RINGrid contract no. 031891

Evaluation and Requirements for Infrastructures

Deliverable 3.1

State of the Art in Networks and Grid Infrastructures

Document Filename:	RINGRID-WP3-D3_1-JKU-State of the Art in Networks.odt
Work package:	WP3
Partner(s):	JKU, PSNC, GRNET, TUI, CLARA, CNIT, CNR/ISTI, UNIS
Lead Partner:	JKU
Document classification:	Public

Abstract

In order to connect and access instruments located at various places on earth one needs a special grid infrastructure connected via high-speed, low-latency networks. The requirements of the network are dictated by the instruments themselves which are deployed within the grid.

There are several possibilities how to achieve the desired properties of a network: Using special hardware, using the right protocols in the network protocol stack, and selecting appropriate middleware.All of these options will be discussed in this deliverable.

There also exist several projects and testbeds related to the idea of remote instrumentation. This deliverable will have a look at the most important ones, in order to gain information which would otherwise not be possible to retrieve during RINGrid's short project lifetime.

Delivery Slip

	Name	Partner	Date	Signature
From	Thomas Prokosch	JKU		
Verified by				
Approved by				

Document Log

Version	Date	Summary of changes	Authors
1	19-Dec-2006	First draft version, document structure established	Thomas Prokosch
2	24-Jan-2007	Chapters 1-6 finished	Thomas Prokosch, Dieter Kranzlmüller, Constantinos Kotsokalis, Tasos Zafeiropoulos, Michael Stanton, Davide Adami, Davide Dardari, Luca Caviglione, Franco Davoli, Antonio-Blasco Bonito, Alberto Gotta, Damian Kaliszan, Tomasz Rajtar, Romeo Ciobanu, Cristina Schreiner, Lei Liang
3	11-Feb-2007	All chapters finished	Thomas Prokosch, Dieter Kranzlmüller, Constantinos Kotsokalis, Tasos Zafeiropoulos, Afrodite Sevasti, Michael Stanton, Marcela Larenas Clerc, Davide Adami, Davide Dardari, Luca Caviglione, Franco Davoli, Antonio-Blasco Bonito, Alberto Gotta, Damian Kaliszan, Tomasz Rajtar, Romeo Ciobanu, Cristina Schreiner, Zhili Sun, Lei Liang
4	19-Feb-2007	Finished proofreading	Thomas Prokosch, Franco Davoli, Lei Liang, Antonio-Blasco Bonito, Constantinos Kotsokalis, Tasos Zafeiropoulos, Romeo Ciobanu, Damian Kaliszan
5	13-Mar-2007	Considered changes suggested by two internal reviews	Thomas Prokosch, Michael Stanton, Tasos Zafeiropoulos

Table of Contents

List of Figures

Index of Tables

1. Introduction

1.1. Goals

RINGrid is a project which tries to solve the problem of "Remote Instrumentation in Next-Generation Networks". It is often the case that expensive instruments cannot be bought because of the lack of funding, or when they are bought the user base would be too small. Highly specialized equipment means that only few people are in need for it, which also implies that only few people are knowledgeable enough to operate these devices.

A solution to locally not available equipment and expertise would be that scientific institutions, industry and commercial interest groups foster international cooperation and tackle the problem together. The idea would be to buy and operate expensive equipment together and to share the results. As a result it would be easier to raise the money and utilize the resources better.

Another aspect of the problem is that such equipment often has to be built at a specific location, which is true for example for radio telescopes. Telescope sites are often located in Third World countries, which have no knowledge about such instruments. If costs and knowledge were distributed equally among all partners, that is also with the countries on whose soil the device is operated, that would bring expertise to those underdeveloped areas. As a result, new communities would be formed and opportunities created. The benefit for developed countries lies in the fact that scientists will no longer need to travel long distances for conducting their experiments, which clearly brings some cost reduction and generally better availability of the equipment.

With these goals in mind, RINGrid tries to integrate devices into existing grid environments. RINGrid is one of the first projects of its kind and therefore many things are unknown. This is why the project is carried out as a "Specific Support Action" (SSA). That means that it tries to gain as much experience as possible in the domain of remote instrumentation.

Workpackage three deals with the underlying infrastructures for remote instrumentation and is therefore important for workpackage six, which is responsible for the implementation of a prototype. The underlying infrastructure consists of two parts: The network infrastructure, comprising the hardware and the network protocol stack, as well as the grid middleware which is needed for operating the devices and managing the data of the experiments.

Three deliverables are to be written for WP3:

- This very deliverable deals with "State of the Art in (Research) Networks and Grid Infrastructures". It highlights current and future network technologies and grid infrastructures which are built on top of them. While the deliverable tries to introduce all grid infrastructures relevant to remote instrumentation, the main emphasis of the deliverable lies on network technologies.
- The second deliverable resumes the topic of middleware but goes further. It takes user feedback into account and gives a "Status of Grid Middleware and Corresponding Emerging Standards for Potential Usage in Sharing Scientific Instruments via (International) Networks".
- The third deliverable evaluates the technologies presented in the first two deliverables by giving a "Summary of Requirements and Needs to be Currently Fulfilled to Efficiently Introduce the Remote Instrumentation Idea Into Practice". Therefore it lays a corner-stone for the implementation of the prototype conducted in WP6.

The main task of this deliverable is to find out which networking technologies can be used to operate the instruments. However, RINGrid is about using a multitude of instruments as grid

components, and each of the instruments has different characteristics. Some of them may need CPU consuming pre- or postprocessing steps, some of them could need high-bandwidth video streaming for displaying the results, and even others may be in need of a low-latency link to be operated properly. Depending on the device there may be an emphasis on visualization, high-efficiency data processing or transmission of the data to the researcher's institution. If multiple researchers are collaborating there may be a need for storing multiple gigabytes of data or multicasting video streams.

Because of the sheer data volume and the real-time properties of video transmission the demands on the network infrastructure are high. In order to be able to assess whether the currently available technologies are sufficient for reaching our project goals, we need to know more about the instruments we want to connect via the grid. This is where workpackage two comes into play: WP2 has the task of identifying relevant instruments and user communities, and gathering information about these. With the help of WP2 we can for example tell what latency instruments require and which technology we have to choose to achieve our goal. However, since it took very long to retrieve the relevant data, it could not be considered in this deliverable. It will be taken into account in the next deliverable D3.2 and eventually in updated versions of this deliverable.

1.2. Definition of the Term "Grid"

In order to understand what it means to deal with "remote instrumentation in the grid", one needs a proper understanding of the term "grid". Using the term "grid" when we talk of combining computing (or storage) power is relatively new — the grid concept however is known for long by the names "metacomputing" or "distributed supercomputing".

Finding a definition for "grid" is not easy. There are several definitions around, but probably the most comprehensive was given by I. Foster, C. Kesselman and S. Tuecke in 2001 [FOKT01]:

> The real and specific problem that underlies the grid concept is coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organisations. The sharing that we are concerned with is not primarily file exchange but rather direct access to computers, software, data, and other resources, as is required by a range of collaborative problemsolving and resource-brokering strategies emerging in industry, science, and engineering. This sharing is, necessarily, highly controlled, with resource providers and consumers defining clearly and carefully just what is shared, who is allowed to share, and the conditions under which sharing occurs. A set of individuals and/or institutions defined by such sharing rules form what we call a virtual organization.

In analogy to sharing resources like processing power or storage space, also resources of instruments can be shared, like observation time or results. Integrating these resources into the grid is one of the main tasks of RINGrid.

2. Classification of Networking Technologies

2.1. Research Approach

Because of the fact that high-speed networking is a diverse topic which evolves quickly, it can be hard to cope with even for those who have been studying this subject for years. Because of this we had to choose an approach that is not quite common for scientific work. It resembles the approach which is used when publishing a review article.

First we collected information about existing networking technologies. This was done by looking at networking journals and reading selected articles. While this approach gave us detailed information on several particular technologies, we were missing a broad overview of the topic. Additionally, most papers contained only experimental data without a prototype being publicly available. Therefore we also surfed different web pages, particularly of similar projects like GridCC, DataTAG, VLab and so on. We had a look at their underlying technologies, which gave better results since these projects have to use existing technology for implementing their prototype. That does not mean that we do not consider research technology, but for implementing our own prototype we are dependent on existing technology. Our abilities to engineer new middleware or even hardware are limited because of the short project duration.

The second step in researching the networking technologies was to find a suitable categorization for all the information we gathered so far. This was a difficult step since networking hardware has several properties which could all serve as a criterion for classification:

- The most obvious property is the speed of the network. While this property can easily be compared, it is hard to pinpoint a certain technology or a group of technologies. There exist several networking technologies, which can be operated at different speeds, and sometimes it is even hard to give a speed range. Additionally, grid middleware cannot be categorized by speed.
- Latency is the next most important criterion. However, latency can seldom be attributed to a particular technology, but most of the time it is a function of the whole network path. When talking about latency one must also consider the load on the network which comes from sources other than remote instrumentation. As a result of this, it is impossible to gather hard facts on latency during this early phase of the RINGrid project. Therefore, while latency is important, it is not usable for classification.
- The occurrence of the technology within the network layer is another option of finding a suitable categorization of the technologies. Indeed, it seems to be the only viable option since grid middleware can be seen as top-level part of the network stack.

There are other possibilities for classification, for example categorizing by properties such as Quality of Service (QoS). Again, these properties are not shared by all technologies and therefore are not universal enough. Most of the time it is hard to fit grid middleware into the classification.

"State of the Art in Networks" with respect to remote instrumentation is a big research field — it covers nearly everything from networking, the hardware itself, be it cable, fibre or lamba networks, the networking stack provided by the operating system or the grid middleware. The reason for covering all those topics lies in the fact that all network layers influence the result equally. For example, it does not help to use the best hardware available if the software stack introduces lags which are unacceptable for video transmission. Another reason for this detailed reflection is that it is only possible to know why a system behaves in a way it is observed when it is known how technology underneath works.

In the next chapter a short introduction to the classification, the surveyed network technologies and grid infrastructures is given.

2.2. Taxonomy

As previously stated we are basing the classification on network layers. However, not all of the 7 layers of the ISO/OSI model are important for the task of remote instrumentation. So, we consolidated several layers and found the following categories suitable for this deliverable:

- Hardware
  Cable, fibre and lambda networks are all suitable for transferring data to distant places. However, since all major network backbones are all optical we have decided to leave out cable.
- Switching
  Since switching introduces the first latencies, it is an important topic to discuss. When handling gigabytes of data in a relatively short period of time over an international network link it is crucial that every packet has to know where it should go. There are approaches to reduce the switching time (see for example Optical Burst Switching) or reduce routing to switching, which is accomplished forexample by using GMPLS.
- Network Layer
  Routing is basically accomplished by using the Internet Protocol version 4 (IPv4) or version 6 (IPv6). While the use of IPv4 is more common, IPv6 provides several enhancements which are especially useful in our context.
- Transport and Application Layer
  Authentication and grid middleware fall both into this last category. They are responsible for the user experience and therefore serve the same purpose.

The categories, as well as the surveyed technologies, can be depicted graphically as seen in figure 1.

Application, Transport	SCTP: Stream Control Transmission Protocol for MPI	DCCP: Datagram Congestion Control Protocol	GTCP: Globus Teleoperations Control Protocol	UCLP: User-Controlled Light Paths
Network	IPv4	IPv6		
Switching	OBS: Optical Burst Switching	POS: Packet over SONET/SDH	GMPLS	RPR: Resilient Packet Ring	DRAGON
Hardware	CWDM, DWDM	G.709 ODU: Optical Data Unit	Lambda Networks	FSAN on G-PON

Figure 1: Surveyed network technologies

3. High-Speed Networking Base Technologies

3.1. Wave Division Multiplexing

Nowadays, and this will continue for the foreseeable future, the principal means of transferring high bandwidth information over medium and long distances are based on optical transmission systems using optical fibres as the transmission medium. Such systems can operate at transmission rates of up to dozens of Terabits per second (Tbps = 10^{12} bps). However, the electronic systems currently used as sources and sinks of the transmitted information are currently only capable of signalling rates of up to dozens of Gigabits per second (Gbps = 10^9 bps).

As we shall see in this section, the great potential capacity of optical transmission can be exploited by multiplexing multiple information streams onto a single optical fibre, while each of these streams uses a fairly narrow spectral bandwidth within the available optical spectrum. As these optical channels are identified by the wavelength of light used, the technique is called wavelength division multiplexing, and the channel is commonly referred to as a lambda, from the name of the Greek letter λ commonly used as a symbol for wavelength.

A good general reference for this section is the book by Ramaswami and Sivarajan [RAM02].

3.1.1. Optical Transmission Systems

The first study of propagation of light in optical fibres was published by Kao and Hockham in 1966 [KAO66], and it was soon realised its enormous potential for high bandwidth transmission. The first operational system was demonstrated in 1977, with the transmission of 140 Mbps [MID00]. Because of advances in optical technology during the last ten years, we will be able to exploit the full potential of optical transmission in the near future, with transmission systems operating in the Tbps range.

An optical transmission system is shown in figure 2. This is a simplex system, capable of accepting an input electrical signal at the transmitter and generating a corresponding output electrical signal at the receiver. At the transmitter, the output of an optical source is modulated to encode the information content of the electrical input signal. The resulting optical beam is injected into the transmission medium, which is an optical fibre. Depending on the distance of propagation, it may be necessary to regenerate the optical signal at one or more intermediate points. At the receiver, an optical detector is used to convert the optical signal once again into an electrical signal.

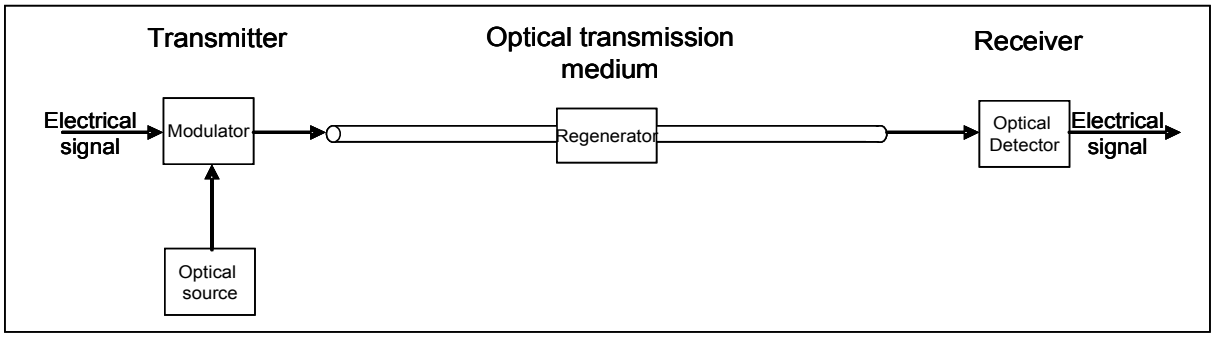


Figure 2: Optical transmission system

3.1.2. Optical Fibres

An optical fibre behaves as a waveguide for light, and most commonly consists of a slender flexible strand of silica (highly pure glass), known as the core, and encased in a coating of silica with lower refractive index, known as cladding. Typically many (up to several hundred) strands are bundled together in a single optical cable with appropriate mechanical properties to support and protect the optical fibres and possible to permit aerial suspension.

An optical fibre behaves like a waveguide due to the difference of refractive index between the core and the cladding, which, in accordance with Snell's Law, provides total internal reflection for angles on incidence sufficiently parallel to the axis of the fibre (see figure 3). Typical diameters are 50 or 62.5 microns (10^{-6} metres) for the core and 125 microns for the cladding. For comparison, human hair is between 40 and 120 microns in diameter. This kind of fibre is called multimode (MMF), and uses a light-emitting diode (LED) as an optical source. In these fibres, many modes (light beams with different angles of incidence) can coexist.

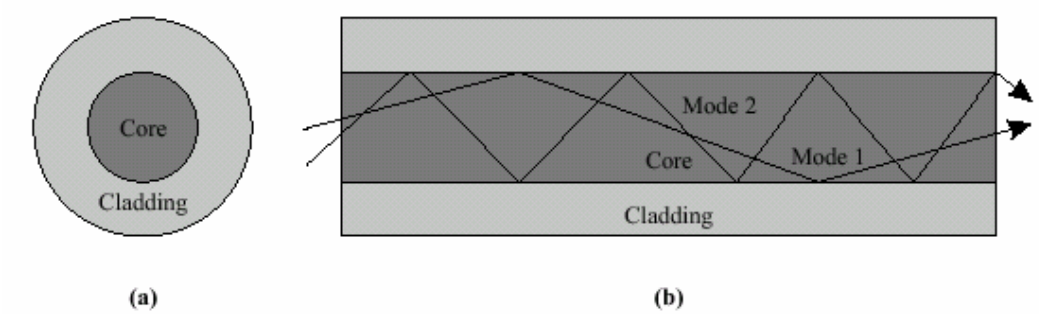


Figure 3: Transverse (a) and longitudinal (b) sections of multimode optical fibre

If the core is reduced in diameter to 9 or 10 microns (see figure 4), then for the wavelengths of light commonly used in optical transmission systems, the entire optical energy is concentrated in a single mode of transmission, parallel to the axis of the fibre. This kind of fibre is called single mode (SMF) and is the most common in use today. In this case, a laser must be used as the transmitter light source.

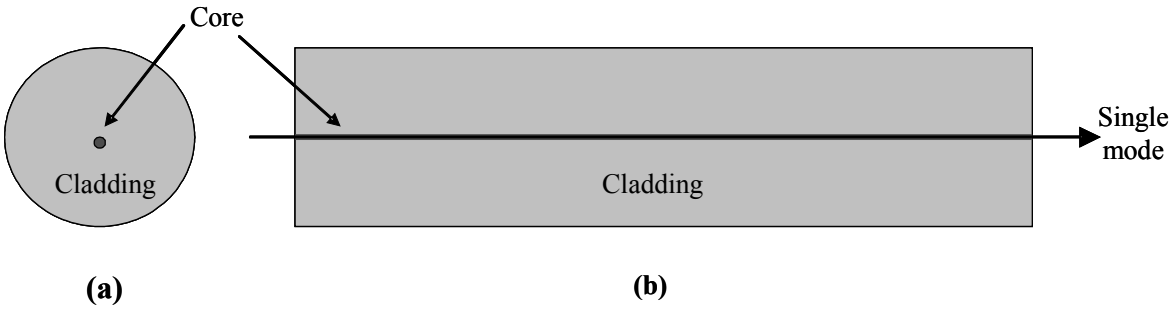


Figure 4: Transverse (a) and longitudinal (b) sections of single mode optical fibre

The choice of the transmission frequency, or wavelength, which is equivalent, is chosen to optimise the transmission. In figure 5 (see [MUR02, RAM02]) the attenuation characteristics are shown for standard silica fibres (G.651 and 652). It will be clearly noticed the existence of two windows of least attenuation of around 2 dB/km (50% attenuation in 16 km) around the

wavelengths of 1310 and 1550 nanometres (nm, or 10^{-9} m). These two optimum windows have a combined width of around 50 THz, from which we deduce a theoretical limit of 50 Tbps for the transmission bandwidth.

It is also important to note that, although we use the terms "light" and "optical", the interesting part of the electromagnetic spectrum of interest to us here is in the near infrared, and is thus not visible to us humans. Visible light has wavelengths between 400 and 700 nm.

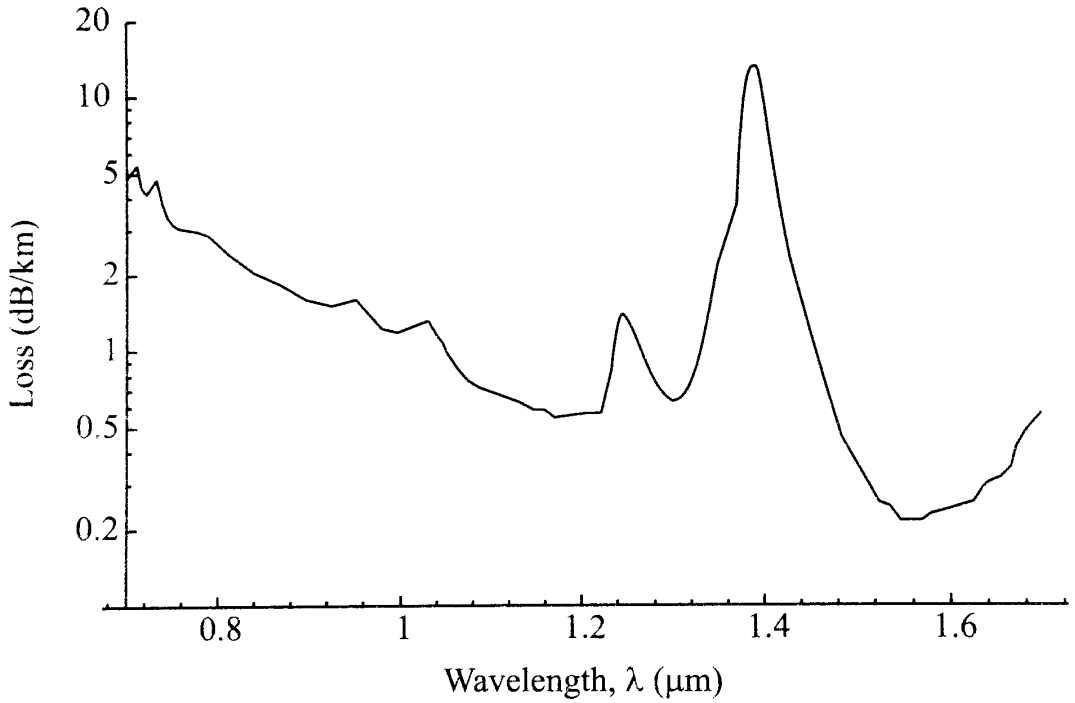


Figure 5: Light attenuation in an optical fibre

Table 1 shows the names which are often used to refer to spectral bands within the region of minimum attenuation. It should be noted that new kinds of fibre are now manufactured eliminating the attenuation peak in the E band.

Band	Descriptor	Range (nm)
O band	original	1260 to 1360
E band	extended	1360 to 1460
S band	short wavelength	1460 to 1530
C band	conventional	1530 to 1565
L band	long wavelength	1565 to 1625
U band	ultralong wavelength	1625 to 1675

Table 1: Spectral bands used in transmission in optical fibres

3.1.3. Limitations to Optical Transmission: Attenuation, Dispersion and Nonlinear Effects

Several separate factors limit the effective use of transmission in optical fibres. One of the most important of these is attenuation which reduces the power of the transmitted signal and places an upper limit on transmission distances. As we have seen, this depends on the wavelength used, and there is generally an upper limit of around 100 km with current detectors.

However, other factors may reduce the effective maximum transmission distance to well below this number. The various phenomena known collectively as dispersion also degrade the transmitted signal. Dispersion is due to variation in the propagation speed of different components of the optical signal, leading to a spreading out of initially well defined pulses so that they become more difficult or impossible to detect after a certain distance. Figure 6 illustrates the effect of dispersion. Clearly, the effect of dispersion depends on the signalling rate and on the distance travelled.

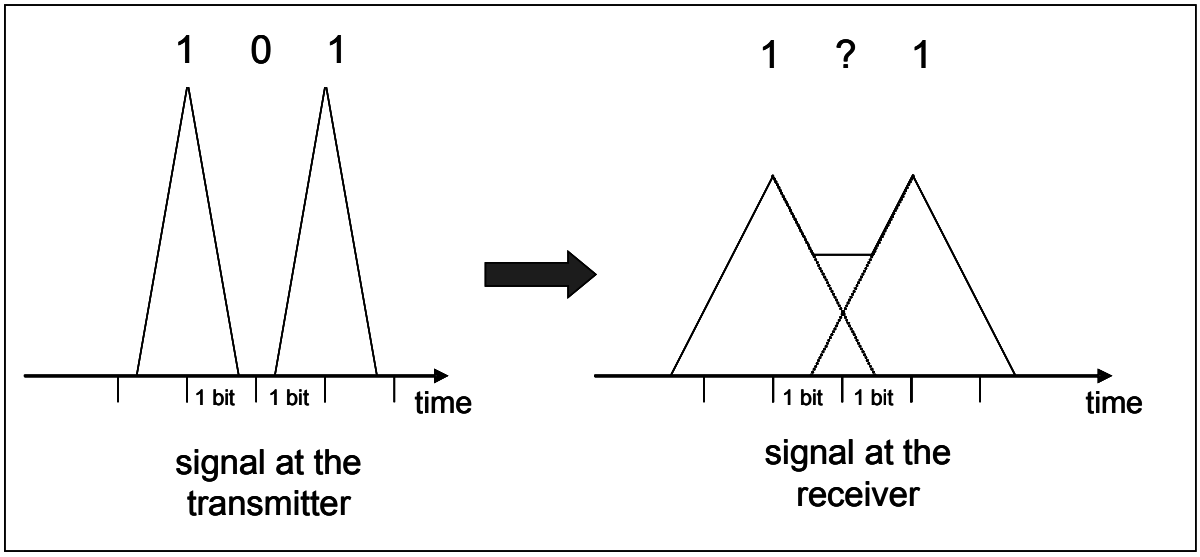


Figure 6: Dispersion. At the receiver the second bit value is uncertain

There are three common kinds of dispersion: intermodal, chromatic and polarisation-mode.

Intermodal dispersion is confined to transmission in multimode fibres, and is due to the distribution of the optical energy between different transmission modes, as shown in figure 6. The problem is that for modes at angle θ to the axis, the propagation distance is effectively L/cos θ, where L is the length of the optical link. In practical terms, this limits GigE transmissions to about 500 m (with GBICs SX), and 10 GigE to about 50 m, when MMF is used.

Chromatic dispersion is due to the variation of the speed of propagation of different colours of light in silica. Even with the use of laser sources, not all the light energy is concentrated at the central wavelength. In a cheaper laser, of the kind known as MLM (Multiple Longitudinal Mode), or Fabry-Perot, several sidebands exist at different wavelengths, and hence propagation speeds, as shown in figure 7. Such lasers are adequate for medium distances, to around 10 or 15 km, for GigE transmissions (using GBICs LX).

If we wish to support longer distances, or higher transmission rates, or both, we need to invest in better lasers, of the kind known as SLM (Single Longitudinal Mode) or DFB (Distributed Feedback). These concentrate all the transmitted energy at a single wavelength, as shown in

figure 7. Such lasers are used in GigE transmission at up to 100 km (using GBICs ZX). Such GBICs are currently between 4 and 6 times the cost of those using MLM lasers.

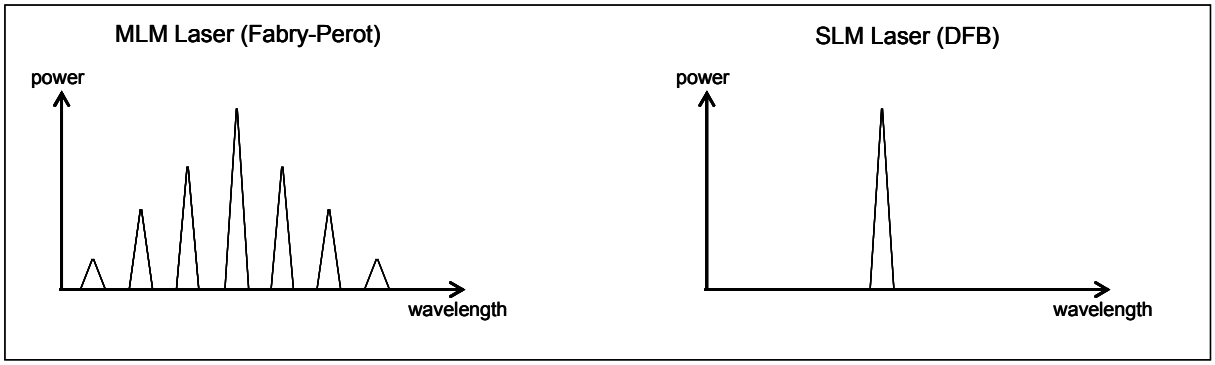


Figure 7: Power spectra of MLM and SLM lasers

Another way of reducing the impact of chromatic dispersion is to use a different quality of optical fibre. In standard single mode fibre (ITU standard G.652), chromatic dispersion is at a minimum in the O band, near to 1310 nm. To provide better support for DWDM (see below), most long-distance transmission systems nowadays utilise the C and L bands. To reduce chromatic dispersion at these wavelengths, the ITU has defined standards G.653 and G.655, in which the minimum chromatic dispersion has been shifted to higher wavelengths. A comparison of G.652 with G.655 shows an increase from 1000 km to 6000 km in the maximum reach of data transmitted at 2.5 Gbps, using just 1R regeneration (see below) to combat signal attenuation.

