDM	Coarse Wave-Division Multiplexing
DANTE network	Delivering Advanced Networking to Europe, runs the GÉANT2 network
dB	Decibel
DC	Direct Current
DFZ	Default-Free Zone
DHCP	Dynamic Host Control Protocol
DMZ	De-militarised Zone
DNS	Domain Name System
DNSSEC	DNS Security Extensions
DoS	Denial-of-Service
DPI	Deep Packet Inspection
DP-QPSK	Dual-Polarisation Quadrature Phase Shift Keying
DPSK	Differential Phase Shift Keying
DQPSK	Differential Quadrature Phase Shift Keying (DQPSK)
DRAM	Dynamic Random Access Memory
DXC	Digital Cross-Connect
DWDM	Dense Wave-Division Multiplexing
ECMP	Equal Cost Multipath Protocol
EDFA	Erbium-Doped Fibre Amplifier
ENNI	External Network-to-Network Interface
EPON	Ethernet Passive Optical Network

ERO	Explicit Route Object
Ethernet	Packet-based transmission protocol
FCAPS	Fault, Configuration, Accounting, Performance and Security
FEC	Forward Error Correction
FIB	Forwarding Information Base, used by IP routers
FPGA	Field-Programmable Gate Array
FUNET	Finnish University and Research Network
FWM	Four Wave Mixing
GARR	Italian Academic and Research Network
Gbps	Gigabits per second
GE	Gigabit Ethernet
GEANT2	Pan-European Research and Education Network
GENI	Global Environment for Network Innovations, an NSF-supported network research initiative
GFP	Generic Framing Protocol
GMPLS	Generalised MPLS
GN2	European research and education networking project
GPON	Gigabit Passive Optical Network
HDLC	High-Level Data Link Control
HE	Higher Education
HTTP	Hypertext Transfer Protocol
i2CAT	Not-for-profit foundation promoting innovation and research for the next-generation Internet.
IAB	Internet Architecture Board
IANA	Internet Assigned Numbers Authority
ICMP	Internet Control Message Protocol
IdMS	Identity Management System
IdP	Identity Provider
IEEE	Institute of Electrical and Electronic Engineers
IETF	Internet Engineering Task Force
IGMP	Internet Group Management Protocol
I-NNI	Intra Network-to-Network Interface
Internet2	US research and education network
IP	Internet Protocol
IPFIX	Internet Flow Control Export Protocol
IPTV	IP Television
IPng	IP next generation
IPPM	IP Performance Metrics
IPv4	Internet Protocol Version 4
IPv6	Internet Protocol Version 6
ISP	Internet Service Provider
IT	Information Technology
ITU-T	International Telecommunications Union – Telecommunication Standardisation Sector
I-WS	Interface Web Service
JRA	Joint Research Activity
kbps	Kilobits per second
km	Kilometres
kW	Kilowatt
L2TP	Layer 2 Tunnelling Protocol

LAN	Local Area Network
LAPS	Link Access Protocol over SDH/SONET
LBE	Less-than-Best-Effort
LCAS	Link Capacity Adjustment Scheme
LDAP	Lightweight Directory Access Protocol
LED	Light-Emitting Diode
LMP	Link Management Protocol
LP-WS	Lightpath Web Service
LR-WS	Logical Router Web Service
LSL	Logical Session Layer
LSP	Label-Switched Path
LSR	Label-Switched Router
MAC	Media Access Control
MAN	Metropolitan Area Network
Mbps	Megabits per second
MIB	Management Information Base
MLD	Multicast Listener Discovery
MPLS	Multi-Protocol Label Switching
MRTG	Multi-Router Traffic Grapher
MSDP	Multicast Source Discovery Protocol
MTU	Maximum Transmission Unit
NAT	Network Address Translation
NLR	National LambdaRail, a US research network
NMIS	Network Management System
NOC	Network Operations Centre
NORDUnet	Nordic Research and Education Network
NREN	National Research and Education Network
NRZ	Non-Return-to-Zero
OADM	Optical Add-Drop Multiplexer
OAM&P	Operations, Administration, Maintenance and Provisioning
OC-x	Optical Carrier
ODU	Optical Data Unit
OEO	Optical-Electrical-Optical
OSPF	Open Shortest Path First, an IP routing protocol
OTN	Optical Transport Network
OTU	Optical Transport Unit
OWD	One-Way Delay
OXC	Optical Cross-Connect
PBB	Provider Backbone Bridges
PBBTE	Provider Backbone Bridges Traffic Engineering
PCEP	Path Computation Element Communication Protocol
PDFA	Praseodymium-Doped Fibre Amplifier
PDH	Plesiochronous Digital Hierarchy
PERT	Performance Evaluation Response Team
PHP	Penultimate Hop Pop
PIM	Protocol Independent Multicast
PIM-SM	PIM Sparse Mode
PKI	Public Key Infrastructure
PMD	Polarisation Mode Dispersion
PN	Physical Network

PON	Passive Optical Network
PoP	Point-of-Presence
POS	Packet-over-SONET
PPP	Point-to-Protocol
PXC	Photonic Cross-Connect
QoS	Quality-of-Service
RADIUS	Remote Authentication Dial-In User Service
RAID	Redundant Array of Independent Drives
RDF	Resource Description Framework
RedIRIS	Spanish Research and Education Network
RESTENA	Luxembourg Research and Education Network
RFC	Request for Comments
RIB	Routing Information Base,
RIPE	Réseaux IP Européens, an RIR
RIR	Regional Internet Registry, allocates IP addresses and AS numbers
ROADM	Reconfigurable Optical Add-Drop Multiplexer
RP	Rendezvous Point
RSVP	Resource Reservation Protocol
RTT	Round-Trip Time
SAML	Security Assertion Mark-up Language
SANET	Slovakian Academic Network
SCHAC	Schema Harmonisation Committee
SCS	Server Certificate Service
SDH	Synchronous Digital Hierarchy
SIP	Session Initiation Protocol
SLAAC	Stateless Address Autoconfiguration
SNMP	Simple Network Management Protocol
SOA	Service-Oriented Architecture
SONET	Synchronous Optical Hierarchy
SRAM	Static Random Access Memory
SSH	Secure Shell
SSM	Source-Specific Multicast
SSO	Single Sign-On
STM-x	Synchronous Transport Module
SURFnet	Dutch Research and Education Network
SWITCH	Swiss Research and Education Network
Tbps	Terabits per second
TCP	Transmission Control Protocol
TDFA	Thulium-Doped Fibre Amplifier
TDM	Time-Division Multiplexing
TE	Traffic Engineering
TL1	Transaction Language 1
TLS	Transport Layer Security
T-MPLS	Transport MPLS
UCLP	User Controlled Lightpaths
UNI	User Network Interface
uPnP`	Universal Plug-and-Play
Uninett	Norwegian Research and Education Network
UPSR	Unidirectional Path Switched Ring
VCAT	Virtual Concatenation

VCSEL	Vertical-Cavity Surface Emitting Laser
VLAN	Virtual LAN
VLL	Virtual Leased Line
VO	Virtual Organisation
VoIP	Voice-over-IP
VPLS	Virtual Private LAN Service
VPN	Virtual Private Network
VRF	Virtual Routing and Forwarding
WAN	Wide-Area Network
WDM	Wave-Division Multiplexing
WSS	Wavelength Selectable Switch
XKMS	XML Key Management Specification
XML	Extensible Mark-up Language
XPM	Cross-Phase Modulation