Polarisation-mode dispersion (PMD) is a phenomenon which begins to be important for transmission rates of 10 Gbps and above, and is due to irregularities in the physical shape of the fibre core used, which is not perfectly cylindrical. This results in different propagation speeds for the two polarisation mode components of the transmitted energy. Sometimes PMD is corrected by post-processing both of the polarisation modes of the arriving signal.

One way to counteract the effect of attenuation is to increase the laser power used. However, in this case new problems arise, due to the higher power used, and are known as nonlinear effects, which include Kerr effects and scattering effects.

Kerr effects include three phenomena where the refractive index of the fibre becomes a function of the signal power: in self-phase modulation, energy may be transferred from one wavelength to other neighbouring wavelengths; in cross-phase modulation, transmission in two separate wavelengths can interact to transfer energy to neighbouring wavelengths; in four-wave mixing, signals in two or more wavelengths can interact to generate a signal at a new wavelength.

Here are two important effects of scattering: stimulated Raman scattering, where the signal loses energy to molecules in the fibre, which is then retransmitted in longer wavelengths (with less energy); and stimulated Brillouin scattering, where energy is transferred to sound waves, which then go on to generate light waves in other wavelengths.

In general, nonlinear effects interfere with signal propagation, limiting the distance. On the other hand, stimulated Raman scattering has beneficial consequences, which are used in Raman amplification, discussed below. Nonlinear effects are an active research topic.

3.1.4. Regeneration of the Optical Signal

As we have seen, there are limits to the maximum distance that an optical signal can be transmitted, imposed by attenuation, dispersion and nonlinear effects. In small networks, for instance campus or even metro networks, the reach provided by optical systems is sufficient. However, long-distance transmission systems require us to use one or more signal regenerators along the route between the transmitter and the receiver.

Three kinds of signal regeneration can be performed: re-amplification, re-shaping and re-synchronisation. Re-amplification merely boosts signal energy, maintaining any distortion already present; reshaping restores the format of the signal, and resynchronisation adjusts the timing, as shown in figure 8. In general, regenerators either only deal with re-amplification, in which case the are known as 1R regeneration. Complete or 3R regeneration needs extensive electronics and is thus much dearer.

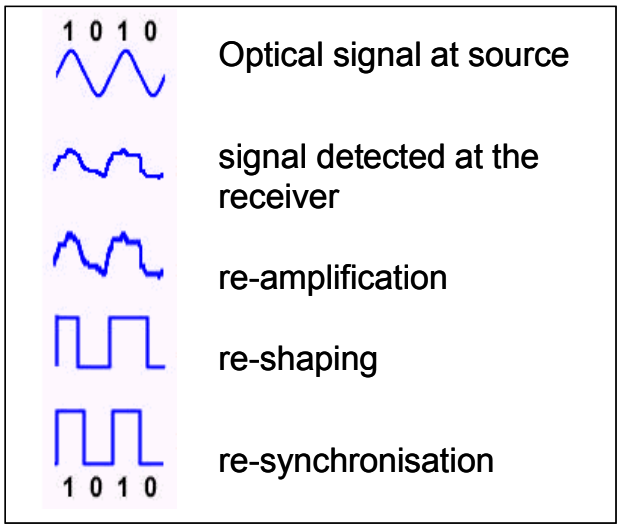


Figure 8: Signal regeneration

The simplest 1R regenerator is a stand-alone transponder, which boosts the incoming optical signal through an OEO (optical-electrical-optical) conversion. Note that such a transponder can also alter the wavelength or even the range of the outgoing signal – this merely depends on the output laser used.

3R regeneration also occurs in an electronic L2 or L3 switch, which may be needed at a particular point for dealing with local traffic needs. In this case, 3R regeneration comes for free in this scenario. This is also an OEO conversion.

Since the 1990s, much interest has been displayed in purely optical regeneration, without the need to convert the optical signal into the electrical domain at all. In this case we are dealing only with 1R regeneration.

The oldest and most interesting of these is Erbium-Doped Fibre Amplification (EDFA). In EDFA, a length of several dozen metres of erbium-doped fibre is spliced into the fibre whose content we wish to regenerate. We also need a pump laser, transmitting at 980 or 1480 nm. The signal to be amplified, which must have a wavelength in the neighbourhood of 1550 nm, is combined with the output of the pump laser and injected into the erbium-doped fibre, as shown in figure 9. The stimulation of erbium atoms by the pump laser radiation causes photon emission in the wavelength of the incoming signal, contributing to its amplification. Figure 10 (taken from [MUK00]) shows the gain as a function of wavelength. It is highly significant that

there is approximately a uniform gain in the range between 1520 and 1560 nm, which coincides with the middle of the second window of low attenuation. Note that EDFA is a broadband amplifier, which permits its simultaneous amplification of independent transmissions on different wavelengths within the C band.

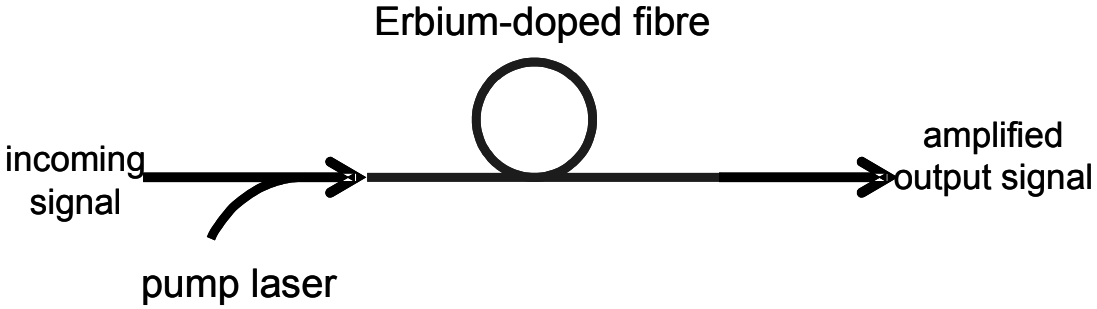


Figure 9: Erbium-Doped Fibre Ampification (EDFA)

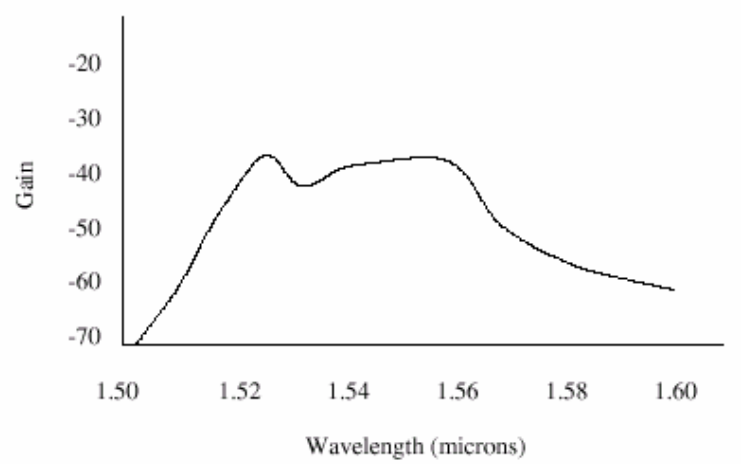


Figure 10: Gain as a function of wavelength for EDFA

A second technique of optical amplification is called Raman. This also uses a pump laser, but does not require the use of special (doped) fibre. The amplification occurs over a distance of several kilometres, and can be undertaken pumping in the same direction as the incoming signal (co-pumping), in the opposite direction (contra-pumping) or in both at the same time (co-contra-pumping). In order to amplify uniformly various wavelengths, it is necessary to pump at different wavelengths.

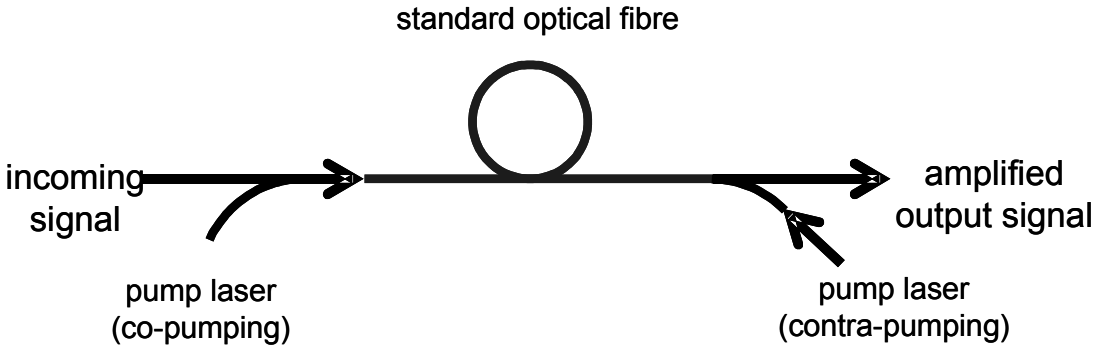


Figure 11: Raman amplification

With 3R regeneration, there are no upper limits to the extent of an optical network. With 1R generation there are some limits. Since 1R regeneration does not fix problems of signal distortion and lack of synchronisation, then there distances beyond which one cannot transmit without requiring 3R regeneration. This limit is receding. For standard G.652 fibres, it is around 800 to 1000 km. For the communications systems of most countries, it is highly unlikely that such long runs will occur without the need for electronic switching, with associated 3R generation. The serious problems come with very long-distance communications, such as across the oceans. Such transmission systems use recently developed fibres, such as G.653 and G.655, where it is possible to transmit longer distances without the need for regeneration.

3.1.5. Increasing the Bandwidth: More Fibre, Higher Signalling Rate, Multiplexing

As we have already pointed out, the theoretical bandwidth limit of optical transmission is around 50 Tbps. Nevertheless, such capacities have never been used for single communication channels between electronic endpoints, since no one has yet discovered how to make electronic circuits which work at these speeds. Current L3 routers do not go beyond 10 Gbps, SDH switches using 40 Gbps are already available, and 100 Gbps Ethernet is on the way. But these are orders of magnitude slower than the potential available.

In principle, we can generate higher total traffic by aggregating several separate parallel high-bandwidth channels. There are two ways of doing this. If we have abundant fibre at our disposal, we can simply allocate separate fibre strands to the different channels we are working with. If the fibre is already deployed, then this is straightforward. However, if we need to install the extra fibre, we have a serious problem, and we should look for another alternative, namely multiplexing individual communications channels onto a single fibre, using a different wavelength, or lambda, for each one. In any case, as we shall discover, there are situations in which so-called wavelength division multiplexing (WDM) is actually economically more interesting than maintaining multiple parallel communications channels in separate fibres. It should be noted that WDM is essentially the same thing as Frequency Domain Multiplexing (FDM), commonly used in telephone access networks, as well as in broadcast radio and television, and in cable TV systems, where is channel is permanently allocated a fixed-size chunk of the frequency spectrum.

There are two varieties of WDM. Coarse WDM is used for short distances (up to 100 km), such as in metropolitan networks, where no regeneration is needed. Dense WDM is used in long-distance networks, where regeneration is essential, and can be provided cheaply using EDFA or, more recently, Raman amplification.

Figure 12 shows the basic components of a WDM system, although details vary between them. In this figure, line amplifiers are shown, but these are only used with Dense WDM, because of the limitations on where broadband amplification can be used.

Typically the lasers used in DWDM systems are transponders that are an integral part of the transmission system. In CWDM systems for GigE networks, it often possible to buy "coloured" GBICs to generate the appropriate wavelength directly by the terminal equipment. It is important to note also that the two couplers (multiplexer and demultiplexer) are passive devices, and are therefore relatively cheap.

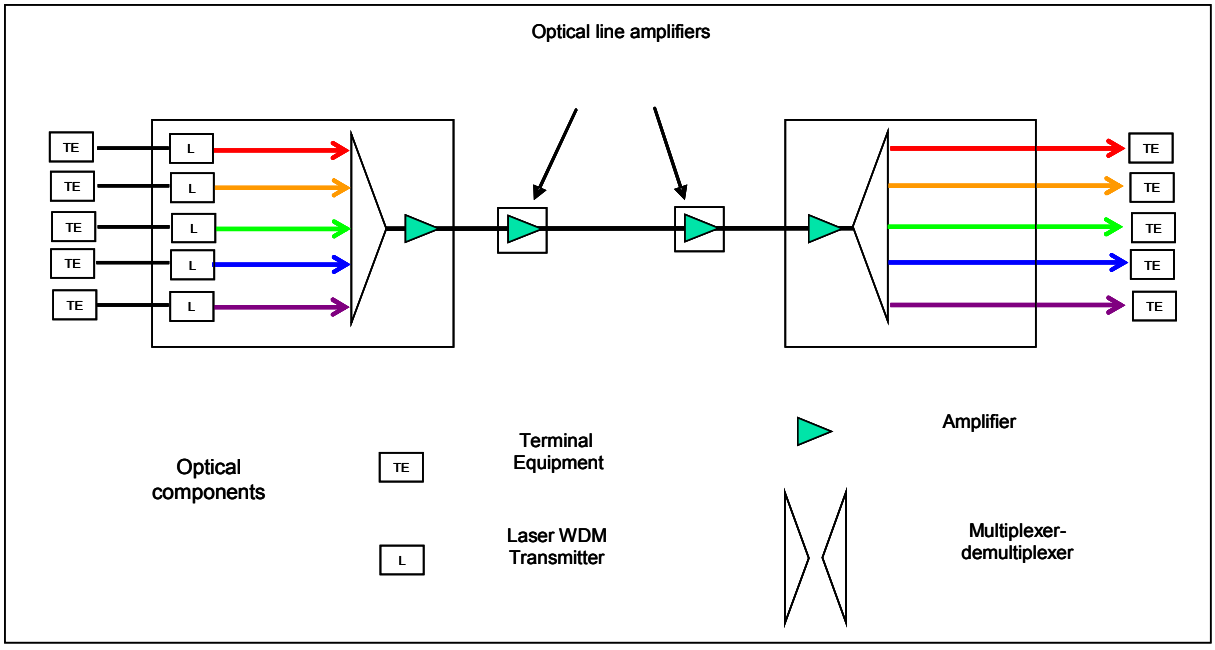


Figure 12: WDM transmission system

3.1.6. Coarse Wave Division Multiplexing (CWDM)

CWDM has been designed to be low-cost. This is achieved by using a wide spacing between the different channels so that cheap (MLM) lasers can be used. The standard inter-channel spacing is 20 nm, and most systems provide up to 8 channels in the E, C and L bands of the second minimum attenuation window (1450 to 1630 nm). In case of need, a further 4 channels can be located in the O band. New fibres, such as Allwave from Lucent (1998), where the attenuation peak in the E band has been eliminated, manage also to support a further 4 channels in this region. Thus it is feasible to install up to 16 channels in a CWDM transmission system.

As already mentioned, the price of a CWDM system can be lessened if the terminal equipment can be fitted with coloured transponders.

As no regenerators are used in a CWDM system, its maximum extension is usually no more than 80 to 100 km. For transmission distances greater than about 10 km, to combat chromatic dispersion, it will be necessary to utilise the more expensive SLM lasers.

3.1.7. Dense Wave Division Multiplexing (DWDM)

The invention of EDFA made broadband amplification of many wavelengths in the C band between 1520 and 1560 nm feasible. As a result, DWDM works with closely packed wavelengths in this spectral range. ITU has standardised channel spacing at 200 GHz (1.6 nm

between lambdas), 100 GHz (0.8 nm) and 50 GHz (0.4 nm). Typically the 50 GHz standard is used to support up to 45 channels, which facilitates the manufacture of EDFA amplifiers. Theoretically, higher total system capacity is theoretically possible, but so far this has not proved necessary in practice.

It should be noted that for DWDM the use of SLM lasers is essential. These lasers also require tight temperature control to maintain the stability of the wavelength of the emitted light.

3.2. Lambda Networks

Lambda networks is the name given to circuit-switched networks, where the switches are connected by WDM transmission links, and the unit of switching is an entire optical channel or lambda. Such networks are sometimes called lambda-switched. A lightpath is an end-to-end optical circuit in a lambda network. Lightpaths can be used to transport high-bandwidth packet-based communications over very long distances, without the need for intermediate packet-switching, either at L2 or L3, thus reducing the overall transport costs. An interesting combination of lambda networks with simultaneous L2 switching has made possible the deployment of hybrid L1/L2 networks, where one can effectively create end-to-end L2 virtual circuits, whose bandwidth is a fraction of the capacity of a single optical channel.

3.2.1. Electronic Switching: Circuit and Packet Switching

In electronic switching of communication links, two styles are common: circuit-switching (CS) and packet-switching (PS). In circuit-switched networks, such as the telephone system, a continuous end-to-end circuit must be set up before any communication can take place. This can be achieved manually, by operator intervention, or automatically, as in today's digital networks, where a signalling protocol (SS7) is used to reserve the required circuit, link by link. Call reservation will fail if there is no free channel to allocate on any of the links along the call path. Circuit-switched networks can lead to inefficient use of bandwidth resources, as the allocated bandwidth is available end-to-end for the duration of a call, independent of time-dependent user demand.

Packet switching was proposed to combat this very obvious inefficiency for bursty traffic in data networks, and achieves this by statistical multiplexing of packets along the links between the (packet) switches. Packets contain header information indicating the destination or route to be followed, and each packet is received, its header analysed, and then it is queued for transmission on the next link along its end-to-end path or route. Packet-switching has been a great success for data networks, and is the basis of the Internet.

3.2.2. Switching in Optical Networks: O-E-O and O-O-O

The first optical networks used electronic switching at internal nodes. The switches used are referred to as O-E-O, as an abbreviation for Optical input – Electronic switching – Optical output. Both circuit- and packet-switched networks can easily be set up with appropriate switches. It should be noted that such switches naturally perform 3R regeneration on the optical input signal. It is also possible to change the colour (wavelength) of the optical channel used on the next link, as the optical signal is recreated at every switch.

An interesting alternative is the all-optical network, where the switching is performed without expensive conversion to the electrical domain of the entire information stream. Such switches are referred to as O-O-O.

With the advent of all-optical networks, both of circuit switching and packet switching have been tried out. Optical circuit switching, usually accompanied by WDM links between the switches, has led to lambda networks. In lambda networks, the circuits are usually set up using out-of-band network management, but recent developments, such as Multi-Protocol Lambda

Switching (MPλS) and Generalised Multi-Protocol Label Switching (GMPLS), have introduced a signalling protocol for this purpose. It should be noted that the resulting lightpath is an analogue end-to-end channel, permitting the use of any rate of signalling which does not lead to dispersion problems in the optical transmission infrastructure. Note that, if no measures are taken to avoid this, the same colour (wavelength) of optical channel will be used on every link. This situation can be avoided by the use of an optical wavelength converter. This will be discussed below.

All-optical packet-switched networks have never appeared outside a small number of networking testbeds, as there are very severe conditions for handling packet headers in a high-speed optical network. Firstly, the header must be read and decoded at line speed, which requires expensive electronics. During this process, the content of the packet body must be stored until it is possible to retransmit the packet along the appropriate output link and channel. No simple means of storing optical packets is known, and the best available alternative is an optical delay line [MAS93]. Experimental switches, such as in the KEOPS project [GUI98], tackled both these problems. In KEOPS, packets and packet headers were of fixed length, and the header information was coded at a significantly lower bit-rate than the packet body, thus lowering the demands on the (electronic) header processing. The packet body circulated in a fixed-time delay line.

To eliminate the need for packet storage, whilst maintaining the flexibility of packet switching, some authors have built burst-switched networks. In these, a set of packets for the same final destination is injected into the network in a single burst. In order to prepare the internal switches to correctly route the burst, an out-of-band burst control packet is dispatched ahead of the burst. This packet has to reach all the switches along the burst route sufficiently before the arrival of the burst. If it is not possible to redirect the burst at a given switch, because no output channel is available, the entire burst is lost. Optical burst switching is less demanding than optical packet switching, but neither has led to commercial products. Optical burst switching is discussed in more detail in a later section of this report.

3.2.3. Switching Elements in a Lambda Network

There are basically two kinds of switching element used in lambda networks: the optical add-drop multiplexer (OADM) and the optical cross-connect (OXC). These function in a similar fashion to the corresponding switches in SDH networks, but at the granularity of an entire optical channel.

An OADM is normally used to connect terminal equipment to a lambda network, by switching a single lambda (or possible several neighbouring lambdas), to a derivative output link (drop). It can also simultaneously insert one or more input lambdas of the same wavelength(s) from the derivative link, to maintain the complement of lambdas in the main fibre. Figure 13 (a) shows the adding and dropping of lambda 2 from the fibre A.

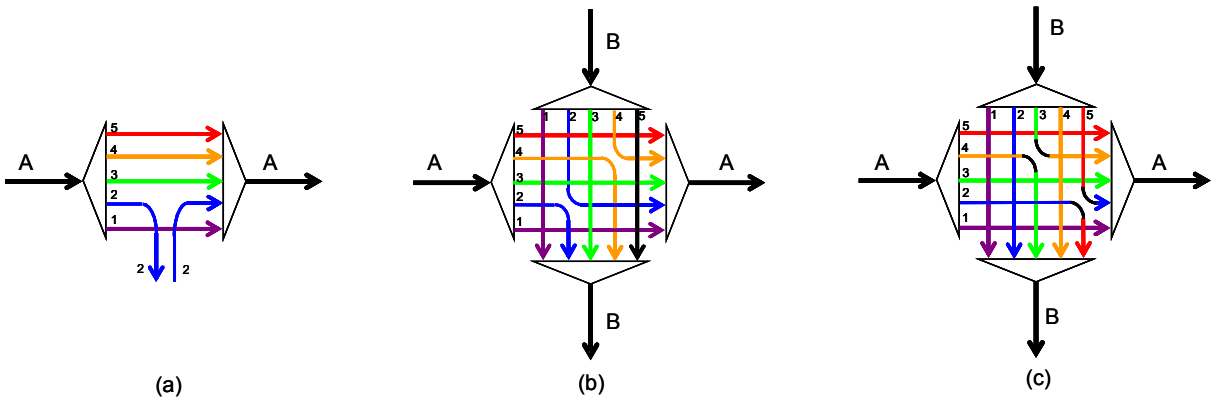


Figure 13: (a) OADM, (b) OXC, (c) OXC with wavelength conversion

A variation on the OADM is the Reconfigurable OADM or ROADM, where the choice of which lambda(s) to add and drop can be altered dynamically. This makes for a much more flexible way of building lambda networks.

An OXC is a circuit switch with two input and two output fibres, and lambdas from the input fibres can be switched to either of the output fibres, as shown in figure 13 (b). Here we have to be careful, as we cannot direct two input lambdas of the same colour to the same output fibre. One solution to this problem is the use of wavelength converters, which alter the wavelength of a given input lambda, as shown in figure 13 (c).

Various alternatives are known for manufacturing OADMs and OXCs. These include multiplexer and demultiplexor components made from thin film filters, fibre gratings with optical circulators, free space grating devices and integrated planar arrayed waveguide gratings. The switching functions range from the manual fibre patch panel to a variety of switching technologies including MEMS, liquid crystal and thermo optic switches in planar waveguide circuits. The use of MEMS (Micro Electro-Mechanical Systems) devices are shown in figure 14, where tiny mirrors are raised and lowered in (a), or are deflected in (b) [DOB02].

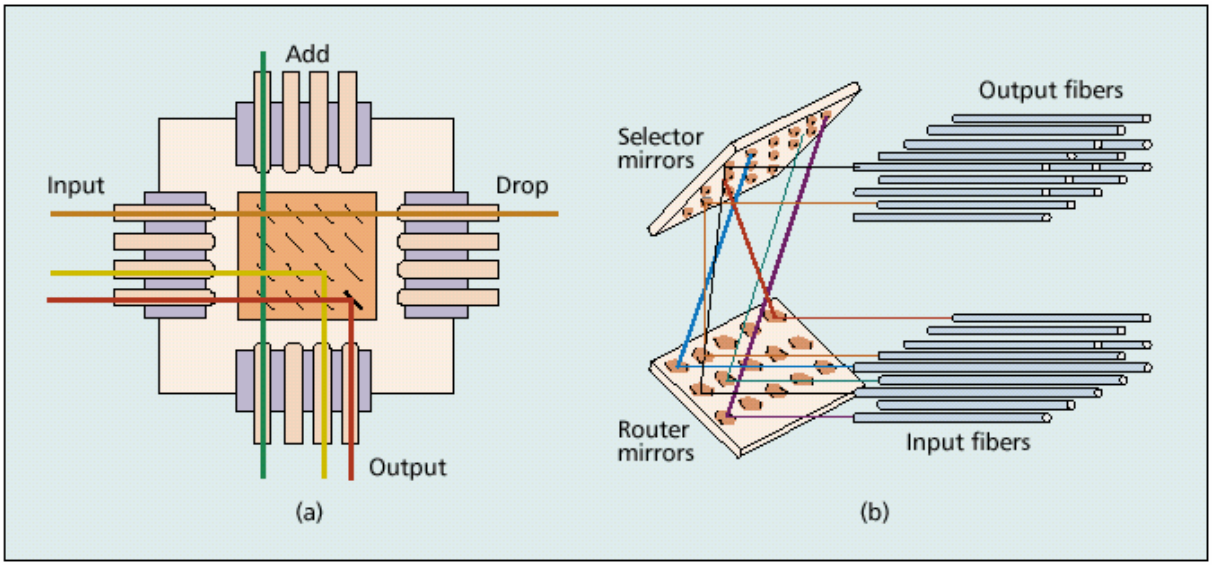


Figure 14: Use of MEMS devices for (a) OADM, (b) OXC

3.2.4. GLIF: An Operational Lambda Network

The GLIF website www.glif.is/about uses the following three paragraphs to describe itself:

"GLIF, the Global Lambda Integrated Facility, is an international virtual organisation that promotes the paradigm of lambda networking. The GLIF participants are National Research and Education Networks (NRENs), consortia and institutions working with lambdas. Participation in GLIF is open to any organisation that subscribes to the GLIF vision and can contribute to the GLIF actvities. GLIF was established at the 3rd LambdaGrid Workshop in Reykjavik, Iceland, in August 2003.

The GLIF community shares a common vision of building a new grid-computing paradigm, in which the central architectural element is optical networks, not computers, to support this decade's most demanding e-science applications. This paradigm is based on the use of parallelism, as in supercomputing a decade ago. However, with GLIF the parallelism is in multiple wavelengths of light, or lambdas, on single optical fibres.

GLIF is interested in developing "application-empowered" networks, in which the networks themselves are schedulable Grid resources. These application-empowered deterministic networks, or "LambdaGrids", complement the conventional networks, which provide a general infrastructure with a common set of services to the broader research and education community."

In networking terms, GLIF has set up an international lambda network, based around a number of lambdas, contributed by the GLIF participants who own or lease them, which are interconnected through a series of exchange points, known as GOLEs (GLIF Open Lightpath Exchanges). GOLEs are usually also operated by GLIF participants, and are comprised of equipment that is capable of terminating lambdas and performing lightpath switching, such as Nortel HDXc or Cisco ONS 15454. This way, different lambdas can be connected together, and end-to-end lightpaths established over them.

The list of GOLEs and lambdas can be found at www.glif.is/resources and at the time of writing includes the following participating sites:

- AMPATH - Miami
- CANARIE-StarLight - Chicago
- CANARIE-PNWGP - Seattle
- CENIC - Los Angeles
- CERN - Geneva
- CzechLight - Prague
- HKOEP - Hong Kong
- KRLight - Seoul
- MAN LAN - New York
- NetherLight - Amsterdam
- NorthernLight - Stockholm
- Pacific Northwest GigaPoP - Seattle
- StarLight - Chicago
- T-LEX - Tokyo
- UKLight - London

Each GOLE will be connected to one or more neighbours through optical links. The two most connected GOLEs in GLIF are Starlight and Netherlight. Figure 15 shows the connection topology of Netherlight (taken from [NET06]).

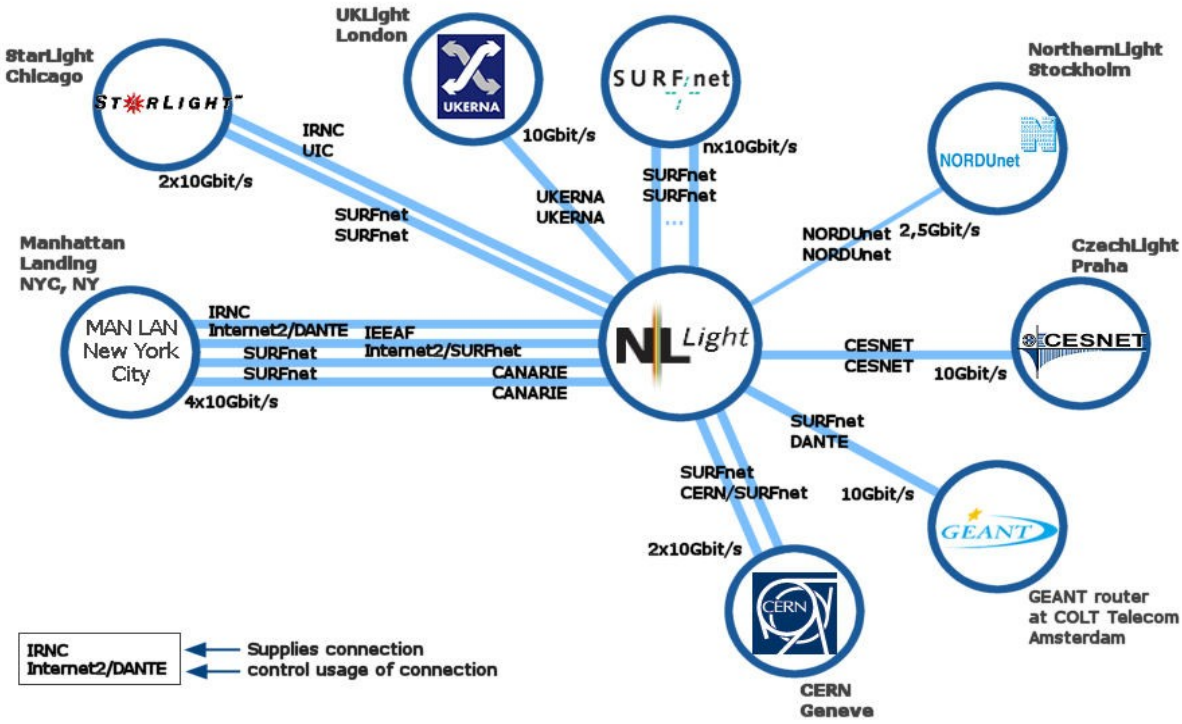


Figure 15: Topology of the Netherlight GOLE in Amsterdam, Netherlands

Each GOLE also provides access to a national or regional lambda network, and can switch lambdas between this local network and the GLIF international network, to enable an end-to-end lightpath between clients in two different countries. In the Netherlands, SURFNET6 maintains a national network based on a Common Photonic Layer (CPL) constructed from one closed and three open DWDM rings, or subnetworks, connecting one or two sites in Amsterdam with the other clients. Each client has dedicated access to several lambdas in its subnetworks, which connect it to one or both of the Amsterdam sites (see figure 16, taken from [NEG04]). Note this figure also shows a fifth subnetwork interconnecting the client sites at Eindhoven and Nijmeegan. At the Amsterdam sites, lambdas can be switched to other national clients, or to the Netherlight facility, in the case of international lightpaths (see figure 17, taken from [NEG04]).

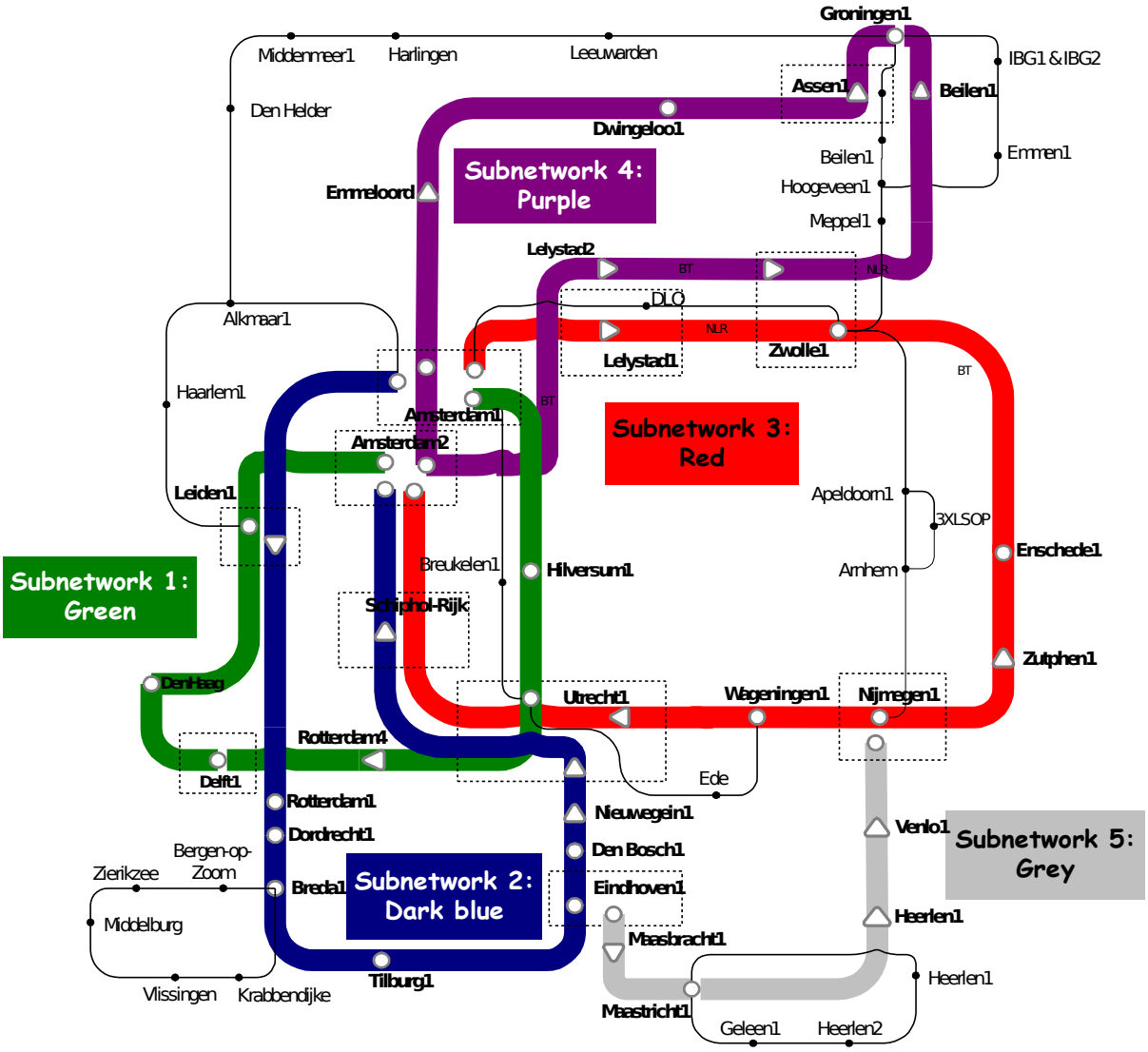


Figure 16: Common Photon Layer of SURFNET6, with subnetworks

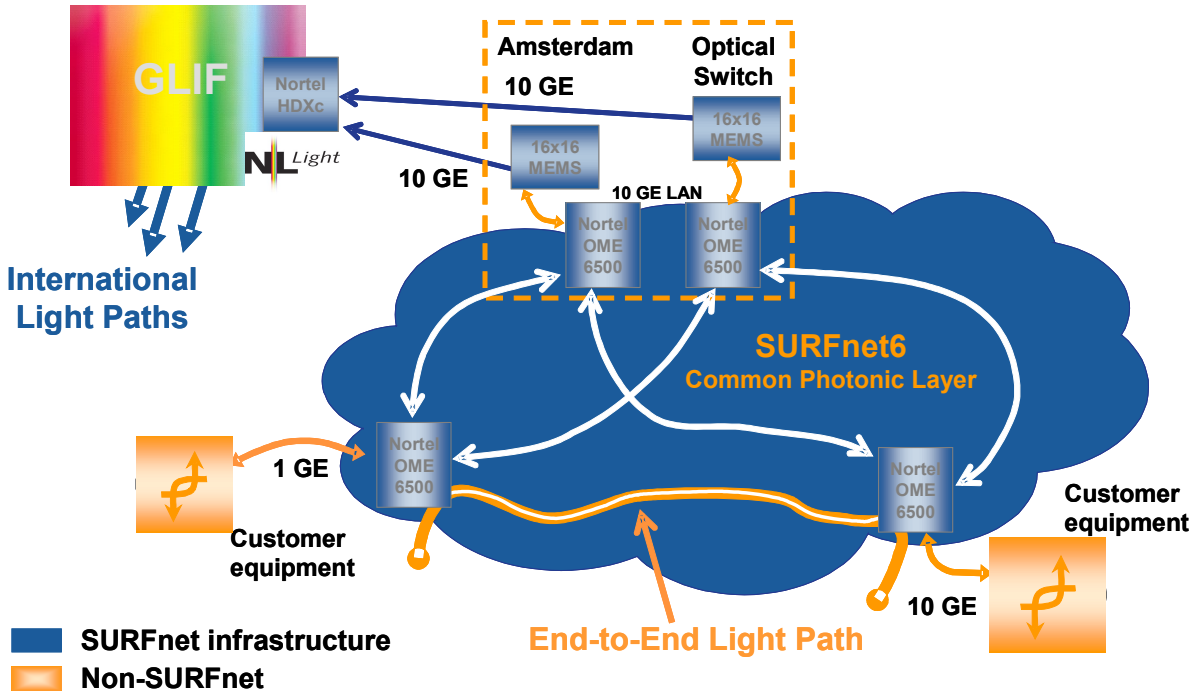


Figure 17: End-to-end lightpaths within SURFNET6 orusing Netherlight

3.2.5. Hybrid L1/L2 Networking

GLIF lightpaths provide end-to-end optical connectivity at high capacity, typically 2.5 or 10 Gbps. It is theoretically possible that this bandwidth be dedicated to a single information flow, but this would be very unusual. What is more common is that such a communication channel is typically occupied by an aggregate of smaller capacity "pipes", which share the same endpoints. Thus a GLIF lightpath could carry traffic between different applications at the endpoints, using some scheme of L2 multiplexing, such as Ethernet. The bandwidth challenges held at the annual supercomputing events are such an example, where the lightpath aggregates multiple end-to-end Ethernet links between one or more pairs of computers at the endpoints of the lightpath. So far, we are considering that this traffic is only demultiplexed at the destination endpoint of the lightpath.

The very interesting alternative of switching sublambda capacities at intermediate points has become very popular in recent years, because of the greater flexibility that it offers to lambda network providers. Essentially, each 10 Gbps capacity lambda is sliced into 8 Gigabit Ethernet (GigE) channels, and each 2.5 Gbps lambda into 2 GigE channels. In this view, the unit of bandwidth that can be allocated to end-to-end flows is 1 Gbps, corresponding to a single GigE channel. Then the switching nodes, instead of being limited to pure lambda switching, are also supplied with complementary GigE and 10GigE switches which are used to switch 1 GigE channels between different lambdas, or to aggregate multiple GigE channels into a larger 10GigE channel. A network with such switching nodes is referred to as a L1/L2 hybrid network.

Such networks have been set up in several different countries and continents. In the USA, networks like Pacific Wave and (more recently) Atlantic Wave have provided distributed exchange points in this way. Internet2's HOPI (Hybrid Optical and Packet Infrastructure) and GEANT2's end-to-end services are widely deployed examples of L1/L2 hybrid networking in the style described here. In both Internet2 and GEANT2, these services are being offered in parallel with the traditional L3 IP networks.

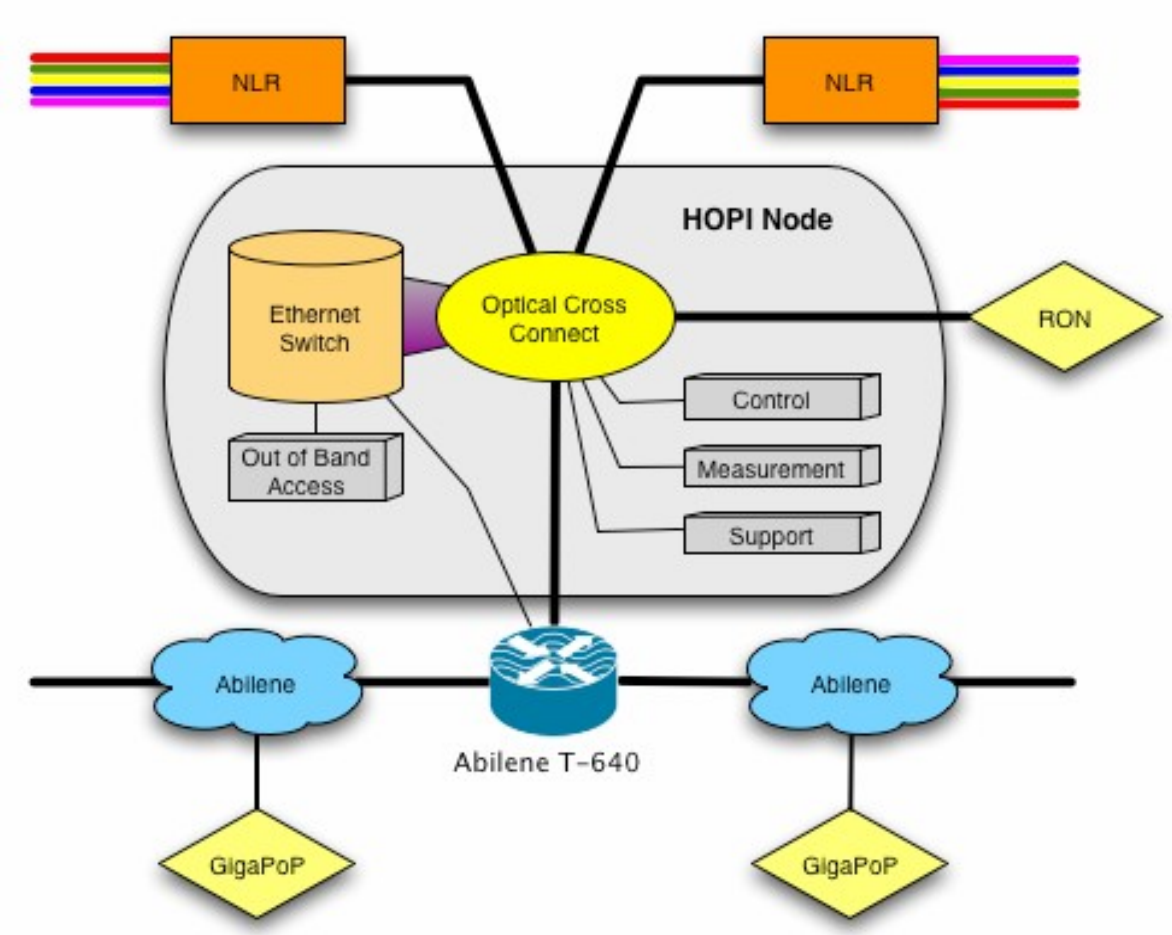


Figure 18: HOPI node in Internet2's national infrastructure

Figure 18 (taken from [SUM06]) shows the optical infrastructure provided by National Lambda Rail (NLR) and a Regional Optical Network (RON), and the L3 IP network Abilene. The HOPI node permits both lambda and sublambda switching within the optical infrastructure, and also between this and the IP network. The equivalent switching nodes in GEANT2 provide both switched IP and end-to-end services for the user community: In this case, the end-to-end services are extended into the NREN, when this is able to support the same hybrid model.

3.3. G.709 Optical Data Unit

The amount of data traffic relative to voice traffic on optical networks and the total traffic volume keeps increasing. These factors are the drivers behind emerging, flexible technologies to supplement the mature, voice optimized, SONET/SDH transport infrastructure and help manage network complexity. At the edge of the network, where data and voice combine in a common infrastructure, new data-centric applications have emerged. An example is the combination of virtual concatenation (VCAT), which provides flexible bandwidth groupings for SONET/SDH, Link Capacity Adjustment Scheme (LCAS), which provides dynamic bandwidth settings, and Generic Framing Procedures (GFP), which provides a protocol agnostic frame container. In the transport core, bandwidth requirements spawned the creation of the Optical Transport Network (OTN) described in general terms in ITU-T G.872. ITU-T G.709 provides the network interface definitions.

G.709 improves transport network performance and facilitates the evolution to higher backbone bandwidths. The G.709 OTN frame includes transport overhead that provides operation, administration, and maintenance capabilities, and Forward Error Correction (FEC). The FEC helps reduce the number of transmission errors on noisy links, which enables the deployment of longer optical spans.

In essence, the OTN consists of the following parts, which are often referred to as layers:

1. Optical Transport Section (OTS)
2. Optical Multiplex Section (OMS)
3. Optical Channel (OCh)
4. Optical Transport Unit (OTU)
5. Optical Data Unit (ODU)
6. Optical Channel Payload Unit (OPU)

Each of these elements and their functions are distributed along the network and activated when they reach their termination points, which are illustrated in figure 19 below.

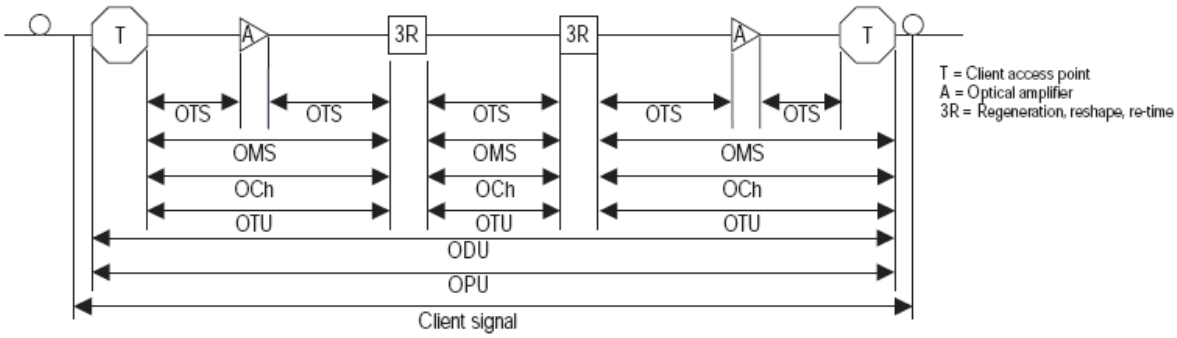


Figure 19: OTN layer termination points

3.3.1. Interfaces and Payload

G.709 defines standard interfaces and rates. These rates have been derived from the existing SONET/SDH rates where the G.709 overhead and FEC information have been taken into account. The resulting interfaces thus operate at line rates, roughly 7% higher, than the corresponding SONET/SDH that becomes the OTN payload. Table 2 lists the G.709 line rates and the matching SONET/SDH interfaces. An additional interface type, which is not part of the G.709 recommendation, applies to 10 Gigabit Ethernet LAN clients. In this case, the same overhead structure and FEC is applied resulting in a line rate of 11.095 Gbps.

G.709 Interface	Line Rate	Corresponding SONET/SDH Rate	Line Rate
OTU-1	2.666 Gbps	OC-48/STM-16	2.488 Gbps
OTU-2	10.709 Gbps	OC-192/STM-64	9.953 Gbps
OTU-3	43.018 Gbps	OC-768/STM-256	39.813 Gbps

Table 2: G.709 line rates

1. OC-48/STM-16 is transported via OTU1

2. OC-192/STM-64 is transported via OTU2
3. OC-768/STM-256 is transported via OTU3
4. Null Client (All 0s) is transported via OTUk (k = 1, 2, 3)
5. PRBS 2^{31}-1 is transported via OTUk (k =1,2,3)

In order to map client signals via ITU G.709, they are encapsulated using the structure illustrated in figure 20.

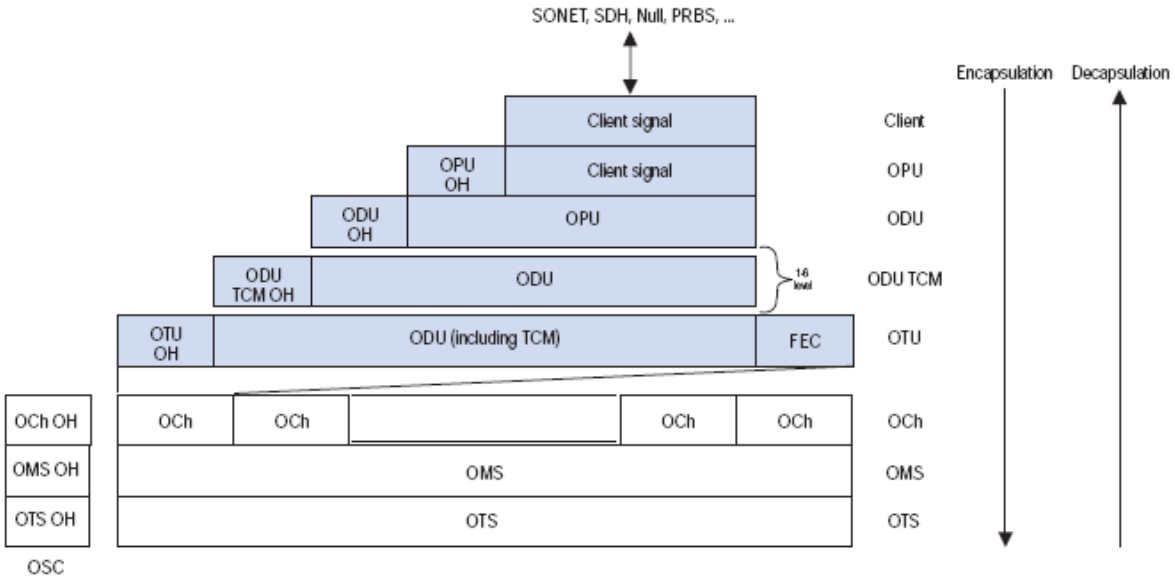


Figure 20: Basic OTN transport structure

3.3.2. ODUk Frame Structure

The ODUk (k = 1,2,3) frame structure is shown in figure 21. It is organized in an octet-based block frame structure with four rows and3824 columns.

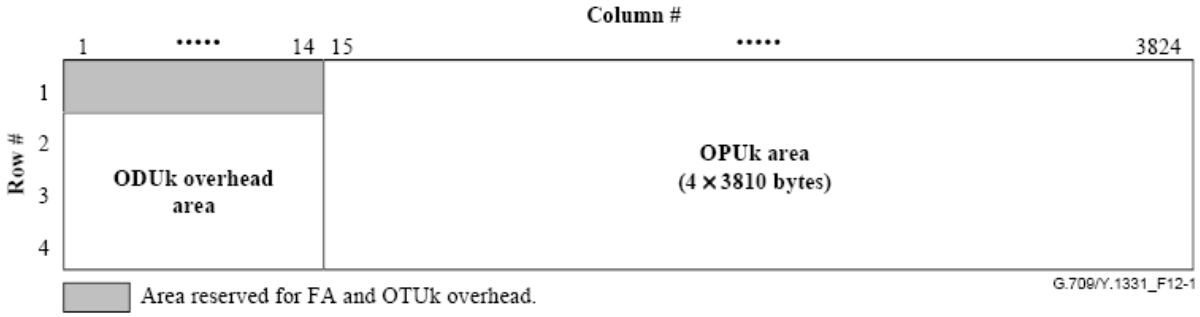


Figure 21: ODU frame structure

3.3.3. Optical Data Unit (ODU) Overhead

The ODU overhead is broken into several fields: RES, PM, TCMi, TCM ACT, FTFL, EXP, GCC1/GCC2 and APS/PCC.

- The reserved **(RES)** bytes are undefined and are set aside for future applications.
- The path monitoring (PM) field is used for the trail trace identifier (TTI), parity (BIP-8) and the backward error indicator (BEI), backward defect indicator (BDI), and Status (STAT).
- There are six tandem connection monitoring **(TCMi)** fields that define the ODU TCM sub-layer, each containing TTI, BIP-8, BEI/BIAE, BDI and STAT sub-fields associated to each TCM level (i=1 to 6). The STAT sub-field is used in the PM and TCMi fields to provide an indication of the presence or absence of maintenance signals.
- The tandem connection monitoring activation/deactivation **(TCM ACT)** field is currently undefined in the standards.
- The fault type and fault location reporting communication channel **(FTFL)** field is used to create a message spread over a 256-byte multiframe. It provides the ability to send forward and backward path-level fault indications.
- The experimental **(EXP)** field is a field that is not subject to standards and is available for network operator applications.
- General communication channels 1 and 2 **(GCC1/GCC2)** fields are very similar to the GCC0 field except that each channel is available in the ODU.
- The automatic protection switching and protection communication channel (APS/PCC) supports up to eight levels of nested APS/PCC signals, which are associated to a dedicated-connection monitoring level depending on the value of the multiframe.

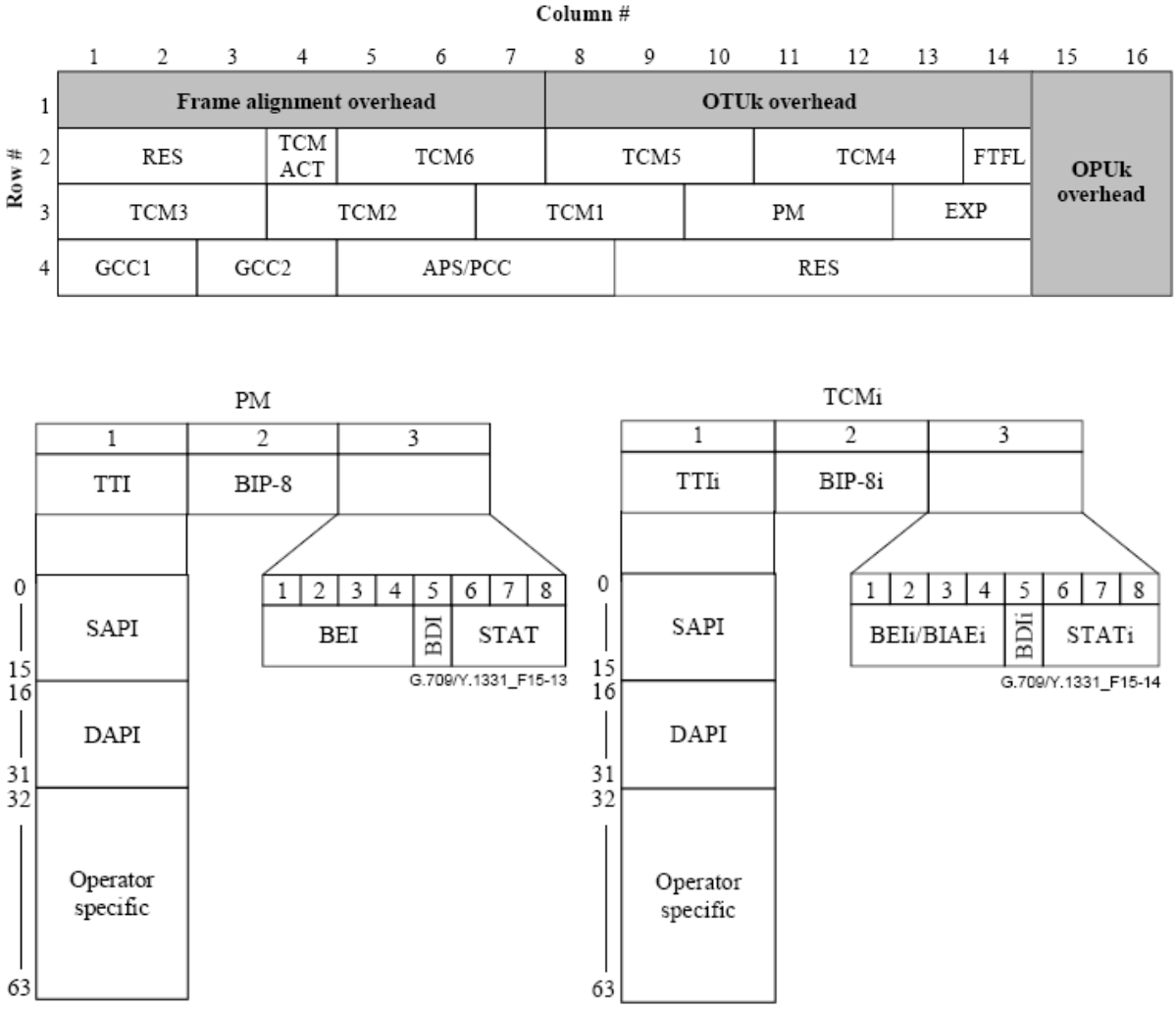


Figure 22: ODU overhead

3.3.4. Tandem Connection Monitoring (TCM)

TCM enables the user and its signal carriers to monitor the quality of the traffic that is transported between segments or connections in the network. SONET/SDH allowed a single level of TCM to be configured, while ITU G.709 allows six levels of tandem connection monitoring to be configured. The assignment of monitored connections is currently a manual process that involves an understanding between the different parties. There are various types of monitored connection topologies: cascaded, nested and overlapping. Examples of these topologies are provided in figure 23.

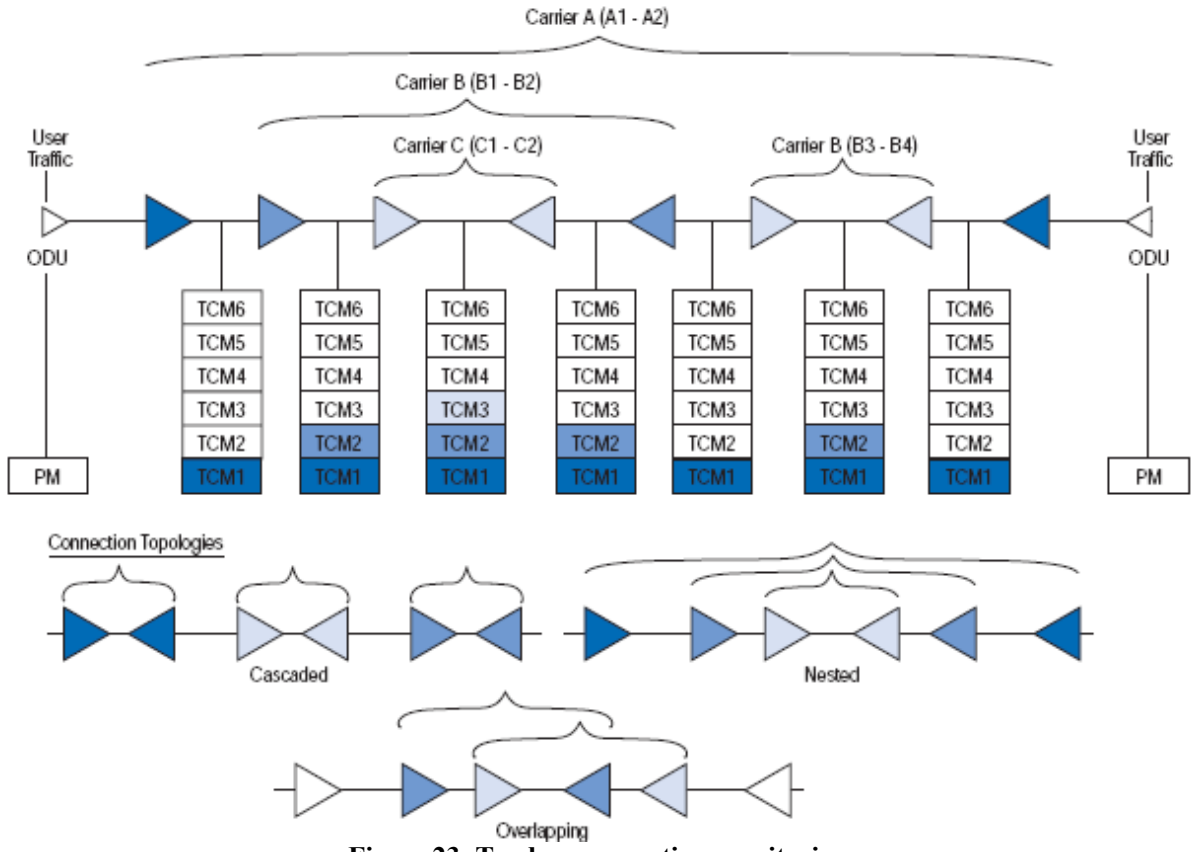


Figure 23: Tandem connection monitoring

Each of the six TCMi fields in the ODU overhead is assigned to a monitored connection. There can be from zero to six levels or connections that can be configured for each ODU trail. In the example from figure 23, there are three different levels that are actually monitored. Carrier C, due to its location, can monitor three TCM levels as the ODU passes through its portion of the network. The other carriers also monitor the levels according to agreement and location.

In addition to monitoring maintenance signals, using the STAT sub-field associated with each TCM level, the TCM connection also monitors the BIP-8 and BEI errors for each connection level. Maintenance signals are used to advertise upstream maintenance conditions affecting the traffic; errors, on the other hand, provide an indication of the quality of service offered at each segment of the network, which delivers a valuable tool for the user and carrier to isolate faulty sections of the network.

3.4. Full Service Access Network based on Gigabit Passive Optical Network Technologies

The Passive Optical Network (PON) provides an interesting option for operators to deliver very high-speed broadband access in a cost-effective way. Following the work done by the FSAN (Full Service Access Network) Group, the ITU-T published the first PON specifications in the G.983 series, which is now known as broadband PON (B-PON) and it is widely deployed [UEDPON, ITUPSV]. Then, following user migration to services based on variable-length packets, the IEEE introduced the Ethernet/gigabit Ethernet PON (EPON/GE-PON) standard as a PON system exploiting native Ethernet as the encapsulation layer. At the same time, FSAN was discussing an evolution of B-PON that could efficiently support all services with gigabit

line rates [FSANWEB]. Such a system, named gigabit-capable PON (G-PON), has been ratified by the ITU-T with the G.984 series [ITUGPON]. It is seen as the ultimate standard that could build an efficient optical access platform for operators to deliver high bandwidth with quality of service (QoS) guarantees to all types of customers, regardless of user protocol.

ITU-T recommendation G.984.1 [ITUGPON] specifies the generic characteristics of G-PON architecture. These are in part based on the G.983 series in order to allow backward compatibility with existing outside plant environments.

The G-PON architecture is based on an optical tree architecture where a central office equipment called OLT (optical line terminal) is passively linked to multiple user-side optical devices (called ONTs/ONUs, optical network terminations/optical network units) through a passive optical splitter.

In the downstream direction, the splitter broadcasts the optical signal coming from the OLT to all the subtended ONTs/ONUs. Although all ONTs/ONUs receive all downstream frames, secure channels are established to ensure each ONT only recovers the data flows destined for it. Each ONT/ONU is in charge of extracting flows addressed to it and ofdiscarding all others.

In the upstream direction, the splitter combines all flows coming fromthe multiple ONTs/ONUs of the same optical distribution network to structure a single flow in the feeder fiber without any overlap. A time division multiple access (TDMA) mechanism is used to avoid collision of optical signals that could happen at the splitter level. Each ONT/ONU has to restrict its transmission only to predefined time-slots by operating in burst mode under the arbitration of the media access control (MAC) protocol, which can statically or dynamically allocate one or several time-slots to each ONT/ONU. As specified in the G.984.3 [ITUTCONV] (framing format, ranging procedure, MAC protocol and security) a fixed 125 µs framing is used in the downstream as basis timing in the whole network. In the upstream, due to different propagation delays (different distances) among ONT/ONUs and OLT, a ranging procedure is performed by the OLT during the set-up phase to synchronize all the uplink signals.

One of FSAN's objectives is to remotely control each user traffic stream from the OLT by means of the MAC protocol to assure fairness, strict QoS guarantees, and traffic-profile control according to the service level agreement (e. g. per customer or per service stream). Thus, each flow at the ONT/ONU can be logically separated down to afine resolution level if required. The OLT has an effective control granularity of 64 kb/s in assigning bandwidth. The bandwidth allocation can be the same in every frame according to a static allocation, or different in every frame in response to dynamic traffic fluctuations. This latter option is the basis of dynamic bandwidth allocation (DBA), which enables dynamic/efficient sharing of upstream bandwidth. By operating the bandwidth distribution algorithm, the OLT's MAC controller executes the assignment of both the guaranteed and surplus parts of the bandwidth to the active queues. Each burst payload consists of a variable number of ATM and/or G-PON encapsulation method (GEM) frames. With GEM, it is possible to carry both Ethernet frames and TDM traffic in their native formats, as it is a modification of the packet transport protocol coming from SDH/SONET framework called generic framing procedure (GFP) and specified in ITU-T G.7041 standard.

The G.984.2 specifications [ITUPMD] focus on the physical media dependent (PMD) layer. As was the case for B-PON, G-PON may be either a one- or two-fiber system. The operating wavelength for the downstream direction on a single fiber system is in the range of 1480–1500 nm in conformance with G.983.3. For a two-fiber system, the downstream operating wavelength is in the range of 1260–1360 nm. The operating wavelength for the upstream direction is in the range of 1260–1360 nm on either a one or two fiber system. G-PON has seven transmission-line rates combinations: 1.244 Gb/s in downstream combined with 155 Mb/s; 622 Mb/s or 1.244 Gb/s in upstream or 2.488 Gb/s in downstream combined with

155 Mb/s; and 622 Mb/s, 1.244 Gb/s, or 2.488 Gb/s in upstream. The physical reach of the G-PON is up to 20 km and the maximum differential logical reach is also 20 km. G-PON also supports a logical reach up to 60 km. Hence, it includes the capability to support future extended-reach systems using, for example, optical amplifiers. Up to 1:64 split ratios are envisioned (e. g. 1:16, 1:32 or 1:64), but anticipation of a split ratio of 1:128 is taken into consideration in the MAC protocol. Consequently, as soon as the optical components permit such a high split ratio, currently defined G-PON products could be adapted to support this evolution in the number of ONTs/ONUs managed by a single OLT.

The use of PON technology brings several advantages [MAESTDT, CAUGPON]. Since the system is passive (passive splitters), there is no requirement to install and maintain active components in the access distribution network with consequent increased reliability and lower cost. It is possible to share a single optical interface at the OLT among several customers. This helps minimizng space requirements in the central office. In addition, the cost of deploying an optical access infrastructure is lower, compared for example with cost saving relative to the multiple optical interfaces required for point-to-point access solutions.

4. Switching in Networks

4.1. Optical Burst Switching

4.1.1. Introduction

The explosive growth of Internet traffic in the last decade has resulted in the deployment of DWDM (Dense Wavelength Division Multiplexing) in the backbone networks. With increase in high bandwidth applications DWDM, which offers multigigabit rates per wavelength, is going to become the core technology for the next-generation internet.

There are two basic drivers for optical internetworks. One is the explosion of the multimedia (mainly data) traffic over the Internet, especially the World Wide Web, which as many discovered recently, can be bursty at all time scales and various multiplexing levels. The other is the continuing advances of WDM optical networking technologies, which offer many opportunities to streamline both software (protocols) and hardware (electronic equipment) for reduced latency and cost. This section contains the description of a novel switching paradigm for WDM (Wavelength Division Multiplexed) optical networks, called optical burst switching (OBS). Due to the complexity of the topic only the main assumptions will be presented.

The main motivation for considering optical burst switching (OBS) is that some traffic in broadband multimedia services is inherently bursty. More specifically, in addition to traffic in a local Ethernet and between remote Ethernets (i.e. WAN traffic), traffic generated by web browsers, wide-area TCP connections (including FTP and TELNET traffic carried over TCP connections), and variable-bit-rate (VBR) video sources are all self-similar (or bursty at all time scales). More importantly, some studies have concluded that, contrary to the common assumption based on Poisson traffic, multiplexing a large number of self-similar traffic streams results in bursty traffic.

Some of the existing switching paradigms in optical networks are not suitable for supporting bursty traffic. Specifically, using optical circuit-switching via wavelength routing, a lightpath needs to be established first from a source node to a destination node using a dedicated wavelength on each link along a physical path. Therefore, the bandwidth would not be efficiently utilized unless the subsequent data transmissions are long enough in relation to the set-up time of the lightpath. In addition, given that number of wavelengths available is limited, not every node can have a dedicated lightpath to every other node, and accordingly, some data may take a longer route and/or go through O/E and E/O conversions. Furthermore, the extremely high degree of transparency of the lightpaths limits the network management capabilities (e. g. monitoring and fast fault recovery).

An alternative to optical circuit switching is optical or photonic packet/cell switching in which a packet is sent along with its header. While the header is being processed by an intermediate node, either all-optically or electronically (after an O/E conversion), the packet is buffered at the node in the optical domain. However, high-speed optical logic, optical memory technology, and synchronization requirements are major problems with this approach. In particular, the limited buffering time that can be provided to optical signals prevents worm-hole routing and virtual cut-through routing, which are popular in systems with electronic buffers, from being deployed effectively in optical networks.

In order to provide high-bandwidth transport services at the optical layer for bursty traffic in a flexible, efficient, as well as feasible way, a switching paradigm is needed that can leverage the attractive properties of optical communications, and at the same time, take into account its limitations. Optical burst switching is intended to accomplish exactly that.

4.1.2. OBS Overview

There are several major OBS variants. They differ in a number of ways: (i) how they reserve resources (e. g. 'tell-and-wait', 'tell-and-go'), (ii) how they schedule and release resources (e. g. 'just-in-time', 'just-enough-time'), (iii) hardware requirements (e. g. novel switch architectures optimized for OBS, commercial optical switches augmented with OBS network controllers), (iv) whether bursts are buffered (using optical delay lines or other technologies), (v) signaling architecture (in-band, out-of-band), (vi) performance, (vii) complexity, and (viii) cost (capital, operational, $/Gbit, etc.).

In OBS, a first control packet is sent to set up a connection (by reserving an appropriate amount of bandwidth and configuring the switches along a path), followed by a burst of data without waiting for an acknowledgement for the connection establishment. In other words, OBS uses one-way reservation protocols similar to tell-and-go (TAG), also known in ATM as fast reservation protocol (FRP) or ATM Block Transfer with Immediate Transmissions (or ABT-IT). This distinguishes OBS from circuit-switching as well as from other burst-switching approaches using protocols such as Reservation/scheduling with Just-In-Time switching (RIT) and Tell-And-Wait (TAW), also known in ATM as ABT-DT (Delayed Transmissions), all of which are two-way reservation protocols. OBS also differs from optical or photonic packet/cell switching mainly in that the former can switch a burst whose length can range from one to several packets to a (short) session using one control packet, thus resulting in a lower control overhead per data unit. In addition, OBS uses out-of-band signaling, but more importantly, the control packet and the data burst are more loosely coupled (in time) than in packet/cell switching. In fact, they may be separated at the source as well as subsequent intermediate nodes by an offset time as in the Just-Enough-Time (JET) protocol to be described later. By choosing the offset time at the source to be larger than the total processing time of the control packet along the path, one can eliminate the need for a data burst to be buffered at any subsequent intermediate node just to wait for the control packet to get processed.

Alternatively, an OBS protocol may choose not to use any offset time at the source, but instead, require that the data burst go through, at each intermediate node, a fixed delay that is no shorter than the maximal time needed to process a control packet at the intermediate node. Such OBS protocols will be collectively referred to as TAG-based, since their basic concepts are the same as that of TAG itself.

One way to support IP over WDM using OBS is to run IP software, along with other control software as a part of the interface between the network layer and the WDM layer, on top of every optical (WDM) switch. In the WDM layer, a dedicated control wavelength is used to provide the "static/physical" links between these IP entities. Specifically, it is used to support packet switching between (physically) adjacent IP entities which maintain topology and routing tables. To send data, a control packet is routed from a source to its destination based on the IP addresses it carries (or just a label if MPLS is supported) to set up a connection by configuring all optical switches along the path. Then, a burst (e. g. one or more data IP packets, or an entire message) is delivered without going through intermediate IP entities, thus reducing its latency as well as the processing load at the IP layer. Note that, due to the limited "opaqueness" of the control packet, OBS can achieve a high degree of adaptivity to congestions or faults (e. g. by using detection-routing), and support priority-based routing as in optical cell/packet switching, as to be discussed later.

In OBS, the wavelength on a link used by the burst will be released as soon as the burst passes through the link, either automatically according to the reservation made (as in JET) or by an explicit release packet. In this way, bursts from different sources to different destinations can effectively utilize the bandwidth of the same wavelength on a link in a time-shared, statistical multiplexed fashion. Note that, in case the control packet fails to reserve the bandwidth at an intermediate node, the burst (which is considered blocked at this time) may have to be dropped.

OBS can support either reliable or unreliable burst transmissions at the optical layer. In the former, a negative acknowledgement is sent back to the source node, which retransmits the control packet and the burst later. Such a retransmission may be necessary when OBS is to support some application protocols directly, but not when using an upper layer protocol such as TCP which eventually retransmits lost data. In either case, a dropped burst wastes the bandwidth on the partially established path. However, since such bandwidth has been reserved exclusively for the burst, it would be wasted even if one does not send out the burst (as in two-way reservation). Similar arguments apply to optical or photonic packet switching as well. In order to eliminate the possibility of such bandwidth waste, a blocked burst (or an optical packet) will have to be stored in an electronic buffer after going through O/E conversions, and later (after going through E/O conversions), relayed to its destination. Fiber-optical delay lines (FDLs) providing limited delays at intermediate nodes, which are not mandatory in OBS when using the JET protocol, would help reduce the bandwidth waste and improve performance in OBS, as to be discussed next. Note that, when using TAG-based OBS protocols (or optical/photonic packet switching), FDLs (or optical buffers) are required to delay each optical burst when the control packet (or the packet header) is processed, but do not help improve performance. Summarizing the above discussions, switching optical bursts achieves, to a certain extent, a balance between switching coarse-grained optical circuits and switching fine-grained optical packets/cells, and combines the best of both paradigms, as illustrated in table 3.

Optical Switching (paradigm)	Bandwidth Utilization	Latency (set-up)	Optical Buffer	Proc./Sync. Overhead (per unit data)	Adaptivity (traffic & fault)
Circuit	low	high	not required	low	low
Packet/Cell	high	low	required	high	high
Burst	high	low	not required	low	high

Table 3: Comparison between three optical switching paradigms

4.1.3. Polymorphic Control

In order to put optical burst switching (OBS) in perspective, a framework called polymorphic control, of which OBS is an integral part, needs to be described.

The framework of polymorphic control is a product of integrating many individual research ideas and results on optical network architectures, control and management. As mentioned earlier, under this framework, an optical layer is "sliced" into static and dynamic virtual optical networks (VONs), which apply self-reconfiguration and on-demand reconfiguration, respectively.

One of the basic forms of self-reconfiguration is scheduled communications. When the bandwidth (e.g. in terms of the number of wavelengths) in a static VON is limited, the set of communicating node pairs may be partitioned into a number of subsets such that the node pairs in each subset can communicate at the same time. Each subset is to be allocated a super time-slot during which data can be transmitted or received between the communicating nodes in that subset, and the number of time slots determines the schedule length. A schedule specifying, among other things, the time slot during which a given node pair can communicate, and the path and wavelength it will use, is then determined. Based on such a schedule, the VON can go through a pre-determined sequence of configurations by appropriately changing the switch settings in between two super time-slots. In this way, external electronic control and its associated implementation overhead and performance degradation are minimized. In scheduled

communications, two important performance measures are the schedule length and the bandwidth (i. e. wavelength) requirement, which relate to each other. With sufficient bandwidth, scheduled communications become embedded communications as a special instance, where the schedule length is one, or in other words, communications among the entire set of communicating nodes are accommodated at the same time. In a similar approach, which may also be considered as a form of self-reconfiguration, a logical topology (analogous to a static VON) containing the set of communicating nodes is devised and embedded even when bandwidth is limited, such that these nodes may communicate at the same time, but a message from its source to its ultimate destination may go through more than one lightpaths, thus requiring O/E and E/O conversions at the nodes where two lightpaths meet.

In VONs adopting on-demand reconfiguration, where the performance measures include throughput, utilization, delay and blocking probability, dynamically changing traffic patterns are supported by transferring data in two basic fashions, namely circuit-switching and packet-switching. With circuit-switching, connections (or lightpaths) between source and destination pairs are established before data is transferred, and released after the transfer is completed. Both centralized control and distributed control have been studied, and in either case, it is common to use out-of-band signaling (i. e. a separate control network with a dedicated wavelength).

With packet-switching, each intermediate node stores an incoming packet, and then forwards it to the next node based on its header and a locally stored routing table. Distributed control is natural and in-band signaling is more often used than out-of-band signaling. Note that alternately, a flow of packets can be switched based on the match between a label carried by each packet's header and a label stored at each node, which is set up either by previous packets of the same flow (as in IP-switching) or by the network (as in tag switching). A bursty VON is a dynamic VON that adopts a novel paradigm (OBS), which can be used to support MPLS (Multi-Protocol Label Switching) in an IP over WDM environment.

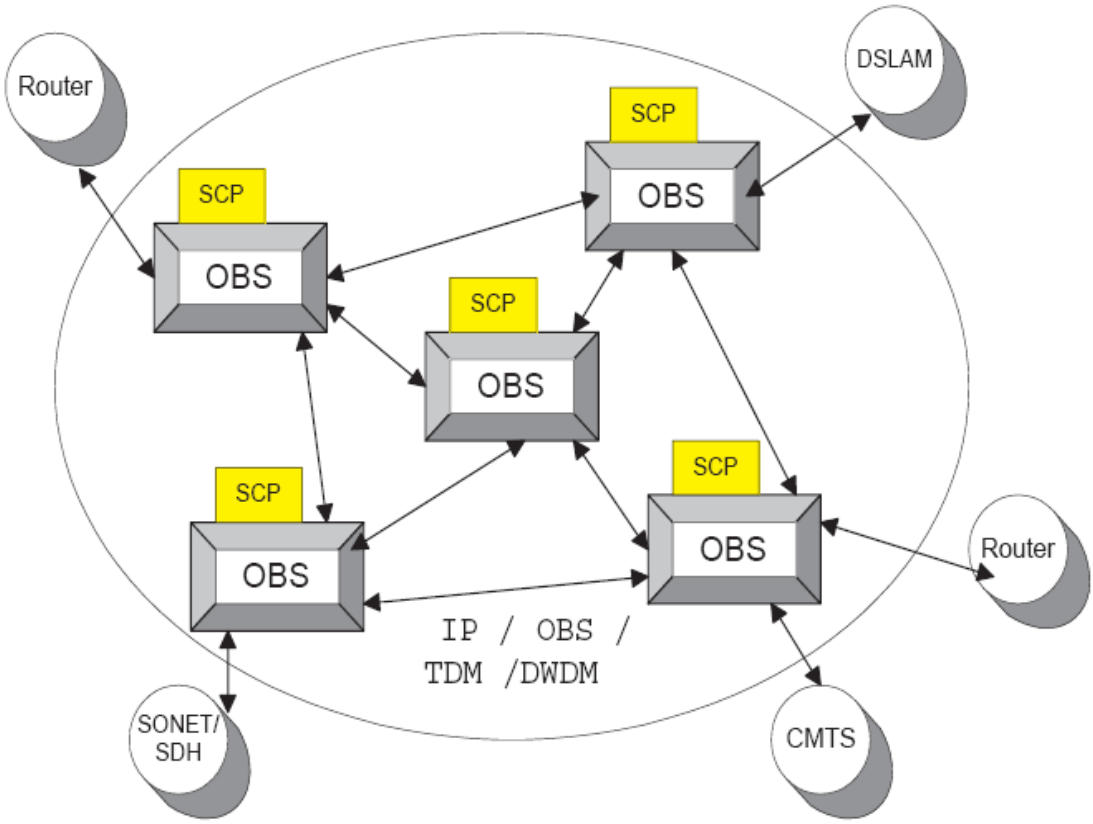


Figure 24: Basic architecture of an OBS network

4.1.4. Just-Enough-Time Protocol

The proposed Just-Enough-Time (JET) protocol for OBS has two unique features, namely, the use of delayed reservation (DR) and the capability of integrating DR with the use of FDL-based buffered burst multiplexers (BBMs). These features make JET and JET-based variations especially suitable for OBS when compared to TAG-based OBS protocols and other one-way reservation based OBS protocols that lack either or both features.

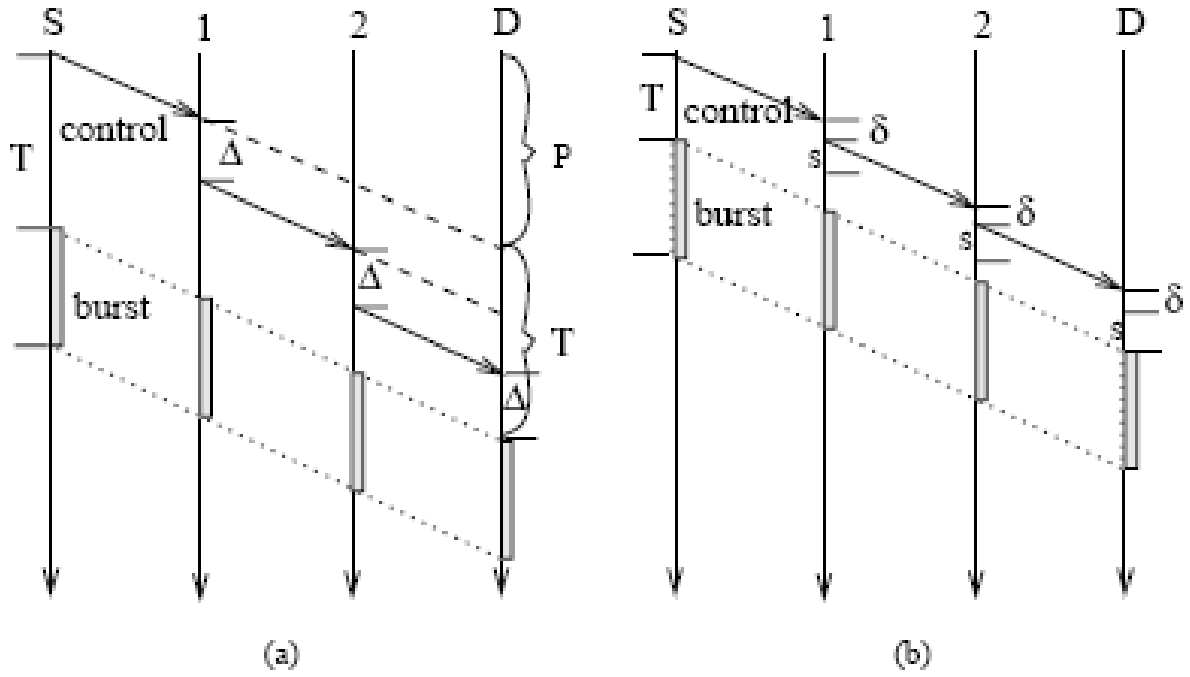


Figure 25: OBS using the JET protocol

Figure 25 illustrates the basic concept of JET. As shown, a source node having a burst to transmit first sends a control packet on a signaling channel (which is a dedicated wavelength) towards the destination node. The control packet is processed at each subsequent node in order to establish an all-optical data path for the following burst. More specifically, based on the information carried in the control packet, each node chooses an appropriate wavelength on the outgoing link, reserves the bandwidth on it, and sets up the optical switch. Meanwhile, the burst waits at the source in the electronic domain. After an offset time T, the burst is sent in optical signals on the chosen wavelength (for example at 2.5 Gb/s).

It is important to note that the burst can be sent without having to wait for an acknowledgement from its destination. At 2.5 Gb/s, a burst of 500 Kbytes (or 4,000 average-sized IP packets) can be transmitted in about 1.6 ms. However, an acknowledgement would take 2.5 ms just to propagate over a distance of merely 500 km. This explains why one-way reservation protocols are generally better than their two-way counterparts for bursty traffic over a relatively long distance. Once a burst is sent, it passes through the intermediate nodes without going through any buffer, so the minimal latency it encounters would be the same as if the burst is sent along with the control packet as in optical packet switching. Of course, if a burst is extremely small, one may just as well send the data along with the control information using packet-switching.

4.1.5. Just-In-Time Signaling Protocol

Just-in-time (JIT) is a reservation protocol for OBS networks that features out of band signaling, eliminating the buffering of data bursts in intermediate switches. Signaling messages are sent ahead of the data burst to setup the intermediate switches; only the signaling messages undergo OEO conversion at every hop. The cross-connects inside the optical switches are configured before the arrival of the data burst, minimizing the waiting time before the data burst is transmitted. The network infrastructure is independent of the data format and, therefore, data bursts travel transparently through the configured path. There are several ways to make

reservations of data channel bandwidth. Baldine et al. [OBS4] outlined the general aspects of the JumpStart signaling architecture, the message format, the message flows, and discussed several JIT signaling schemes. The implementation presented supports the explicit setup and explicit release JIT signaling scheme shown in figure 26. Five message types are supported: SETUP, SETUP_ACK, CONNECT, RELEASE and KEEPALIVE. A SETUP message is sent ahead of the data burst by the calling host, the message is processed at every hop and the cross-connects are configured along the path to the called host. The calling switch uses a delay estimation mechanism to determine an appropriate delay value for the incoming burst and sends it back to the calling host in the SETUP-ACK message. A CONNECT message confirms the establishment of the path, however the calling host does not wait for the CONNECT message and it starts sending the data burst after an estimated time. The path can be torn down explicitly by a RELEASE message or implicitly through timeouts; therefore, if the burst is long the calling host maintains the path by sending KEEPALIVE messages. Any switch along the path might fail to establish the connection, due to limited wavelengths' capacity or because of transmission (CRC) errors. Because FAILURE messages are not supported in hardware, the intermediate switches are left to timeout and it is left for the higher level protocol to notify the calling host.

In all the Just-In-Time signaling approaches to burst-switching is the lack of the roundtrip waiting time before the information is transmitted (TAG scheme) when the cross-connects inside the optical switches are configured for the incoming burst as soon as the first signaling message announcing the burst is received. The variations on the signaling schemes mainly have to do with how soon before the burst arrival and how soon after its departure, the switching elements are made available to route other bursts.

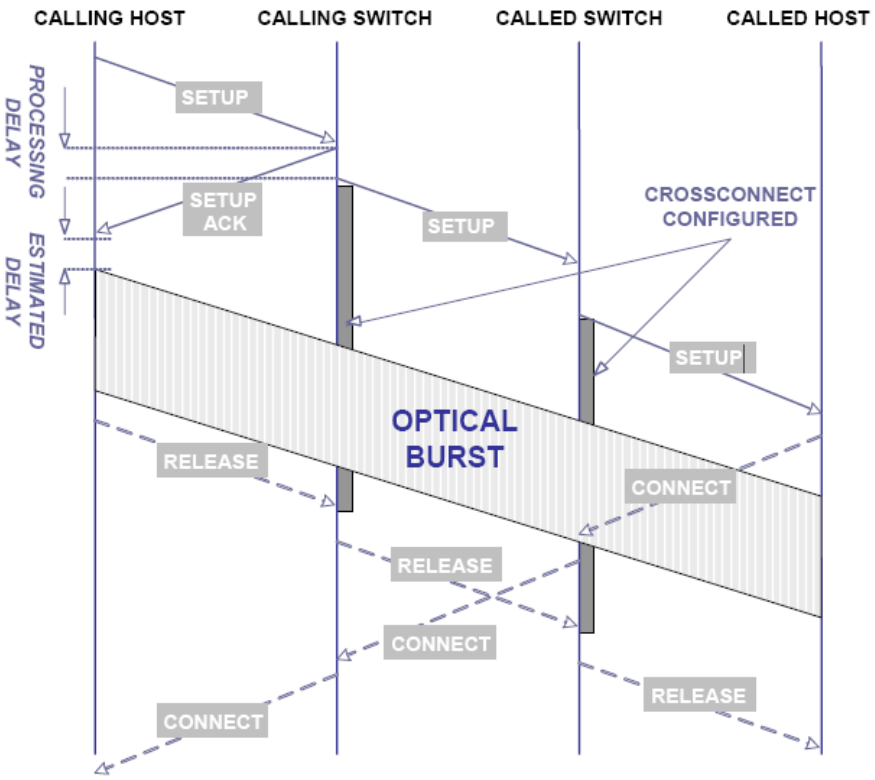


Figure 26: Explicit setup and explicit release JIT signalling

4.1.6. Time Sliced Optical Burst Switching

Time Sliced Optical Burst Switching (TSOBS) is a proposed variant of optical burst switching that replaces switching in the wavelength domain with switching in the time domain. While time-domain switching does require the use of optical buffers, the amount of storage needed is less than 1% of that needed for conventional packet switching, greatly hanging the cost tradeoffs. Like burst switching, TSOBS separates burst control information from burst data. Specifically, Burst Header Cells (BHC) are transmitted on separate control wavelengths on each WDM link. These wavelengths are converted to electronic form at each switch, while all remaining wavelengths are switched through in optical form. The data wavelengths carry information in a Time-Division Multiplexed (TDM) format, consisting of a repeating frame structure, which is sub-divided into time slots of constant length. A repeating sequence of time slots in successive frames, at a fixed position within the frame is referred to here as a channel. Each BHC "announces" the imminent arrival of a data burst, and includes address information plus the wavelength and channel on which the burst is arriving. It also includes an offset, which identifies the frame in which the first timeslot containing data from the burst appears, and a length, which identifies the number of timeslots used to transmit the burst.

4.1.7. QoS Support

One of the primary objectives in the design of an OBS network is to provide QoS. It can be done by minimizing packet loss. Packet loss occurs primarily due to the contention of bursts in the bufferless core. Approaches for resolving contention include wavelength conversion, optical buffering and deflection routing. In wavelength conversion, if multiple bursts try to use the same wavelength on the same output port at the same time, then the bursts are shifted to another free wavelength on the same link. In buffering, fibre delay lines are used to provide the required delay to resolve the contention. In deflection routing, the burst is deflected to an alternate port in case of a contention on the primary port. Deflection in the network results in several side effects, including looping of bursts, and out-of-order packet arrival at the destination.

The above contention resolution techniques are reactive techniques that attempt to resolve contentions rather than avoiding the contentions. Also, these contention resolution techniques attempt to minimize the loss based on the local information at that node. An alternative to resolving contention when it occurs, is to prevent contention before it happens. In contention avoidance, the goal is to reduce the number of contentions, by policing the traffic at the source, or by routing traffic in a way that the congestion in the network is minimized.

4.1.8. Contention-Based Limited Deflection Routing

Some other techniques and protocols have also been proposed in order to ensure the proper routing. The reason is that the above mentioned ones do not address the issue of how routing to an alternate path should be done, given that some constraints may apply to the selection of an alternate path. One of them is a novel Contention based Limited Deflection Routing (CLDR) protocol, which mitigates and resolves contention with significantly better performance as compared to techniques currently known in the literature. While several variants of the basic deflection routing scheme have been proposed, they all lack the ability to determine the alternate route based on clear performance objectives. An on-demand deflection routing scheme sequentially performs the following: Based on certain performance criteria, (i) it dynamically determines if the burst should be deflection routed or retransmitted from the source, (ii) if the decision is to deflection route, then the same is done using a path that is based on the minimization of a performance measure that combines distance and blocking due to contention.

The proposed CLDR scheme prevents injudicious deflection routing. The simulations results show that the scheme proposed has much superior performance both in terms of burst loss probability and increased network throughput. It has been proposed that the network nodes

should periodically re-compute and store optical paths, with the aim of staying optimal in the face of changing node and link congestion measures. This allows for deflection routed bursts to traverse the alternate optical paths that are not necessarily shortest paths but are optimized for best performance (i. e. blocking and delay). This technique calls for monitoring the link and node congestion and updating the same in a periodic manner so that the path computation can be as optimal as possible (albeit with some minor lag in the updates).

Typically, the traffic originating from the edge nodes of the network would be correlated and such correlations would have a significant impact on the burst contentions at the edge as well on internal links in the network.

A model accounts for these correlations (including various parameters that help quantify the correlations) in the prediction of burst loss ratio or probability. The analytical model results are compared with simulation results. Additionally, the analytical modeling results are also used to create some relevant input in the design of the simulation experiments for studying CLDR and comparing it with other known schemes.

4.2. GMPLS

Generalized Multi Protocol Label Switching (GMPLS) is an evolution of the classical MPLS architecture. It was introduced to establish a complete separation of the control and data planes in a network. In fact, today carriers face the problem of bandwidth provisioning and QoS requirements in an even more complex scenario. This reflects in the need of adopting complex signaling and routing protocols. In this perspective, MPLS (spawned as an industrial standard, rather than a product of academic research) allows a big advancement in such direction. However, MPLS did not separate in a bold manner the different planes available in a network. Moreover, the core at the basis of the data transport remained unavailable. It must be emphasized that GMPLS has not been introduced to replace MPLS.

Roughly, it is possible to find out the GMPLS genesis by investigating its parent, the MPLS protocol. MPLS provides the ability of managing label switching and offers support to exploit traffic engineering. The latter is often developed by using other protocols in conjunction with MPLS. For instance, the Resource reSerVation Protocol for Traffic Engineering (RSVP – TE) is often used on the overlay available when employing the MPLS protocol. GMPLS extends the MPLS concept of label switched paths and its traffic engineering capabilities to the control of TDM, lambda, and fiber switched networks. Extensions are directly rooted in the Traffic Engineering (TE) extensions of the standard MPLS (known as MPLS–TE).

GMPLS aims at providing a single, unified control plane architecture for multiple switching layers, by adapting existing MPLS signaling and IP routing protocols for transport networks. GMPLS is based on the IP routing and addressing models. Hence, IPv4 and/or IPv6 addresses are used to identify interfaces, as well as network elements. Relying on the IP layer allows the reuse of the IP routing protocols. The common control plane promises to simplify network operation and management by automating end-to-end provisioning of connections, managing network resources, and providing the level of QoS that is expected in applications.

It is possible for service providers to manage a wide range of network elements, ranging from packet switches to digital and optical cross-connects, in a uniform way. GMPLS restores the coherency and transparency needed to maintain complex networks with a reasonable effort. By using GMPLS it is possible to automate almost all operations devoted to circuit planning and resource provisioning, thus reducing operational complexity, time, and costs, and potentially transforming the entire provisioning process, or practical subsets of it, into a real-time process. Owing to its ability of hiding heterogeneities present in the carrier, GMPLS treats and manages packets, light-paths and circuits in the same way. Then, it is possible also to establish an inter-

network of different technologies, allowing also legacy networks to benefit of end-to-end provisioning and traffic engineering.

The biggest addition in the GMPLS protocol suite is a new signaling protocol, Link Management Protocol (LMP), to establish, release and manage connections between two adjacent GMPLS capable nodes. Other protocols are employed: RSVP–TE, Open Shortest Path First – Traffic Engineering (OSPF–TE), Constraint Routed – Label Distribution Protocol (CR–LDP) and Intermediate System to Intermediate System – Traffic Engineering (IS–IS–TE), where OSPF–TE and IS–IS–TE are extended from the original versions to conform with the GMPLS framework. The next table presents a summary of the protocols employed in the GMPLS framework.

Functionality	Protocols	Description
Routing	OSPF-TE IS-IS-TE	Routing protocols for the auto-discovery of network topology, advertise resource availability
Signaling	RSVP-TE CR-LDP	Signaling protocols for the establishment of traffic-engineered LSPs.
Link-Management	LMP	**Control-Channel Management:** established by negotiating link parameters (such as how often a "keep alive" message is sent), also ensures the health of a link using "hello protocol" **Link-Connectivity Verification:** ensures the physical connectivity between switches using a "ping" test message **Link-Property Correlation:** identifies link properties of neighbor switches **Fault Isolation:** isolates faults in the optical domain

Table 4: Protocols employed in the GMPLS framework

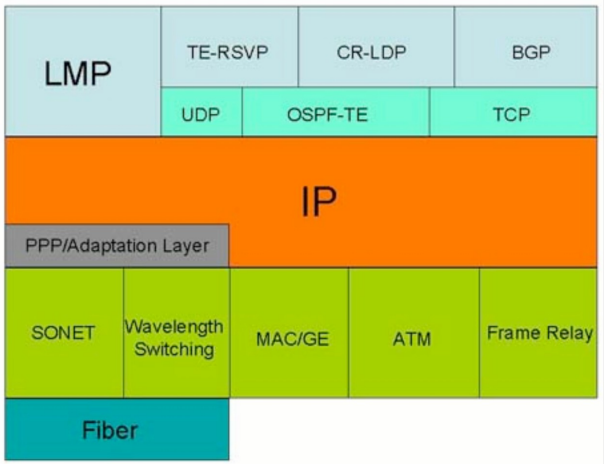


Figure 27: GMPLS protocol architecture

Figure 27 depicts the GMPLS protocol architecture. Note that the IS–IS–TE routing protocol stack is similar to OSPF–TE with the exception that, instead of IP, the connectionless network protocol (CLNP) is used to carry information generated by the IS–IS–TE.

As stated previously, the core technology adopted in the GMPLS control plane is IP-based, but the data plane (sometimes also known as the traffic plane) can now handle directly different technologies. As a consequence, GMPLS extends the MPLS functionalities to handle (in the sense of establishing and provisioning) paths for:

- **Time Division Multiplexing (TDM):** Time slots are treated as labels. Such concept has been specifically tweaked for the Synchronous Optical NETwork (SONET) technology.
- **Frequency Division Multiplexing (FDM):** Every electromagnetic frequency is managed as a label. When employing an optical network, a given light wave (or a group of light waves) is handled with a proper label.
- **Space Division Multiplexing (SDM):** Space division relies on the physical position of the data in a given portion of the space. Thus, a label indicates the physical – spatial position of the related data. This technique is often used for managing data in photonic cross-connects with GMPLS.

The core of the architecture being based on labels, it is important to state how they are distributed and managed. To be able to support devices that switch in different domains, GMPLS introduces new additions to the format of the labels.

The new label format is referred to as a generalized label containing information to allow the receiving device to program its switch and forward data regardless of its construction (packet, TDM, lambda, etc.). A generalized label can represent a single wavelength, a single fiber, or a single time-slot. Traditional MPLS labels (e. g. ATM based) are also included to assure compatibility with legacy devices and domains.

The information that is embedded in a generalized label includes:

- **LSP encoding type:** Indicates what type of label is being carried.
- **Switching type:** Indicates the switching capabilities of a node. For instance, if a node is able to switch packets, time-slots, wavelengths or fibers.
- **A general payload identifier:** Indicates what payload is being carried by the LSP.

Similar to MPLS, label distribution starts from the upstream LSR requesting a label from the downstream LSR. GMPLS takes this further by allowing the upstream LSR to suggest a label for a LSP that can be overridden by the downstream LSR. Besides, establishing an LSP in a GMPLS network is similar to that of MPLS networks. Then, the following operations are performed: i) label creation and distribution; ii) table creation at each GMPLS capable router; iii) label-switched path creation; iv) label insertion/table lookup; v) packet forwarding.

4.2.1. GMPLS Implementations

The actual status of GMPLS-enabled devices is something difficult to retrieve. However, the last exhaustive survey about GMPLS implementations and adoptions from manufacturer is 4 years long. In fact, in December 2002, IETF (and the ISOC) spawned a survey to understand whether or not GMPLS had been introduced in manufacturers' pipelines.

The survey concluded that GMPLS is a reality, even if many vendors had only beta versions of their devices. However, owing to the increased market demand for MPLS/GMPLS devices, it is now possible to buy GMPLS compliant devices at a reasonable cost, or to upgrade recent devices via a software update (of course, not for free).

Table 5 summarizes the outcome of the survey available in [PAPSUR], updated with information gathered on the mostimportant manufacturers' websites.

Agilent	YES	On-Sale (Device)
Alcatel	YES	On-Sale (Device)
Calient	N/A	On-Sale (Code)
Data Connection	YES	On-Sale (Code)
Juniper	YES	On-Sale (Device)
Marconi	YES	On-Sale (Device)
Movaz	YES	On-Sale (Device)
NEC	YES	On-Sale (Device)
NetPlane	YES	On-Sale (Code)
Wipro	YES	On-Sale (Code)

Table 5: Survey on the availability of GMPLS implementations

4.2.2. Adoption of GMPLS in Other Projects Interesting for RINGrid

Two different projects rely on GMPLS for supporting the next-generation of grids. In order to provide a comprehensive overview, the following sections will briefly discuss how GMPLS has been exploited, while other sections will deal in detail about the functionalities of the experimental deployments, as well as with prototypes.

4.2.3. CHEETAH: Circuit-switched High-speed End-to-End Transport ArcHitecture

Project CHEETAH's [CHEWEB] gigabit Ethernet switches are used to allow multiple users to share the network resources. In order to enable dynamic provisioning of dedicated end-to-end circuits GMPLS-enabled switches have been developed. CHEETAH's network testbed relies on core nodes of the network, e. g. Sycamore 16K SONET cross-connects that are located at MCNC, Raleigh NC, ORNL in Tennessee, and SOX PoP in Atlanta. The enterprise Ethernet switches act as concentrators of traffic onto the CHEETAH backbone links. Equipping these

Ethernet switches with GMPLS engines would then allow them to participate in setting up and releasing end-to-end connections all the way to the end machines.

One of the most interesting aspects of the project concerns the design and specifications of a GMPLS engine implemented to support the Ethernet switches. Such engine has been called Cheetah Virtual Label Switching Router (CVLSR), which is based upon the VLSR code developed by the DRAGONteam.

Lastly, the CHEETAH technology has been also deployed to provide a control plane in grid applications acting on high-performance networks.

4.2.4. The Dynamic Resource Allocation over GMPLS Optical Networks (DRAGON) Project

The DRAGON Project [DRAWEB] conducts research and development to enable dynamic provisioning of network resources on an inter-domain basis across heterogeneous network technologies. A DRAGON network architecture and control plane is defined, which aims at leveraging the maturity of network technologies (such as WDM, Ethernet, and next generation SONET) to demonstrate the power and flexibility of a "hybrid" packet and circuit switched network infrastructure. A key element is the extension of the GMPLS IP control plane to enable multi-domain, multi-layer, multi-service provisioning with robust levels of Authentication, Authorization and Accounting (AAA).

The most interesting feature is that the DRAGON project released open-source GMPLS software. The latter is a key component of the IP control plane forecast in DRAGON, and it allows provisioning across domain boundaries and multiple networks.

4.3. Dynamic Resource Allocation via GMPLS Optical Networks

4.3.1. Introduction

Dynamic allocation of resources is an important topic not just in allocating grid computing resources, but also in allocating networking resources that are necessary to interconnect nodes/grids together. Depending on the type of grid application, requests for resources (both grid computing and networking) could be immediate or scheduled. Further features includes duration of connections (short holding times — seconds to a few minutes —, long holding times — tens of minutes to hours) and bandwidth granularity (fine bandwidth pipes — tens to a few hundred Mbps —, large bandwidth pipes — 1–10 Gbps). The dynamic demands of grid applications need that the underlying networks be capable of setting up and releasing end-to-end connections that may cross traditional administrative networks' demarcations. Moreover, it is relevant to devise resources scheduling mechanisms that can cost-effectively allocate networking resources to applications, while maintaining efficient allocation of these resources.

Generally speaking, dynamic provisioning of end-to-end deterministic bandwidth pipes is an essential capability to enable grid computing applications.

4.3.2. Networking Requirements for Grid Applications

Emerging eScience applications have some networking requirements that can be summarized as follows:

- **Cost effectiveness.** Offering users quality guarantees at affordable prices implies that network providers maximize the productivity of the network assets without sacrificing quality. This requires new modes of operations, in which deterministic bandwidth paths are dynamically provisioned. Large capacity pipes have traditionally been statically provisioned, which has been a very costly solution. The main idea is that demands for these long pipes are usually time dependent; hence, statistical sharing of these pipes

among many users will significantly decrease the cost. Of course, there will be a chance of blocking. Therefore, the network resources (interfaces, ports, wavelengths and time slots) can be appropriately sized in order to maintain an acceptable rate of blocking. Connection-oriented technologies, such as MPLS and GMPLS control planes, are the right enabling tools to provide users with on-demand provisioning capabilities.

- **High degree of flexibility for users communications needs.** Applications require variable amounts of bandwidth pipes and for varying durations. Near-real-time signalling is needed for "application-driven" rapid provisioning of communications paths. The GMPLS distributed control plane is a good fit for this requirement.
- **Interdomain and internetworking dynamic provisioning.** Provisioning of heterogeneous paths across multiple networks is an important requirement, since supercomputers, instrumentation tools and other storage facilities form grids that cross network demarcations. Devices that are not equipped with GMPLS control capabilities (e.g. Ethernet switches) should be equipped with such capability.
- **Short-term and long-term as well as low-capacity and large-capacity connections should be scheduled.** This requirement will ensure that the networking resources are efficiently utilized in response to the applications' demands. Dynamic reconfiguration of network resources should also be scheduled.
- **Distributed fault management mechanisms.** Dynamic provisioning of large bandwidth pipes requires mechanisms for a rapid restoration even across multiple networks.
- **End-to-end QoS.** Currently it is hard to achieve that QoS is supported end-to-end. However, edge-to-edge QoS is provided instead in many cases.

The GMPLS distributed control plane has three main capabilities that allow it to meet most of the above requirements:

1. distributed routing
2. near-real-time connection set-up using signalling
3. network services and topology auto-discovery mechanisms

Many areas, such as scheduling connections for future times, interdomain and internetworking rapid provisioning are currently being investigated in the research community.

It is worthwhile noting that the above requirements draw a new picture of networking that is quite different from what happens today. At present, users interact with networks according to a client-server model:

1. Users provide carriers with requirements for static networking demands.
2. The carrier, by using a CNM (Centralized Network Management) system provides static paths to meet the customers' requests.

The centralized approach is not scalable for BoD (Bandwidth-on-Demand) services, whereas it is suitable to book-ahead large bandwidth pipes, where significant resources are to be committed.

A hybrid approach is likely to be the best solution: Functionalities such as long-term capacity planning, analysis for fault management, AAA (Authentication, Authorization, Accounting) are better suited for a CNM system. At the same time, the CNM may benefit from some of the capabilities of the control plane, such as auto-discovery, because this capability will significantly improve the accuracy of several important databases in the CNM system (e. g. network inventory assets, service management).

4.3.3. DRAGON

The DRAGON (Dynamic Resource Allocation in GMPLS Optical Networks) project is a research and experimental framework for high-performance networks required by grid computing and e-science applications (http://dragon.maxgigapop.net).

The DRAGON project is a collaboration amongst the following institutions:

1. Mid-Atlantic Crossroads (MAX)
2. University of Southern California (USC) Information Sciences Institute (ISI) East
3. George Mason University (GMU)

DRAGON is developing technology and deploying network infrastructure that allows dynamic provisioning of network resources on an inter-domain basis across heterogeneous network technologies, in order to establish deterministic paths in direct response to end-users requests. This also includes multi-domain provisioning of traffic-engineering paths, using a distributed control plane across heterogeneous network technologies.

The DRAGON architecture is similar to contemporary research and education (R&E) networks in the context of topology and administrative organization. It consists of a number of autonomous network domains, each able to unilaterally define internal traffic management policy, and enter into bi-lateral peering arrangements with other external domains where desirable and mutually agreed. The fundamental difference is that the DRAGON architecture is based on switching and forwarding nodes that support the GMPLS label hierarchy.

The GMPLS label hierarchy comprises packet labels (e. g. MPLS), TDM labels (e. g. SONET), wavelength labels, and fiber labels. Switching nodes that operate on traffic according to GMPLS labels are referred to as Label Switch Routers (LSR). An LSR is not required to support all GMPLS label types; instead, it will support a subset based on equipment and network capabilities.

The DRAGON architecture further assumes that each LSR runs an intra-domain GMPLS routing protocol (e. g. GMPLS-OSPF) and a GMPLS path signaling protocol (e. g. GMPLS-RSVP). Alternate routing and signaling protocols may also be used. The combination of label switched routers and associated signaling allows for the dynamic control and instantiation of Label Switched Paths (LSPs), which provide end-to-end connections for multi-service traffic at the transport level consistent with the specified label type.

Intra-domain establishment of lambda based LSPs is reasonably mature due to standardized extensions to the Interior Gateway Protocols (IGP) and signaling protocols. However, inter-domain instantiation of end-to-end LSPs, which can be dynamically provisioned across multiple administrative domains and heterogeneous network technologies, is still problematic. As depicted in figure 28, there are several capabilities and technologies missing in order for this capability to be readily available to end systems:

- no standardized inter-domain routing architecture for LSPs
- no simple application interface
- no end to end instantiation (with proper authentication, authorization, and accounting)
- no ability to signal through non-GMPLS enabled network segments.

Within the DRAGON project, an architecture for inter-domain end-to-end GMPLS transport has been proposed.

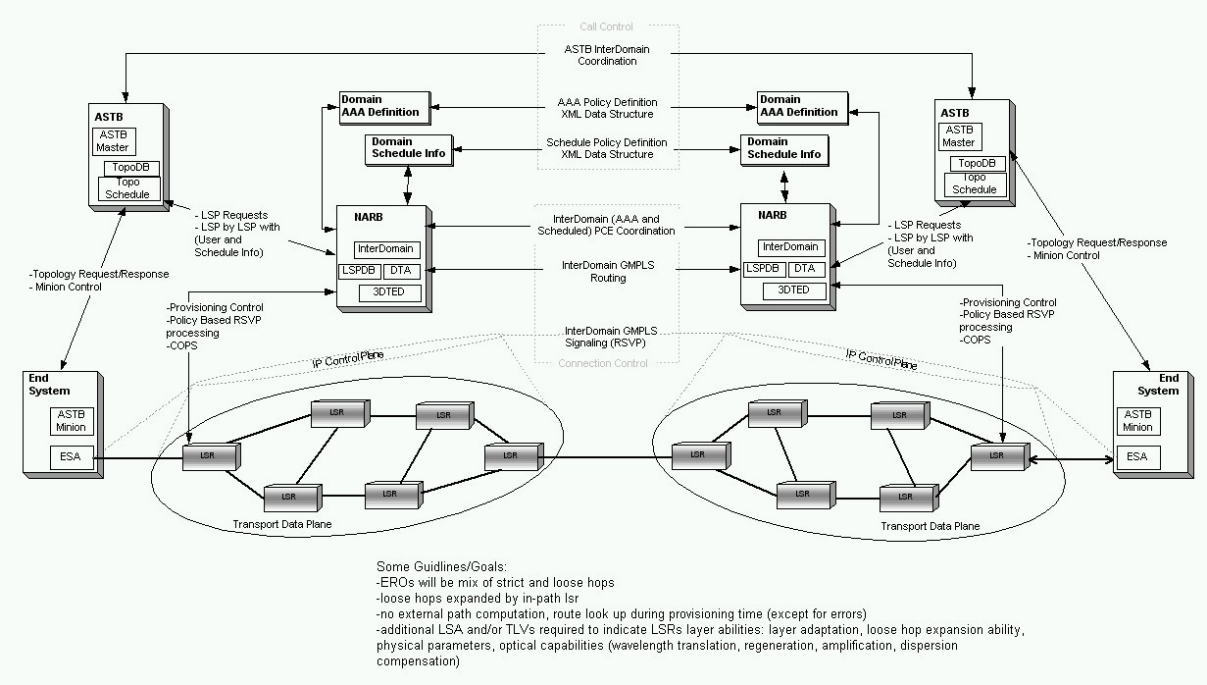


Figure 28: DRAGON control plane architecture

The DRAGON Architecture will provide the missing components in the form of Network Aware Resource Broker (NARB), Virtual LSR (VLSR) and an Application Specific Topology Definition Language (ASTDL).

- **NARB:** In order to establish end-to-end inter-domain LSPs, interconnected networks must exchange information corresponding to the label types each offers, which destinations can be reached via which label types, and which users are allowed to use which label types. DRAGON proposes the concept of a Network Aware Resource Broker (NARB) that represents an autonomous domain and exchanges this information with NARB instantiations representing other domains. It is this inter-domain exchange that enables end-to-end LSP routing. An integral part of this exchange will be authentication, authorization and accounting information.
- **VLSR:** The concept of a Virtual Label Switch Router to translate standard GMPLS protocols into device specific protocols, to allow dynamic reconfiguration of non-MPLS aware devices is proposed. This is a pragmatic issue, in that it allows non-GMPLS switching devices to be managed via SNMP or another proprietary protocol and still be an integral part of the end-to-end GMPLS environment.
- **ASTDL** is used to formalize network service definitions and to simplify description of complex network topologies. DRAGON also proposes to develop tools and middleware libraries to interface user applications with the routing, signaling, and security framework of the proposed architecture.

The DRAGON testbed (figure 29) is situated in the Washington DC metropolitan area, building on previous experience with operation of the MAX GigaPOP and other regional testbed networks.

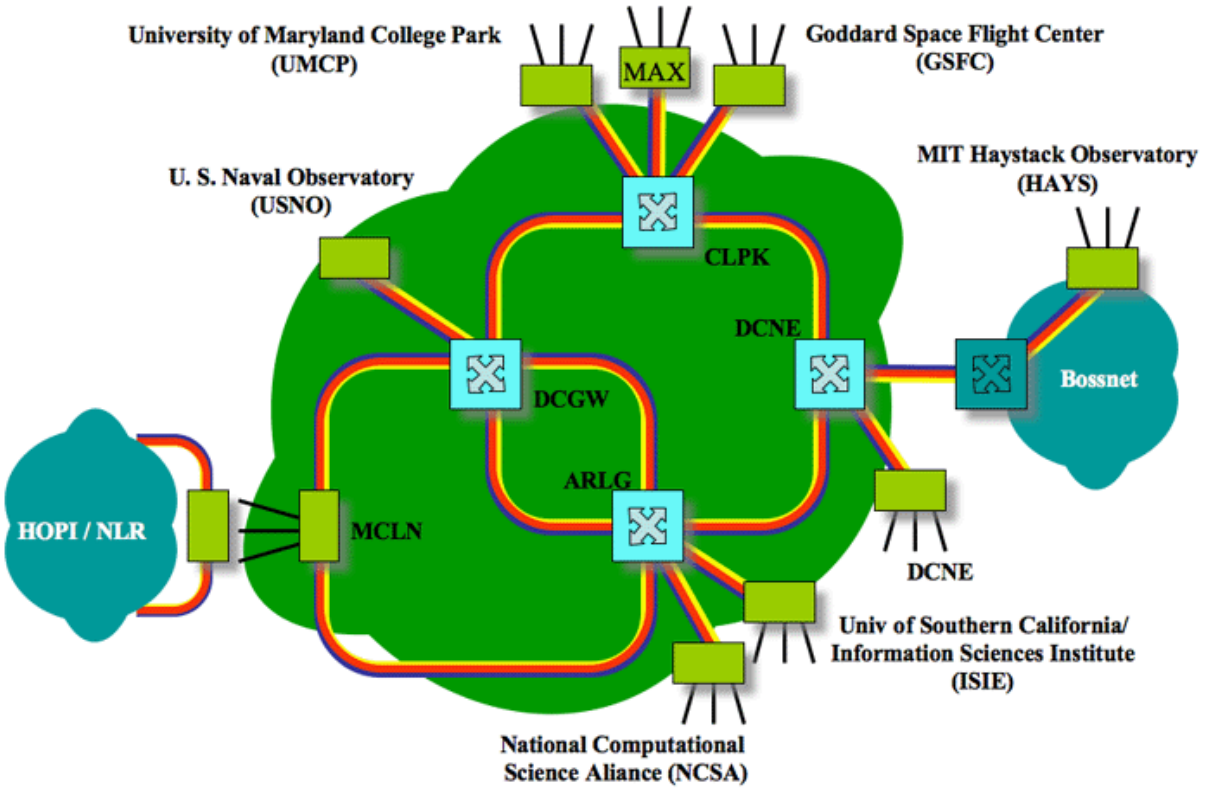


Figure 29: DRAGON optical switched testbed in the Washington DC metro-area

It is built around a multiprotocol all-optical metro-area WDM infrastructure to allow for maximum flexibility in terms of the type of end-systems and applications which can be supported. It consists of two interlocking dark fiber rings interconnecting 5 core PoPs. The total diameter is about 150 km.

4.4. Resilient Packet Ring

The IEEE 802.17 Resilient Packet Ring Working Group (RPRWG) develops standards to support the development and deployment of Resilient Packet Ring networks in local, metropolitan, and wide area networks for resilient and efficient transfer of data packets at rates scalable to many gigabits per second. These standards build upon existing physical layer specifications. IEEE 802.17 is a unit of the IEEE 802 LAN/MAN (Metropolitan Area Network) standards committee.

In metropolitan and wide area networks, fiber optic rings are widely deployed. These rings are currently using protocols that are neither optimized nor scalable to the demands of packet networks, including speed of deployment, bandwidth allocation and throughput, resiliency to faults, and reduced equipment and operational costs. In response to the need for an Ethernet-like MAN solution, the IEEE 802.17 resilient packet ring (RPR) working group (http:/grouper.ieee.org/groups/802/rprsg) is to develop a new standard for MANs using an optical fibre ring.

IEEE 802.17 [802 17] established five criteria that have guided the RPR work:

1. Broad market potential. Fiber-optic rings are already widely deployed, but not optimized for data traffic. With the growing demand for IP-based data traffic in the metropolitan area, there is a potentially large demand for an effective MAN solution.
2. Compatibility. RPR must be compatible with existing IEEE 802 standards, especially those governing QoS and network management.
3. Distinct identity. No current effort within IEEE 802 addresses the requirements of a high-data rate, resilient MAN to provide QoS for a variety of traffic types.
4. Technical feasibility. Implementations of candidate proposals for an RPR already exist, demonstrating the feasibility of this general approach.
5. Economic feasibility. Fiber-optic and related costs are so high that the RPR will provide a cost-effective solution to MAN requirements.

The RPR is intended to serve a variety of needs. Local exchange carriers can use RPR to offer a shared MAN capability available to small and medium-sized businesses in a metropolitan area. Larger organizations, with facilities spread over a campus or metropolitan area, can support their own private MAN using RPR. Because of its high data rate and robustness, RPR will also be attractive within a local area of Internet service providers and Internet server farms.

IEEE 802.17 is backed by the RPR Alliance (http://www.rpralliance.org), consisting of system vendors, silicon vendors, carriers, service providers, and individual networking consultants. RPRA is participating in establishing the objectives and the plans to develop the standard and demonstrate interoperability. Whereas the 802.17 working group is focused on technical issues, RPRA is focused on marketing issues.

RPR implements a media access control (MAC) protocol, for access to the shared ring communication medium, which has a client interface similar to that of Ethernet.

At the MAC level, the primary objective is to provide enhanced services for the transmission of Ethernet packets over a ring-based interconnect topology. This implies a simple mapping from the Ethernet frame format to the RPR frame format.

In addition, the MAC protocol provides a flexible QoS capability to support various classes of real-time and non–real-time traffic. Two major classes are supported:

- synchronous, for low-latency requirements, voice, and video
- asynchronous, for high-bandwidth cost-effective LAN support

The asynchronous class provides latency-insensitive subclasses, such as the following:

- **Unfair.** High-priority traffic, such as prioritized HTTP traffic.
- **Fair.** Lower-priority traffic, such as management messages.
- **Bulk.** Opportunistic traffic, such as file transfers.

Key features of the MAC protocol include the following:

- Each node on the ring has a unique MAC address. As in other IEEE 802 MAC protocols, each MAC frame includes a destination address and a source address.
- For a unicast frame (only one destination), it is the responsibility of the destination to remove the frame. The effect is that multiple frames can be circulating on the ring at any one time, providing efficient spatial reuse of the fiber capacity.
- For a broadcast or multicast frame, each destination copies the frame as it goes by, and the source removes the frame after a complete circulation.

- The frame header includes a class-of-service indicator that can be used to implement priority handling of frames.

Unlike FDDI and the IEEE 802.5 token ring, RPR does not use a token to control access to the medium. With the high data rates and need for flexible QoS, a token-passing medium access scheme would be inadequate. Instead, IEEE 802.17 relies on a capacity-management scheme based on channels. Although the full details of the channel architecture are not yet in place, the following general characteristics apply to RPR:

- A channel is a reserved portion of the capacity of the link from a sender to one or several receivers. Channels are multiplexed on the fiber or frequency band when WDM equipment is used.
- A channel is a dynamic resource that is managed in quantum steps, up to the full capacity of the link. The bandwidth allocated to the channel can be changed dynamically during its lifetime.
- A channel can have multiple receivers, enabling true multicast operations (as well as unicast and broadcast).
- Channels are simplex, making it possible to guarantee resources both upstream and downstream, and thus provide inherent support for asymmetric traffic patterns with high bandwidth utilization.

4.4.1. Ring Network Basics

In RPR, the transit methods supported are cut-through (the station starts to forward the frame before it is completely received) and store-and-forward. To prevent frames, with a destination address recognized by no station on the ring, from circulating forever, a time to live (TTL) field is decremented by all stations on the ring. The receiver of a frame will remove the frame from the ring and release the bandwidth back to the sender. This is generally known as spatial reuse. Figure 30 shows an example scenario where spatial reuse is obtained on the outer ring; station 2 is transmitting to station 4 at the same time as station 6 is transmitting to station 9. Figure 30 (b) shows a station's attachment to one ringlet, showing the "insertion buffer" or "transit queue", which stores frames in transit, while the station itself adds a frame.

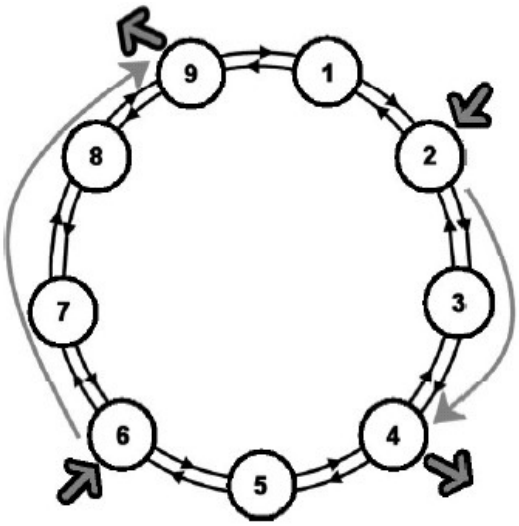


Figure 30 (a): Destination stripping and spatial reuse illustrated on ringlet 0

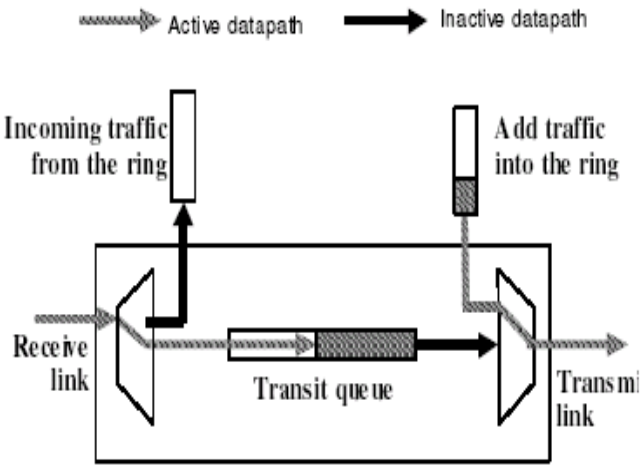


Figure 30 (b): A station's attachment to one ringlet

4.4.2. Station Design and Packet Priority

The RPR ring implements a medium access control (MAC) protocol that controls the stations' access to the ring communication medium. The clients can call using this MAC protocol to send and receive frames and status information.

RPR provides a three level, class based, traffic priority scheme: class A, B and C. The class A has the highest priority to achieve lowest latency and jitter. But RPR will not drop any frames regardless of what class the frames are. So all frames will reach the destination.

Class A traffic is divided into classes A0 and A1, and class B traffic is divided into class B-CIR (Committed Information Rate) and B-EIR (Excess Information Rate). The two traffic classes C and B-EIR are called Fairness Eligible (FE), because such traffic is controlled by the "fairness" algorithm, described in the next section.

The class A0 traffic has a "reserved" bandwidth in the ring network and can only be utilized by the station holding the reservation. Each station will broadcast its reserved bandwidth to the network and calculate the total reserved bandwidth using this kind of messages it received from all other stations. The class A1 and class B-CIR have the "reclaimable" bandwidth that can be also used to send FE traffic.

Each class has a corresponding traffic shaper in each station in the ring to shape the traffic transmitted. There is another shaper in each station for all classes expect the class A0. It is for all unreserved bandwidth to make sure the bandwidth of class A0 will not be used by other classes.

To make sure higher priority traffic being forwarded quicker than lower priority ones, a simpler solution, adopted by RPR, is to optionally have two transit queues. So the class A traffic will always queue in the Primary Transit Queue (PTQ) and class B and class C will wait in the Secondary Transit Queue (STQ). All frames in the PTQ will be transmitted immediately by the station after they arrive. Figure 31 shows the forwarding priority of the different classes of traffic in a station. The numbers in the circles give a very crude indication of transmit link priority.

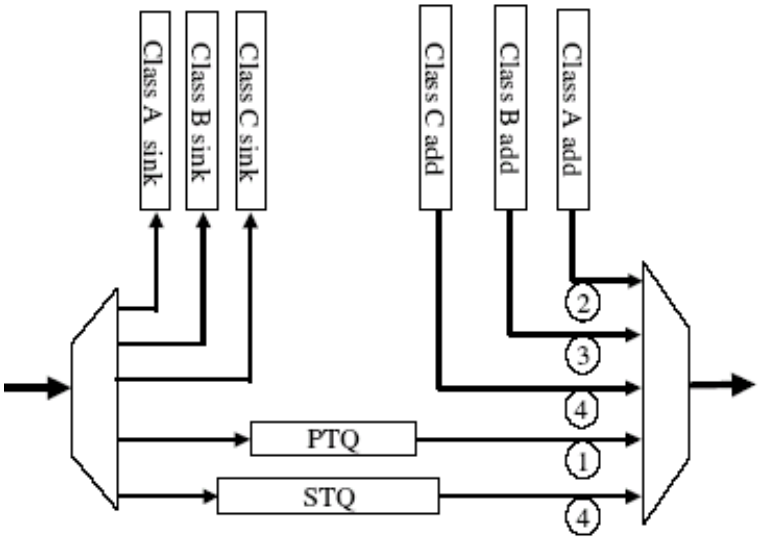


Figure 31: The attachment to one ring by a dual transit queue station

4.4.3. RPR Fairness Algorithm

A fairness algorithm is defined by RPR to distribute unallocated and unused reclaimable bandwidth fairly among the contending stations and use this bandwidth to send class B-EIR and class C traffic, i.e. the fairness eligible (FE) traffic.

The fairness algorithm starts working when there are congestions on the ring. The station that detects congestion should calculate a first approximation to the fair rate, either by dividing the available bandwidth among all upstream stations that are currently sending frames through this station, or by using its own current add rate. The calculated value will be sent upstream to all stations that are contributing to the congestion, and these stations have to adjust their sending of FE-traffic accordingly.

There are two modes for a station to responsd to the congestion it detected. In the "conservative" mode, the station sends a new calculated fair rate value after an estimated time, called the Fairness Round Trip Time. In the "aggressive" mode, a congested station keeps on sending a new fair value to the upstream stations until the congestion disappears. The upstream stations will receive a message of the disappearance of the congestion and gradually increase their transmission rate if the current bandwidth they are using can not fit their demands.

4.4.4. Topology Discovery

All stations in a ring will collect topology information, including connectivity and the ordering of the stations around the ring using a topology protocol. The information collected is stored in a topology database in each station.

When establishing the ring, all stations will send topology discovery messages. These messages will go around the ring so every station in the ring can see them. The senders' status will be carried in the messages and will therefore be learned by all. After a ring is established, if any one wants to join the ring or leave the ring, this will cause a topology change, which will trigger the joining station or the adjacent station of the leaving station to send out a new topology discovery message. This message will then trigger a ripple action of sending topology discovery messages of every station in the ring until the topology information is stable. Stations in the ring

will periodically send topology discovery messages to maintain and update the topology database of all stations.

The topology database is used when a frame is given to the MAC without specifying which direction it will go on the ring. The MAC will forward the frame based on the topology to the shortest path. Information in the topology database is also used when calculating the Fairness Round Trip Time in the conservative mode of the fairness algorithm.

4.4.5. Resilience

When a station is notified that the ring attached is broken by receiving a topology message as described in above section, it starts forwarding the frames to the remaining direction to the receiver. This is called steering.

In a RPR ring, all stations will stop adding packets and discard all transit frames until the topology picture is re-established. Then, they will start steering the frames. This is to ensure that all frames are delivered in order. The time needed for this procedure is the restoration time of the ring. The RPR standard mandates the restoration time to be below 50 ms.

The PRP introduces an option to tolerate the non-ordered frames. A station may steer frames immediately after the failure has been detected. This means the station can immediately start wrapping the frames to the other direction if they are marked as eligible to wrapping.

4.4.6. Frame Formats

Four kinds of frames have been standardized by the RPRWG. They are data, fairness, control and idle frames.

Data frames have two formats, basic and extended. The basic data frame format is shown in figure 32.

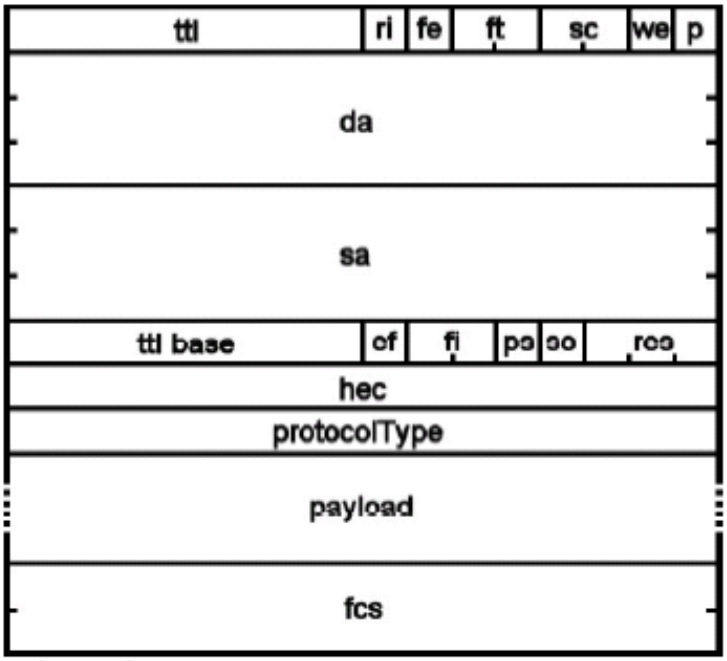


Figure 32: RPR basic data frame format

TTL	two byte "time to live" field

RI	"ring identifier" bit defines which ringlet the packet was inserted into initially
FE	"fairness eligible" bit indicates that the packet has to abide by the rules of the fairness algorithm
FT	two bit "frame type": Data, Fairness, Control, Idle
SC	two bit "service class":A0, A1, B, C
WE	"wrap eligible" bit defines if the frame can be wrapped at a wrap node
P	"parity" bit is reserved for future use in data frames
DA	six-byte "destination address"
SA	six-byte "source address"
TTL base	This field is set to the initial value of the "ttl" field when the packet is initially sourced into the ring. It is used for fast calculation of the number of hops that a packet has travelled.
EF	"extended frame" bit, indicating an extended frame format
FI	two bit "flooding indication" is set when a frame is flooded and if so, on one or both ringlets
PS	The "passed source" bit is set when passing its sender on the opposing ring after a wrap. The bit is used in detecting an error condition where a packet should have been stripped earlier.
SO	The "strict order" bit, if set, identifies that the frame should be delivered to its destination in strict order.
RES	three-bit reserved field
HEC	two byte "header error correction" field, protects the initial 16 bytes of the header

Table 6: RPR frame identifiers

The 16-byte fairness frame mainly provides the advertised "fairRate" and the source of the fairness frame. The information is used in the RPR fairness algorithm.

A control frame is similar to the data frame, but is distinguished by a designated "ft" field value and its controlType field specifies the type of information carried. There are different types of control frames in RPR, for example, topology and protection information and OAM (Operations Administration and Maintenance).

Idle frames are utilized in order to compensate rate mismatches between neighbouring stations.

4.5. Packet over SONET/SDH

SONET (Synchronous Optical Network) is a US standard for the internal operation of telephone company optical networks. It is closely related to a system called SDH (Synchronous Digital

Hierarchy) adopted by the CCITT (now the ITU-T) as a recommendation for the internal operation of carrier (PTT) optical networks worldwide. Despite the name SONET is not an optical networking system. It is an electronic networking system designed to use optical link connections.

SONET Optical Carrier Level	SONET Frame Format	SDH level and Frame Format	Payload bandwidth (kbit/s)	Line Rate (kbit/s)
OC-1	STS-1	STM-0	48,960	51,840
OC-3	STS-3	STM-1	150,336	155,520
OC-12	STS-12	STM-4	601,344	622,080
OC-24	STS-24	STM-8	1,202,688	1,244,160
OC-48	STS-48	STM-16	2,405,376	2,488,320
OC-96	STS-96	STM-32	4,810,752	4,976,640
OC-192	STS-192	STM-64	9,621,504	9,953,280
OC-768	STS-768	STM-256	38,486,016	39,813,120
OC-1536	STS-1536	STM-512	76,972,032	79,626,120
OC-3072	STS-3072	STM-1024	153,944,064	159,252,240

Table 7: SONET/SDH designations and bandwidths

Packet over SONET/SDH, abbreviated PoS, is a communications protocol for transmitting packets in the form of the Point to Point Protocol over SDH or SONET, which are both standard protocols for communicating digital information using lasers or light emitting diodes (LEDs) over optical fibre at high line rates. PoS is defined by RFC 2615 as PPP over SONET/SDH. PPP is the Point to Point Protocol that was designed as a standard method of communicating over point-to-point links. Since SONET/SDH utilises point-to-point circuits, PPP is well suited for use over these links. Scrambling is performed during insertion of the PPP packets into the SONET/SDH frame.

The most important application of PoS is to support sending of IP packets across wide area networks. Large amounts of traffic on the Internet are carried over PoS links. PoS is also one of the link layers used by the IEEE 802.17 Resilient Packet Ring standard.

PoS is a layer 2 technology that uses PPP in HDLC encapsulation, using SONET framing. The PoS solution lowers the cost per megabyte when compared to other wide area networking architectures. The PoS interface supports SONET level alarm processing, performance monitoring, synchronization, and protection switching. This support enables PoS systems to seamlessly interoperate with existing SONET infrastructures and provides the capability to migrate to IP+optical networks without the need for legacy SONET infrastructures. PoS is used in a point-to-point environment, much like the legacy T-carrier architectures, but without the need for TDM.

PoS efficiently encapsulates IP traffic with a low-overhead PPP header. When encapsulated, the traffic is placed inside an HDLC-delimited SONET SPE and transported across SONET. Voice, video, and data can be carried within the IP packets using layer 3 QoS mechanisms to control priority when bandwidth contention occurs.

PoS can be used in tandem with other technologies carried over SONET architectures. PoS is not compatible with these other technologies, but is not aware of them because they are being transported over different time slots. PoS, TDM voice, ATM, and Dynamic Packet Transport (DPT) can each use their required synchronous transport signals, not interacting with each other. PoS interfaces are available in concatenated and nonconcatenated (channelized) options. Channelized interfaces are more costly than concatenated interfaces.

4.5.1. PoS Transport

PoS does not require SONET transport but works in tandem with such as a result of the SONET framing that PoS employs. Two PoS devices can be connected directly with duplex fiber. Because PoS interfaces are layer 3 enabled, PoS interfaces are an example of an IP+optical architecture. Figure 33 displays three different ways in which PoS traffic can be transported. The three mechanisms are explained as follows:

- **Connectivity to SONET ADMs:** SONET circuits are provisioned as point-to-point circuits over SONET rings. Routers with PoS interfaces can be attached to SONET add/drop multiplexers (ADMs). As long as the proper number of STS are provisioned, the PoS interfaces will have connectivity. The PoS traffic is multiplexed with the other traffic that the SONET ADMs are carrying.
- **Connectivity to transponders in a DWDM system:** PoS traffic can be translated to a DWDM ITU-grid wavelength using a transponder. Most transponders support SONET framing. Through the DWDM system, 32 PoS circuits can be multiplexed onto one fiber.
- **Dark-fiber connectivity:** PoS interface can be connected directly over dark fiber using PoS interfaces. Dark fiber is fiber that is leased from a service provider; the customer provides the source (Laser or LED) and destination (photodiode receiver) devices. This process is normally referred to as lighting the fiber. Long spans can be accommodated through standard SONET regenerators that provide regeneration, reshaping, and retiming (3Rs) of the signal. The Cisco 15104 is an OC-48 SONET regenerator that fits this application.

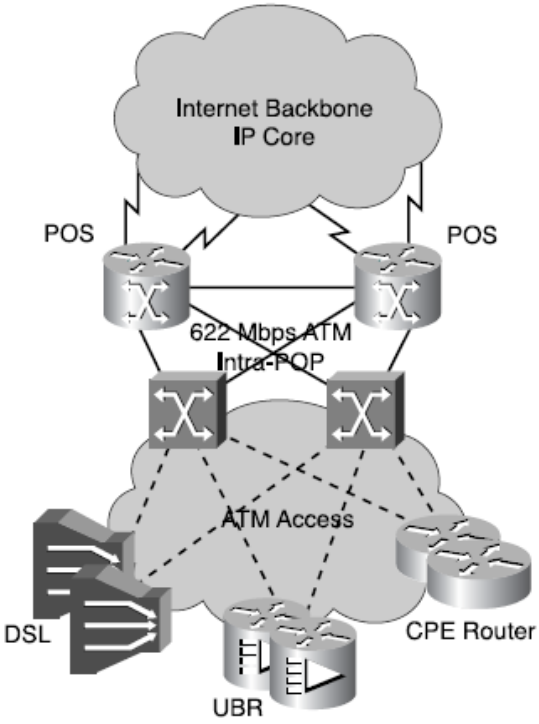


Figure 33: PoS transport options

4.5.2. Packet over SONET Operation and Specifications

The current Internet Engineering Task Force (IETF) PoS specification is RFC 2615 (PPP over SONET), which obsoletes RFC 1619. The PoS RFCs define the requirements that are needed to transport data packets through PoS across a SONET network. These requirements are summarized as follows:

- **High-order containment:** PoS frames must be placed in the required synchronous transport signals used in SONET. An example of this is an OC-12 concatenated PoS interface. This interface requires an STS-12 circuit to contain the required payload of the PoS traffic.
- **Octet alignment:** This refers to the alignment of the data packet octet boundaries to the STS octet boundaries. An octet (byte) defines an arbitrary group of 8 bits. The word byte is defined as usually containing 8 bits. IBM used to define a byte as containing 7 bits. Although both byte and octet are used interchangeably, octet is a more accurate representation for 8 bits because its meaning is aseries of eight.
- **Payload scrambling:** Scrambling is the process of encoding digital 1s and 0s onto a line in such a way that provides an adequate number for a 1s density requirement. The ANSI standard for T1 transmission requires an average density of 1s of 12.5 percent (a single 1 in 8 bits meets this requirement) with no more than 14 consecutive 0s for unframed signals and no more than 15 consecutive 0s for framed signals. The primary reason for enforcing a 1s density requirement is for timing recovery or network synchronization. However, other factors, such as automatic-line-build-out (ALBO), equalization, and power usage are affected by 1s' density. RFC 1619 inadvertently permitted malicious users to generate packets with bit patterns that could create SONET density synchronization problems and replication of the frame alignment. RFC 2615 provides a more secure mechanism for payload scrambling.

4.5.3. High-Order Containment

End stations at customer sites are predominantly TCP/IP-enabled devices. At the edge of the customer's network, the IP packet is encapsulated into a layer 2 format that will be supported on the SP's network. The layer 2 protocols supported by Cisco are PPP and Cisco HDLC, but the PoS standards specify PPP encapsulation for PoS interfaces. The Layer 2 PPP or Cisco HDLC frame information is encapsulated into a generic HDLC header (not Cisco proprietary HDLC) and placed into the appropriate SPE of the Whereas frame. This can be a confusing concept at first. Although HDLC and PPP are different, mutually exclusive layer 2 protocols, HDLC is used as a SPE delimiter in the SONET frame. The encapsulation process of an IP packet to a SONET frame is illustrated in figure 34.

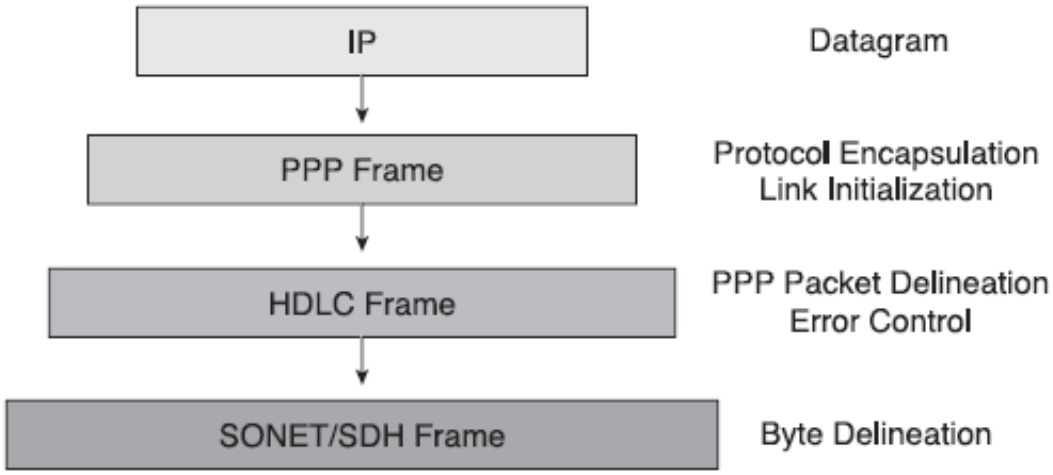


Figure 34: Encapsulating IP into a PoS frame

4.5.4. PPP Frame

RFC 1548 defines a PPP frame that contains the following three components:

- protocol field
- information field
- padding field

The protocol field is used because PPP was designed to be multiprotocol in nature. Multiprotocol encapsulations transport multiple protocols, including IP and IPX. The information field is the protocol data unit (PDU) transmitted, and can be from 0 to 64,000 bytes. The padding field is used to pad the PPP frame if the information field does not contain enough data. The padding field might receive padding up to the maximum receive unit (MRU), which will fill the information field. The default value for the MRU is 1500 octets but can be up to 64,000 octets if negotiated in the PPP implementation. It is the responsibility of the protocol to determine which bits are used as padding. More information about the PPP protocol can be found in RFC 1548 and RFC 1661 at www.ietf.org. Figure 35 illustrates the PPP in HDLC-like frame format.

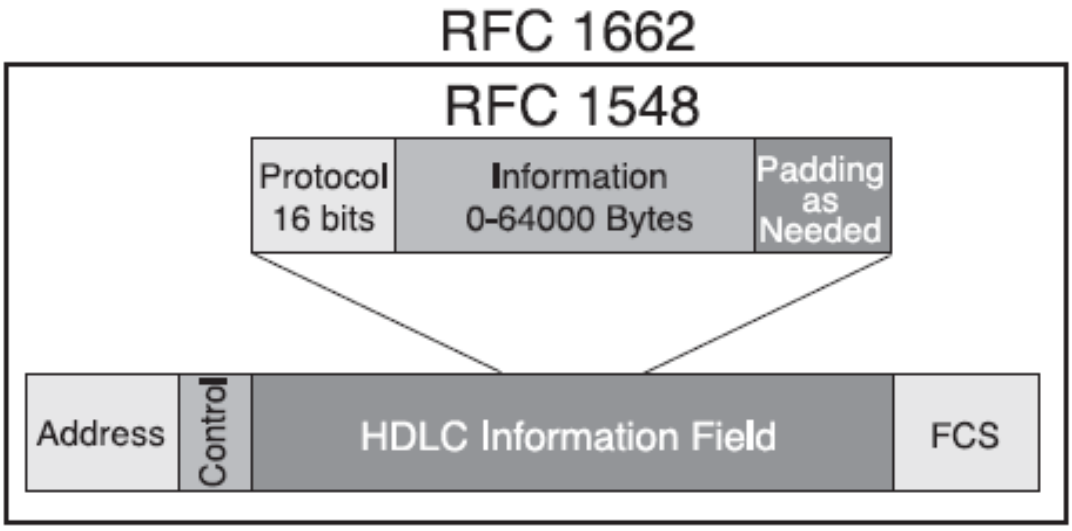


Figure 35: RFC 1662: PPP in HDLC-like framing

Figures 36 and 37 illustrate the values used in the PPP in the HDLC-like framing process. Notice that frame delimiters of hexadecimal 0x7E (126 in decimal) are used to denote the beginning and ending of a frame. The transmitting device generates flags as a time fill when there are no data packets. The address field is always set to 0xFF (255) because every frame is a broadcast frame in PoS. There are only two ends of the point-to-point connection, and the frame always needs to get to the other side. There is no reason to have more than one address because there are no other addressable destinations. The layer 2 mechanism is terminated at the other end of the link because PoS interfaces are layer 3 enabled. A control field of 0x03 (3) is used to denote an HDLC frame. The information field is where the PPP frame is inserted and is variable in nature due to MRU variability. A 16 or 32 bit frame check sequence (FCS) is used as a trailer to the frame. The FCS can be 16 or 32 bits long, but 32-bit CRCs are highly recommended due to the enhanced error recovery that is available using 32 bits. Most interfaces that run at speeds greater than OC-12 use FCS-32 as the default. The FCS is a configurable option, and FCS 32 is always recommended. The FCS field needs to match on both ends of the connection; otherwise, the layer 2 protocol will never come up.

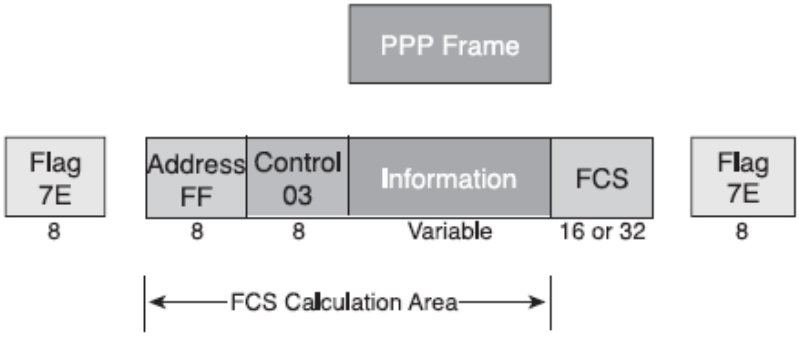


Figure 36: Packet over SONET frame information

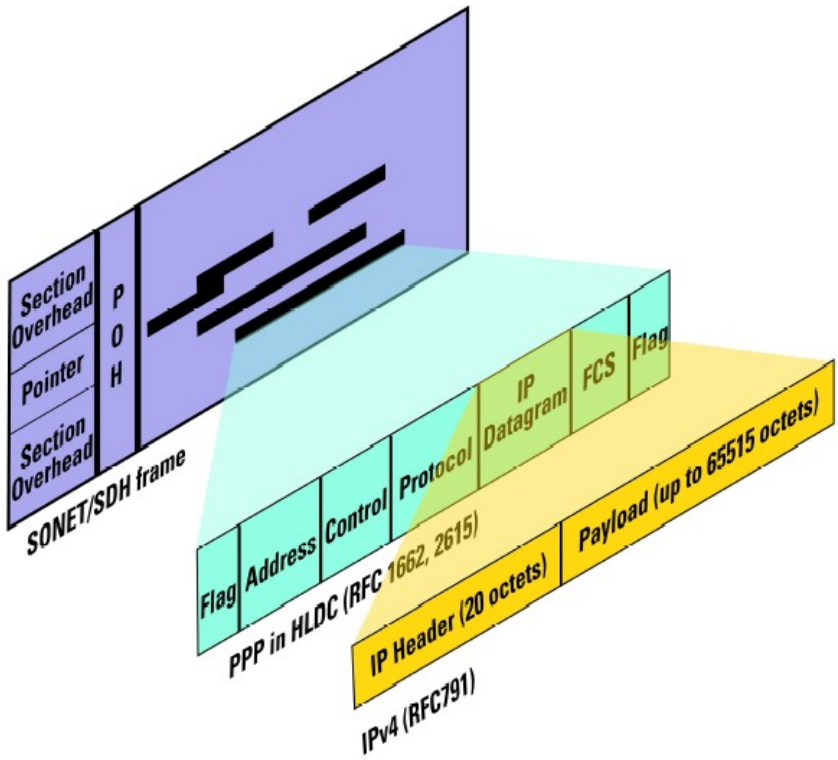


Figure 37: Packet over SONET/SDH

4.5.5. PoS Efficiencies

Overhead efficiency is a critical topic for SPs that charge for the amount of bandwidth capacity customers use. ATM was introduced to the SP market as the technology that would enable converged voice, video, and data traffic to reside on the same infrastructure (because of the intrinsic QoS parameters built in to the technology). The technology is widely used by Internet service providers (ISPs) today, but many of the original intentions behind ATM have not been used due to the complexity of configuring, selling, and maintaining such features. ISPs want to maximize their profits and minimize the costs associated with transporting IP over ATM.

While all packet sizes (and gaps) between the hardware minimum and 65,535 octets are possible, studies show that real Internet traffic exhibits a packet size distribution that is almost trimodal (figure 38).

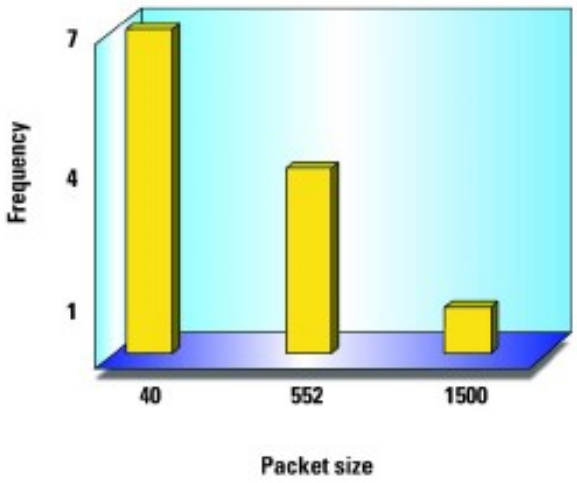
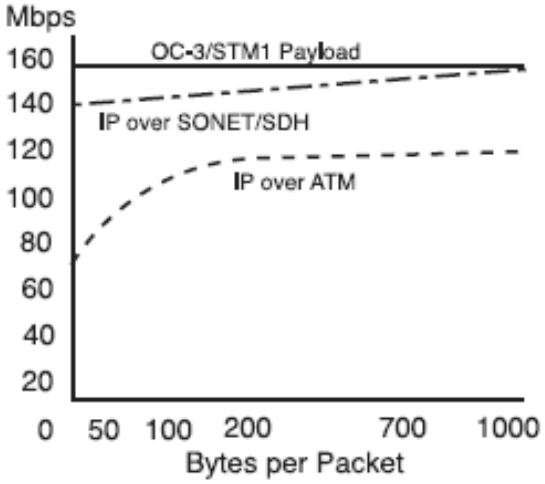


Figure 38: 7, 4, 1 distribution

For example, between 30% and 40% of the packets fall into the 40-byte category. These small packets are typically transmission control protocol (TCP) acknowledgement messages and appropriately considered to be an important corner case in the design of high performance routers. Other modes occur at 552 octets (TCP applications not performing maximum transmission unit discovery) and 1500 octets (maximum segment size for Ethernet). Figure 39 illustrates the PoS efficiency over ATM in both a line graph and table, which compares efficiency based on packet size.



Packet Size (byte)	POS Eff. (SPE %)	ATM Eff. (SPE %)
64	86.8	43
128	94	69
256	97	75
512	98.6	85
1024	99.3	86
1518	99.5	88
2048	99.6	89
4352	99.8	89

Figure 39: PoS efficiencies compared to ATM

5. Network Layer Protocols

5.1. IPv4

5.1.1. IPv4 overview

Internet Protocol version 4 is the fourth iteration of the Internet Protocol (IP) and it is the first version of the protocol to be widely deployed. IPv4 is the dominant network layer protocol on the Internet and apart from IPv6 it is the only network layer protocd used on the Internet.

It is described in IETF RFC 791 (September 1981), which obsoletes RFC 760 (January 1980). The United States Department of Defense also standardized it as MIL-STD-1777.

IPv4 is a data-oriented protocol to be used on a packet switched internet work (e. g. Ethernet). It is a best effort protocol, in that it does not guarantee delivery. It does not make any guarantees on the correctness of the data; it may result in duplicated packets and/or packets out-of-order. These aspects are addressed by an upper layer protocol (e. g. TCP, and partly by UDP).

The entire purpose of IP is to provide unique global computer addressing to ensure that two computers communicating over the Internet can uniquely identify each other.

IPv4 uses 32-bit addresses, which limits the address space to 4,294,967,296 possible unique addresses. However, some are reserved for special purposes, such as private networks (~18 million addresses) or multicast addresses (~1 million addresses). This reduces the number of addresses that can be allocated as public Internet addresses. As the number of addresses available is consumed, an IPv4 address shortage appears to be inevitable.

This limitation leads to the push towards IPv6, which is currently in the early stages of deployment and is currently the only contender to replace IPv4.

When writing IPv4 addresses in human readable form, the most common notation is the dot-decimal notation. There are other notations based on the values of the octets of the IP address.

Notation	Value	Conversion from dot-decimal
dot-decimal notation	207.142.131.235	n/a
dotted octal	0317.0216.0203.0353	each octet is individually converted into octal

Originally, the IP address was divided into two parts:

- **network ID:** first octet
- **host ID:** last three octets

This created an upper limit of 256 networks and led to the creation of classful networks. Under classful networking, 5 classes were created (A, B, C, D and E) with 3 created (A, B and C) with different lengths of network number and rest fields to change the number of IPs in each range: few networks with lots of addresses and numerous networks with only a few addresses. Class D was for multicast addresses and class E is reserved.

The actual assignment of an address is not arbitrary. The fundamental principle of routine is that the address encodes information about a device's location within a network. This implies that an address assigned to one part of a network will not function in another part of the network. A hierarchical structure, created by CIDR (Classless InterDomain Routing) and overseen by the Internet Assigned Numbers Authority (IANA) and its Regional Internet Registries (RIRs),

manages the assignment of Internet address worldwide. Each RIR maintains a publicly searchable WHOIS database that provides information about IP address assignments; information from these databases plays a central role in numerous tools that attempt to locate IP addresses geographically.

Address block	Description	Reference
0.0.0.0/8	current network (only valid as source address)	RFC 1700
10.0.0.0/8	private networks	RFC 1918
14.0.0.0/8	public data networks	RFC 1700
127.0.0.0/8	localnet	RFC 3330
169.254.0.0/16	link-local zeroconf	RFC 3927
172.16.0.0/12	private networks	RFC 1918
192.0.2.0/24	documentation and example code	RFC 3330
192.88.99.0/24	IPv6 to IPv4 relay	RFC 3068
192.168.0.0/16	private networks	RFC 1918
198.18.0.0/15	network benchmark tests	RFC 2544
224.0.0.0/4	multicast (former class D network)	RFC 3171
240.0.0.0/4	reserved (former class E network)	RFC 1700
255.255.255.255	broadcast	

Table 8: Special IPv4 address ranges

In addition to private networking, the IP range 127.0.0.0–127.255.255.255 (or 127.0.0.0/8 in CIDR notation) is reserved for localhost communication. Any address within this range should never appear on an actual network and any packet sent to this address does not leave the source computer, and will appear as an incoming packet on that computer (known as loopback).

5.1.2. The Domain Name System

The Internet resources are most publicly known not by IP addresses but by names (e. g. www.whitehouse.gov, www.freebsd.org, www.mit.edu). The routing of IP packets across the Internet is oblivious to such names. This requires translating (or resolving) names to IP address.

The Domain Name System (DNS) provides such a system to convert names to IP address(es) and IP addresses to names. Much like CIDR addressing, the DNS naming is also hierarchical and allows for sub delegation of name spaces to other DNS servers.

One method to increase both address utilization and security is to use network address translation (NAT). By assigning one IP to a public machine as an internet gateway and using a private network for an organization's computers allows for considerable address savings. This also increases security by making all of the computers on a private network not directly accessible from the public network.

5.1.3. Virtual Private Networks

Since all public routers deliberately ignore private address ranges, it is not normally possible to connect two private networks (e. g. two branch offices) via the public Internet. Virtual private networks (VPNs) solve this problem.

VPNs work by inserting an IP packet (encapsulated packet) directly into the data field of another IP packet (encapsulating packet) and using a publicly routable address in the encapsulating packet. Once the VPN packet is routed across the public network and reaches the endpoint, the encapsulated packet is extracted and then transmitted on the private network just as if the two private networks were directly connected.

Optionally, the encapsulated packet can be encrypted to secure the data while over the public network (see VPN for more details).

5.1.4. Address Resolution

IP is an upper layer protocol to the data link layer. The data link layer of the underlying physical network segment over which two communicating computers are directly connected (typically through a hub or a switch) uses its own addressing scheme at the hardware level. In order to send a packet from computer A to B, A needs to know the hardware address of B. This discovery and mapping of IP addresses onto the hardware addresses is done by using the Address Resolution Protocol (ARP).

A different case is when a computer knows its data link layer address but not its IP address. This is a common scenario in private networks and Digital Subscriber Line (DSL) connections when the IP addresses of the machines are irrelevant. This is usually the case for workstations but not servers. In this case a machine needs an answer to the question: "This is my hardware address, what is my IP address?" RARP was the first method for answering this question. Then, for most cases, BOOTP has been used instead, which in turn was obsoleted by the Dynamic Host Configuration Protocol (DHCP).

In addition to sending the IP address, DHCP can also send the NTP server, DNS servers, and more.

5.1.5. IP Header

An IP packet consists of two sections:

1. header
2. data

The header consists of 13 fields, of which only 12 are required. The 13th field is optional and aptly named: Options. The fields in the header are packed with the most significant byte first, and for the diagram and discussion, the most significant bits are considered to come first. The most significant bit is numbered 0, so the version field is actually found in the 4 most significant bits of the first byte, for example.

+	Bits 0–3	4–7	8–15	16–18	19–31													
0	Version				Header length	Type of Service (now DiffServ and ECN)							Total Length					
32	Identification												Flags	Fragment Offset				
64	Time to Live								Protocol				Header Checksum					
96	Source Address																	
128	Destination Address																	
160	Options																	
160/192+	Data																	

Figure 40: IPv4 packet schematics

Version The first header field in an IP packet is the 4-bit version field. For IPv4, this has a value of 4 (hence the name IPv4).

Internet Header Length (IHL) The second field is a 4-bit Internet Header Length (IHL) telling the number of 32-bit words in the header. Since an IPv4 header may contain a variable number of options, this field specifies the size of the header (this also coincides with the offset to the data). The minimum header size is 20 bytes, so the minimum value for this field is 5 (5×4 = 20 bytes). Being a 4-bit field the maximum length is 15 words or 60 bytes.

Type of Service (TOS) In RFC 791, the following 8 bits were allocated to a Type of Service (TOS) field:

- bits 0-2: precedence
- bit 3: 0 = normal delay, 1 = low delay
- bit 4: 0 = normal throughput, 1 = high throughput
- bit 5: 0 = normal reliability, 1 = high reliability
- bits 6-7: reserved for future use

The TOS field is now used for DiffServ and ECN. The original intention was for a sending host to specify a preference for how the datagram would be handled as it made its way through an inter-network. For instance, one host could set its IPv4 datagrams TOS field value to prefer low delay, while another might prefer high reliability. In practice, the TOS field has not been widely implemented. However, a great deal of experimental, research and deployment work has focused on how to make use of these eight bits. These bits have been redefined, most recently through DiffServ working group in the IETF and the Explicit Congestion Notification codepoints. New technologies are emerging that require real-time data streaming and therefore will make use of the TOS field. An example is Voice over IP (VoIP) that is used for interactive data voice exchange.

Total Length This field defines the entire datagram size, including header and data, in bytes. The minimum-length datagram is 20 bytes (20 bytes header + 0 bytes data) and the maximum is 65,535 — the maximum value of a 16-bit word. The minimum size datagram that any host is required to be able to handle is 576 bytes, but most modern hosts handle much larger packets.

Sometimes sub-networks impose further restrictions on the size, in which case datagrams must be fragmented. Fragmentation is handled in either the host or packet switch in IPv4.

Identification This field is an identification field and is primarily used for uniquely identifying fragments of an original IP datagram. Some experimental work has suggested using the ID field for other purposes, such as for adding packet-tracing information to datagrams in order to help trace back datagrams with spoofed source addresses.

Flags A 3-bit field follows and is used to control or identify fragments. They are (in order, from high order to low order):

- Reserved, must be zero
- Don't Fragment (DF)
- More Fragments (MF)

If the DF flag is set and fragmentation is required to route the packet then the packet will be dropped. This can be used when sending packets to a host that does not have sufficient resources to handle fragmentation.

When a packet is fragmented all fragments have the MF flag set except the last fragment, which does not have the MF flag set. The MF flag is also not set on packets that are not fragmented — clearly an un-fragmented packet can be considered the last fragment.

Fragment Offset The fragment offset field is 13-bits long and allows a receiver to determine the place of a particular fragment in the original IP datagram, measured in units of 8-byte blocks. This method allows a maximum offset of 65,528, which would exceed the maximum IP packet length of 65,535 with the header length counted with it. Therefore the 1st bit of the fragment offset is mostly unused.

Time To Live (TTL) An 8-bit time to live (TTL) field helps prevent datagrams from persisting (e. g. going in circles) on an inter-network. Historically the TTL field limited a datagram's lifetime in seconds, but has come to be a hop count field. Each packet switch (or router) that a datagram crosses decrements the TTL field by one. When the TTL field hits zero, the packet is no longer forwarded by a packet switch and is discarded. Typically, an ICMP message (specifically the time exceeded) is sent back to the sender that it has been discarded. The reception of these ICMP messages is at the heart of how traceroute (a program for tracing routes) works.

Protocol This field defines the protocol used in the data portion of the IP datagram. The Internet Assigned Numbers Authority maintains a list of protocol numbers and was originally defined in RFC 790.

Header Checksum The 16-bit checksum field is used for error checking of the header. At each hop, the checksum of the header must be compared to the value of this field. If a header checksum mismatches, then the packet is discarded.

Source address An IP address is a group of 4 8-bit octets for a total of 32 bits. The value for this field is determined by taking the binary value of each octet and concatenating them together to make a single 32-bit value.

This address is the address of the sender of the packet. Note that this address may not be the "true" sender of the packet due to network address translation. Instead, the NATing machine will translate the source address to its own address. Thus, reply packets sent by the receiver are routed to the NATing machine, which translates the destination address to the original sender's address.

Destination address Identical to the source address field but indicates the receiver of the packet.

Options Additional header fields (called options) may follow the destination address field, but these are not often used. The list of options may be terminated with an EOL (End of Options List) option; this is only necessary if the end of the options would not otherwise coincide with the end of the header.

5.1.6. Fragmentation

To make IPv4 more tolerant of different networks the concept of fragmentation was added so that, if necessary, a device could break up the data into smaller pieces. This is necessary when the maximum transmission unit (MTU) is smaller than the packet size.

The reason fragmentation was chosen to occur at the IP layer is that IP is the highest application-independent layer each packet has to pass while travelling between two different physical networks. If fragmentation were performed on higher layers (TCP, UDP, etc.) then this would make fragmentation/reassembly be redundantly implemented (once per protocol); if fragmentation were performed on a lower layer (Ethernet, ATM, etc.) then this would require fragmentation/reassembly be performed on each hop (could be quite costly) and redundantly implemented (once per link layer protocol). Therefore, the IP layer is the most efficient one for fragmentation.

When a device receives an IP packet it examines the destination address and determines the outgoing interface to use. This interface has an associated MTU that dictates the maximum data size for its payload. If the MTU is smaller than the data size then the device must fragment the data.

The device then splits the data into segments where each segment is less-than-or-equal-to the MTU less the IP header size (20 bytes minimum, 60 bytes maximum). Each segment is then put into its own IP packet with the following changes:

- The total length field will be adjusted to the segment size.
- The more fragments (MF) flag is set for all segments except the last one.
- The fragment-offset field is set accordingly based on the offset of the segment in the original data payload. This is measured in units of 8-byte blocks.

Notice that if MTU minus header length is not a multiple of 8, then only a multiple of 8 number of bytes of data will be included in the datagram, even if that leaves a total datagram size of less than MTU (could only be off by 4 bytes because header is always multiple of 4 bytes).

By some chance if a packet changes link layer protocols or the MTU reduces then these fragments would be fragmented again.

5.1.7. Reassembly

When a receiver detects an IP packet where either of the following is true:

- "more fragments" flag set
- "fragment offset" field is non-zero

then the receiver knows the packet is a fragment. The receiver then stores the data with the identification field, fragment offset, and the more fragments flag. When the receiver receives a fragment with the more fragments flag not set then it knows the length of the original data payload since the fragment offset plus the data length is equivalent to the original data payload size.

Once it has all the fragments then it can reassemble the data in proper order (by using the fragment offsets) and pass it up the stack forfurther processing.

5.1.8. IPv4 Limitations

It is very important to reflect on some key elements of the original IPv4 architecture. All the early papers and practice on internet architecture stress that each computer attached to the Internet will have a globally unique IP address. Typical is this passage from Doug Comer's 1988 text on TCP/IP: "Each host on the Internet is assigned a unique 32-bit Internet address that is used in all communication with that host." (Douglas Comer, Internetworking with TCP/IP: Principles, Protocols, and Architecture, Prentice-Hall, 1988.) Thus, if one speaks of the IPv4 architecture, it is understood that globally unique IP addresses per host is part of that architecture. Further, the applications-level flexibility provided by globally unique addresses helps explain the ongoing vitality of applications innovation within the Internet. If, for example, a hard decision had been made at the outset of the Internet that some hosts would be clients and others would have been servers, then this would have constrained and ultimately weakened the early work on voice over IP, on person-to-person chats, andon teleconferencing.

5.1.9. Lack of IPv4 Address Space and Scalability of Routing

Therefore, ideally, every IP connected device should have a unique IP address. This is currently not practical with the IPv4 32-bit address space. The original IPv4 address space cannot sustain the original IP addressing architecture, given the growth in the number of devices capable of performing as IP hosts, including PDAs, mobile phones, and other appliances. Internet access devices are becoming more capable and powerful, thanks to their faster CPU, increased bandwidth and storage capacity, as well as new peripherals - such as cameras, sensors, GPS, etc. This leads to a new trend that enables such devices to provide application services, in addition to being ordinary clients.

5.1.10. NAT

In order to try and solve the problems arising from lack of addresses, mechanisms like Network Address Translation (NAT) coupled with Application Level Gateways (ALGs), such as web-proxies, are used. An IPv4-NAT access network simply does not allow client devices to run as servers, because of NAT's inability to map incoming connections to its clients' private IP addresses. One workaround to this problem is to deploy an ALG (Application Level Gateway) on the NAT for each service of interest. This solution incurs significant protocol design and service deployment complexity. Furthermore, this is not a scalable solution since it requires protocol-specific changes for each service that needs to go through NAT. Therefore, while ALGs could provide a costly solution for enabling few services through NATs, they fail to restore the much-needed end-to-end transparency that is lost upon NAT deployment.

One of the most critical deficiencies of the IPv4-NAT architecture is its inability to allow hosts to run as the connection-receiving end of a communication session (similar to servers). Therefore, while such mechanisms are useful, they destroy global addressability and disrupt the end-to-end service semantics of application-level communication; which has implications for services such as security, mobility and QoS. That interferes with application innovation by removing the ability of one host to initiate direct communication with another host. Instead, all applications must be mediated by a central server with a global IP address. Apart from this major negative impact on application innovation, there are other negative impacts on performance and network management.

The widespread deployment of NATs as the solution to the lack of IPv4 addresses is architecturally radical, in that it changes the essence of the Internet architecture. NATs destroy both global addressability and end-to-end transparency, a key internet architectural principle. According to the principle of end-to-end transparency, all the routers and switches between a pair of communicating hosts simply pass IP packets along and do not modify their contents

(apart from decrementing the TTL field of the IP header at each hop along the path). This principle is key to the support for new applications. A few examples:

- The new generation of SIP-based interpersonal communications applications, including voice over IP, innovative forms of messaging, presence, and conferencing, makes effective use of central servers to allow users to locate each other, but then also makes effective use of direct host-to-host communications in support of the actual communications. This enables applications flexibility and allows for high performance.
- Other conferencing applications, such as VRVS, also require direct host-to-host communications and break when either user is placed behind a NAT.
- The new grid computing paradigm supports high-speed distributed computing by allowing flexible patterns of computer-to-computer communications. The performance of such systems would be crippled were it required for servers to be involved in these computer-to-computer communications.

NAT also makes network troubleshooting more difficult. When NAT and other middleboxes modify the contents of the packets, it becomes more difficult for applications developers to understand how to get new applications (those not known when the given middlebox was designed) to work. NAT boxes also break a number of tools, such as ping and traceroute, that depend on adherence to the classic Internet architecture and which are key to diagnosing network problems. Both expert ISP engineers and ordinary users have their time wasted trying to debug network problems either caused by the NAT boxes or made more difficult to diagnose by the NAT boxes. Thus, the key issue is not so much IPv4 vs IPv6 per se, but rather classic IP (either v4 or v6 but without NATs in the path) vs NATted IPs.

5.1.11. Services for Grid Computing

IPv4 has already many of the features required for grids; among them there are: uniformity (i. e. its abstraction over every network technology), dynamism (i. e. automatic routing reconfiguration), scalability (although limited by 32-bit addresses) by using its multicast addressing capability, improved security using IPSec, Quality of Service enhancements, automatic configuration (DHCP), reduced management of routing tables (CIDR), and improved mobility support.

Some of these features were a natural evolution of the original IPv4 design, some have been developed by IPv6 network engineers and then, in their entirety or in similar functions, have already been incorporated back into IPv4.

5.1.12. Security Considerations

IPsec has been an important step forward to security at the network service level. Its development and wide deployment will be key to scalable secure communications as the Internet continues to grow. IPsec is important both for pure host-to-host and for support by gateways in a variety of ways. However, it is important to note that:

- When no NATs are in the path, IPv4 can also provide quite good support for IPsec. Thus, statements of the form "IPv4 supports IPsec almost as well as IPv6 does" are correct.
- But when NATs are present in the path, IPv4 will not be able to support IPsec well. Although NATs are expected to be less important in the IPv6 infrastructure, IPv6 NATs are conceivable and, when actually present, they would also defeat support for IPsec.

5.1.13. QoS Support

Grid computing provides a global-scale distributed computing infrastructure for executing scientific and business applications. An important requirement is the need to make this infrastructure appear as a single logical co-ordinated resource – which in reality consists of a variety of resources aggregated across different administrative domains. The aggregation of network resources is often undertaken over a 'best effort' infrastructure as provided by the Internet – however many applications, which necessitate soft-real time constraints, such as collaborative working or remote visualisation, require more stringent traffic guarantees. Managing Quality of Service (QoS) requirements across these aggregated resources therefore becomes an important concern. Such QoS criteria must extend to computational, data and network resources, and are often expressed in a Service Level Agreement, multiples of which may co-exist over the entire collection of resources. Such an SLA must contain descriptions of network, storage and computational resources, and forms the basis of a grid QoS framework.

IPv4 was never designed to support real-time, interactive, QoS-sensitive applications. Currently the Internet treats all traffic equally as 'best effort' and provides no support for any other Quality of Service (QoS). There has been work within the IETF over many years to try and define QoS support and resource control in the network and in IPv4. The Internet Engineering Task Force (IETF) has proposed the Integrated Services (IntServ) and the Differentiated Services architectures (DiffServ) [Barden et al. 1994, Blake et al. 1998]. Both architectures support QoS and provide a type of guarantee, in terms of bandwidth, latency and other data transfer parameters, to deliver network traffic between a source and destination.

5.1.14. IntServ

IntServ enables the user to reserve resources by maintaining per flow admission control, signalling, classification and scheduling at every router on the path from source to destination. Hence, IntServ can provide 'per flow' guarantees to users. Scalability is the main issue negating the deployment of the IntServ architecture on the Internet. Provision of 'maintenance of state' information, for huge number of flows passing through the Internet core routers, needs enormous resources. This approach is therefore virtually impossible to deploy in reality.

5.1.15. DiffServ

DiffServ, however, provides a broad, and flexible, range of services avoiding a 'per flow' state in core routers. DiffServ is based on the use of the DS field (the first 6 bits of the TOS octect in the IPv4 header) [RFC 2474, RFC 2475]. The main goal of the DiffServ architecture is to provide a preferred level of service to particular types of network traffic without increasing overhead in the core routers. Essentially, DiffServ provides an aggregated end-to-end service over a number of separately administered domains. At inter-domain level there has to be a mechanism to exchange critical information among domains about aggregated flows.

A Bandwidth Broker has been introduced [Teitelbaum et al. 1999] as a resource management entity that provides the necessary functionality for allocation of intra-domain resources, and arranging inter-domain agreements.

A Bandwidth Broker (BB) is a logical entity responsible for managing QoS resources in an administrative domain, based on a Service Level Agreement (SLA). A SLA is the contract between two domains, or between a domain and a client, which specifies, to the forwarding service, the amount of traffic the client can receive. Organizational policies can be configured by using the mechanism provided by a BB. On the inter-domain level the BB is responsible for negotiating QoS parameters and setting up bilateral agreements with neighbouring domains. On the intra-domain level a BB's responsibility includes configuration of edge routers, to enforce resource allocation and for admission control. Edge routers can be configured to (mainly) police and classify/mark packets with the corresponding DSCP (DiffServ Code Point). Policing entails

ensuring the received rate does not exceed the agreed-on rate; if exceeded, depending on the adopted policy, excess packets are either discarded or re-marked for delayed discard if there is congestion.

Bandwidth Broker (BB) in DiffServ. The DiffServ architecture supports a simple mechanism to provide QoS to network traffic. The traffic entering a DiffServ domain is classified and conditioned at the boundary of the network and then assigned to different behaviour aggregates. The flows entering a domain are classified into one of many classes based on the value of the DSCP in the header of the packet. All packets with the same DSCP are treated in the same manner, and belong to the same behaviour aggregate (BA). The core routers forward packets, according to the treatment required, on the basis of their BA.

The main resource management entity in a DiffServ domain is the BB, which maintains policies and negotiates SLAs with clients and neighbouring domains. The interactions of BB with other components of a DiffServ domain, and the end-to-end communication process in a DiffServ domain are shown in figure 41. This figure shows that when a flow needs to enter the DiffServ domain, or a local user wants to send traffic, the BB is requested to check related SLAs and the present traffic condition on the network. The BB decides whether or not to allow the traffic, on the basis of the previously-negotiated SLAs – to ensure new traffic does not violate current SLAs. If there is a new flow, the BB might have to negotiate a new SLA with the neighbouring domain(s) depending on the traffic requirements. Once the BB allows the traffic, the edge or leaf router needs to be reconfigured by the BB. SLA negotiation is a dynamic process that needs to take into account the ever changing requirements of network traffic. The BB is responsible for admission control, as it has global knowledge of network topology and resource allocation.

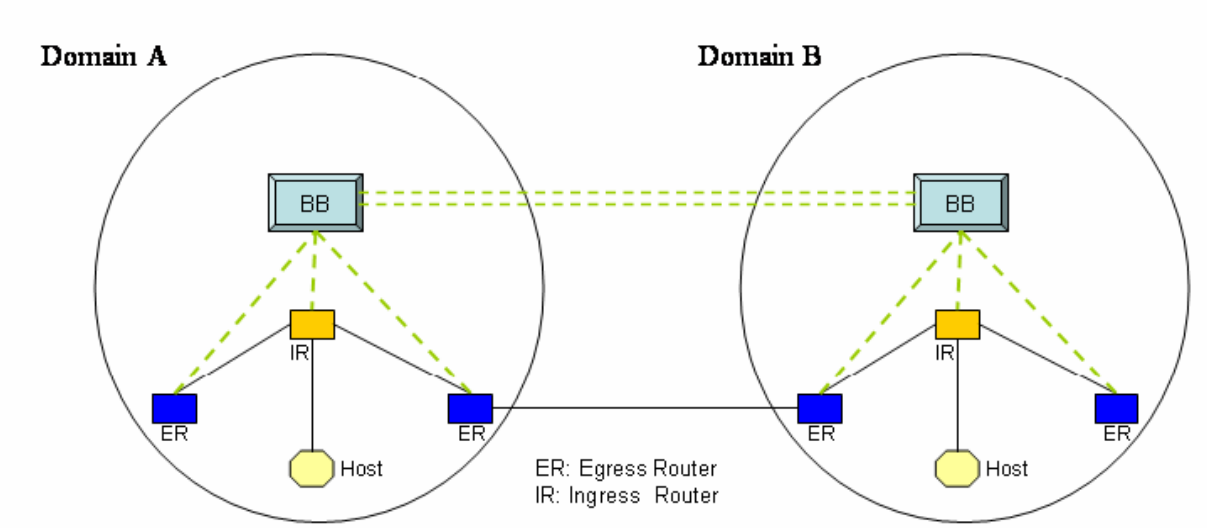


Figure 41: Role of bandwidth broker in DiffServ

A remark: IntServ and DiffServ are applicable to both IPv4 and IPv6; so it could be argued that there will be no advantage to IPv6 in terms of QoS support. However, coupled with the simplified header and better addressing capability, we should see improved performance with IPv6, and the flow-label in its header has the potential to allow additional QoS support.

For example, some grid applications would like to have protected flows in the network. The processing of QoS sensitive flows in a router could be simplified by uniquely identifying end-to-end flows in the IPv6 header, rather than using transport- or link-layer header fields.

Therefore such QoS-sensitive grid applications could benefit from using the IPv6 flow label to identify individual end-to-end flows.

5.1.16. Explicit Congestion Notification

TCP's congestion control and avoidance algorithms are based on the notion that the network is a black-box. The network's state of congestion or otherwise is determined by end-systems probing for the network state, by gradually increasing the load on the network (by increasing the window of packets that are outstanding in the network) until the network becomes congested and a packet is lost. Treating the network as a "black-box" and treating loss as an indication of congestion in the network is appropriate for pure best-effort data carried by TCP, with little or no sensitivity to delay or loss of individual packets.

This behaviour cannot offer constant performance and bandwidth to network applications. For example grid applications generating bursts of heavy network traffic are penalized by TCP's slow start.

The proposal of ECN RFC 3168 Draft-ecnsyn is an extension to the Internet protocols to provide a congestion indication for incipient congestion where the notification can sometimes be through marking packets rather than dropping them. It is based on the use of the DS field (the last 2 bits of the TOS octect in the IPv4 header). That could help in avoiding packet drops by signaling congestion to end-systems. They could react appropriately, thus collaborating in overcoming the incipient congestion.

5.1.17. Network Architectures for QoS in Grid Computing

With the emerging interest in grid computing, several research groups have been actively involved in bringing concepts and mechanisms from the network community to support network QoS for grid applications. Most of these activities are based on the Differentiated Services architecture (DiffServ). Two substantial recent efforts in the grid community aim at addressing the issue of introducing DiffServ-based QoS provision for use in grid applications, namely:

1. General-purpose Architecture for Reservation and Allocation (GARA)
2. Network Resource Scheduler (NRS) project.

5.1.18. GARA

The General-Purpose Architecture for Reservation and Allocation (GARA) is the best known framework for supporting QoS in computational grids, and provides the capability for specifying end-to-end QoS requirements; its advance reservation service treats various types of resources uniformly, such as network, computation and storage, and provides a guarantee that an application initiating a reservation will receive a specific QoS from the Resource Manager. GARA also provides an application-programming interface to create, modify, bind and cancel reservation requests. Network QoS in GARA is designed and built to work with a specific network router, the Cisco 7507, and uses Cisco's Modular QoS Command-line interface (MQC) to configure routers, i. e. a Policy Enforcement Point (PEP), to support DiffServ capability. In a multidomain network, the GARA system must exist in every administrative domain. In making a network reservation for traffic spanning multiple administrative domains, two issues need to be resolved: 1) locating/contacting the GARA system in each domain along the traffic path, and 2) ensuring that the application/user requesting the reservation has secure access to each GARA system along the path. This introduces manageability limitations and constraints on the types of domains in which GARAcan be deployed.

5.1.19. NRS

The Network Resource Scheduler (NRS) on the other hand adopts the Peer-to-Peer (P2P) model, as a NRS exists in every administrative domain and it is assumed there is a trust relationship between neighbouring NRSs. The NRS uses the DiffServ concept, and therefore every neighbouring NRS has a DiffServ Service Level Agreement (SLA). The application/user requesting a network QoS need only negotiate with the local NRS and establish a local SLA. During the negotiation process the local NRS replicates the request to all NRSs along the network path to conduct an admission control check and, subsequently, establish a SLA. NRS, somewhat like GARA, is designed and built to only work with Cisco routers and makes use of Cisco-ISO to configure Cisco's routers (PEP) to support DiffServ capability.

Although NRS has demonstrated its effectiveness in providing DiffServ QoS, it is not clear how a grid application developer would make use of this capability especially as the application programming interface is not clearly defined. The use of a NRS also requires the definition of specific network parameters such as Traffic Specification (TSpec), token bucket size and token bucket rate – which require advancednetworking knowledge.

5.2. IPv6

5.2.1. The IPv6 Protocol

IPv6 was invented to succeed IPv4 as the network layer standard used to address equipment connected to the Internet, allowing for a much larger address space than its predecessor. Even though IPv4 was originally introduced in 1981 (RFC 791), it has not been substantially changed since then. Among the characteristics that made IPv4 being predominately deployed are its robustness, its easy implementation and its interoperability. Nevertheless, the most important reason was that it stood the test of scaling to the size of internet today. However, while IPv4 supports $4.3 \cdot 10^9$ unique addresses, IPv6 can support up to $3.4 \cdot 10^{38}$ of them, solving the problem of address exhaust that would normally occur following the trend of new networked devices such as cell phones and PDAs. Techniques such as Network Address Translation (NAT) and Classless Inter-Domain Routing (CIDR) were deployed in the past to extend IPv4 lifetime. Those techniques appeared to increase the available address space but failed to meet the requirements of the emerging peer-to-peer applications.

IPv6 was adopted by IETF in 1994 (back then as "Internet Protocol next generation" – IPng), and although its market share is still minimal (mainly slowed down by the introduction of Network Address Translation/NAT, which however breaks the end-to-end nature of the Internet) it is expected that at some point it will be broadly adopted. It is quite certain, however, that it will co-exist with IPv4 for the foreseeable future.

IPv6 technologies for transition from IPv4 but also for allowing IPv6-only hosts and networks to reach other (either IPv6 or IPv4) similar entities over IPv4 networks, have stabilized for quite some time now. We expect that it will be straightforward to implement IPv6 over the SEEREN2 backbone, since all routers with a decently recent operating system support it. It is true, however, that research is still ongoing to develop new services that take advantage of the IPv6 features and resolve problems with currently existing such services. Another challenge for the project will be to deploy IPv6 on a wide basis, within the participating NRENs and the universities. RFC 2460 describes the latest IPv6 specification.

5.2.2. IPv6 Features

IPv6 includes a number of improvements over IPv4, the most significant of which are the following:

Larger address space. This was the initial motivation for the specification of IPv6, as explained earlier in this document. This extension from 32-bit addresses to 128-bit addresses has some interesting repercussions: Large address blocks can be allocated, due to the vast number of available addresses, which leads to smaller routing tables and easier management of the address space; also, scanning the IPv6 address space for vulnerabilities is much harder than with IPv4, which makes IPv6 more resilient to attacks.

Stateless autoconfiguration of hosts. When an IPv6-enabled host is connected to a respective network, it issues a broadcast message requesting configuration information. The gateway can then provide a router advertisement with such information included, so that the host is automatically configured. This feature makes renumbering easier and as a result the migration from one ISP to another reduces significantly the network management overheads.

Multicast is a part of the protocol itself – although IPv6 multicast support at the moment is very limited within most routers. IPv6 does not use broadcast at all. The functions previously supported by IPv4 (e. g. router discovery) are directly handled by the protocol itself. IPv6 uses specific multicast group addresses for its various functions. Thus, IPv6 multicast prevents the problems caused by broadcast storms in IPv4 enabled networks.

Jumbograms. IPv6 includes support for packets larger than 64 kb, when the link layer supports this option. Such packets are called jumbograms and can improve performance significantly over high-throughput links.

IPsec. Although not widely deployed at the moment as a feature, IPsec is natively supported by IPv6 and it is part of the protocol suite. IPv6 provides security extension headers, making easier the implementation of encryption and authentication. On top of this, end-to-end security services can be provided omitting the need of additional hardware machines that typically introduce additional administrative overheads and performance bottlenecks.

Last but not least even though QoS in IPv6 is handled the same way as it is currently handled in IPv4 a new field in its header enables QoS devices in the path to take appropriate actions based on this label.

5.2.3. IPv6 Header Structure

The fields of IPv6 header are the following:

1. **Version:** 4 bits for version number (6 in our case).
2. **Class:** The 8-bit traffic class field can be used in Differentiated Services, to provide Quality of Service. This field is a part of the Quality of Service supplement that is not provided in IPv4.
3. **Flow Label:** 20 bits for defining a sequence of packets that a source sends to the destination. The value of this field can define special handling method according to which routers will handle the flow packets. This field is a part of the Quality of Service supplement that is not provided in IPv4.
4. **Payload Length:** 16 bits that hold the length of the extension headers including the transport-level PDU.
5. **Next Header:** 8 bits for identifying the type of the next header.
6. **Hop Limit:** 8 bits for the limited number of hops remaining for a packet. Each router decreases the number by one and when it reaches 0 it is being discarded (similar to time-to-live field in IPv4).
7. **Source address:** 128 bits address of the sender.
8. **Destination address:** 128 bits address of the receiving host.

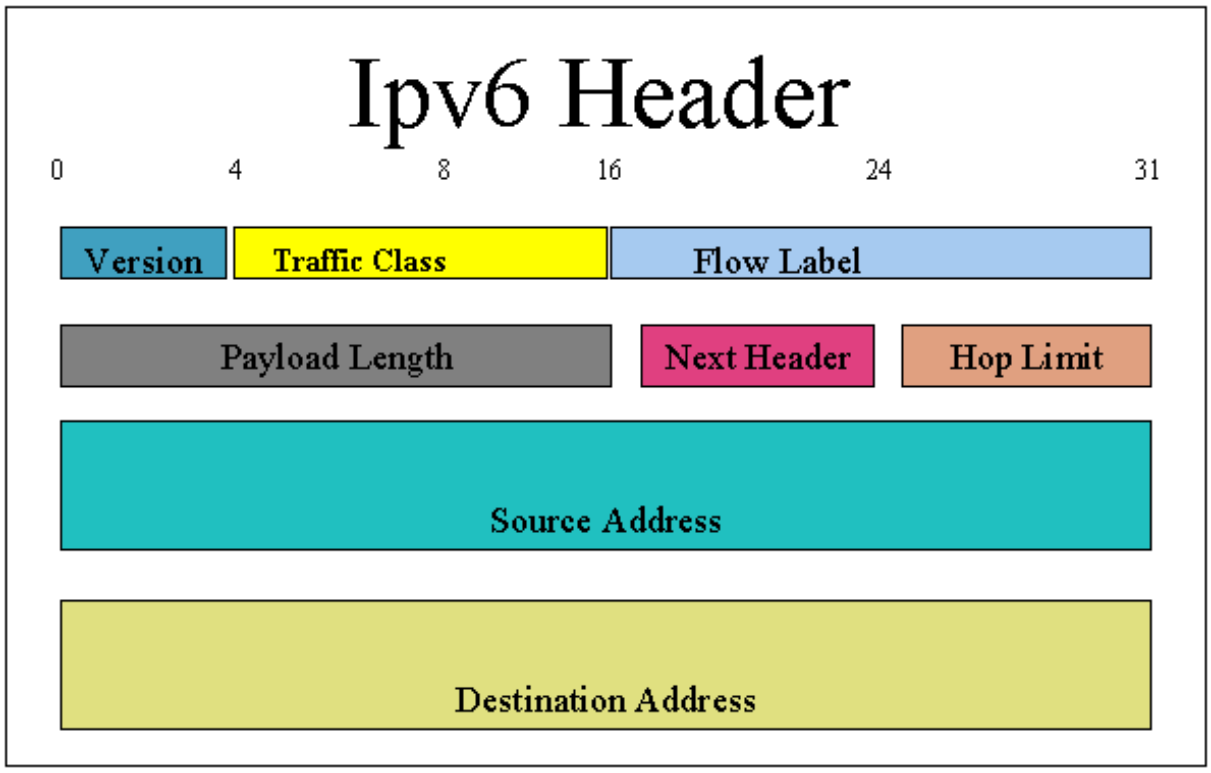


Figure 42: IPv6 header

5.2.4. IPv6 Addressing

The primary change from IPv4 to IPv6 is the length of network addresses. IPv6 addresses are 128 bits long (as defined by RFC 4291), whereas IPv4 addresses are 32 bits; while the IPv4 address space contains 4,294,967,296 addresses, IPv6 has enough room for 340,282,366,920,938,463,463,374,607,431,768,211,456 unique addresses.

IPv6 addresses are typically composed of two logical parts: a 64-bit (sub-) network prefix, and a 64-bit host part, which is either automatically generated from the interface's MAC address or assigned sequentially. Because the globally unique MAC addresses offer an opportunity to track user equipment, and so users, across time and IPv6 address changes, RFC 3041 was developed to reduce the prospect of user identity being permanently tied to an IPv6 address, thus restoring some of the possibilities of anonymity existing at IPv4. RFC 3041 specifies a mechanism by which variable over time random bit strings can be used as interface circuit identifiers, replacing unchanging and traceable MAC addresses.

IPv6 addresses are normally written as eight groups of four hexadecimal digits. For example, 2001:0db8:85a3:08d3:1319:8a2e:0370:7334 is a valid IPv6 address.

If a four-digit group is 0000, the zeros may be omitted. For example, 2001:0db8:85a3:0000:1319:8a2e:0370:1337 can be shortened as 2001:0db8:85a3::1319:8a2e:0370:1337. Following this rule, any group of consecutive 0000 groups may be reduced to two colons, as long as there is only one double colon used in an address. Leading zeros in a group can also be omitted. Thus, the addresses below are all valid and equivalent:

- 2001:0db8:0000:0000:0000:0000:1428:57ab
- 2001:0db8:0000:0000:0000::1428:57ab

- 2001:0db8:0:0:0:0:1428:57ab
- 2001:0db8:0:0::1428:57ab
- 2001:0db8::1428:57ab
- 2001:db8::1428:57ab

Having more than one double-colon abbreviation in an address is invalid, as it would make the notation ambiguous.

A sequence of 4 bytes at the end of an IPv6 address can also be written in decimal, using dots as separators. This notation is often used with compatibility addresses (see below). Thus, ::ffff:1.2.3.4 is the same address as ::ffff:102:304.

In a URL the IPv6-Address is enclosed in brackets. Example:

http://[2001:0db8:85a3:08d3:1319:8a2e:0370:7344]/

This notation allows parsing a URL without confusing the IPv6 address and port number:

http://[2001:0db8:85a3:08d3:1319:8a2e:0370:7344]:443/

Additional information can be found in "RFC 2732 - Format for Literal IPv6 Addresses in URL's" and "RFC 3986 - Uniform Resource Identifier (URI): Generic Syntax".

IPv6 networks are written using CIDR notation. An IPv6 network (or subnet) is a contiguous group of IPv6 addresses the size of which must be a power of two; the initial bits of addresses, which are identical for all hosts in the network, are called the network's prefix.

A network is denoted by the first address in the network and the size in bits of the prefix (in decimal), separated with a slash. For example, 2001:0db8:1234::/48 stands for the network with addresses 2001:0db8:1234:0000:0000:0000:0000:0000 through 2001:0db8:1234:FFFF:FFFF:FFFF:FFFF:FFFF.

Because a single host can be seen as a network with a 128-bit prefix, you will sometimes see host addresses written followed with /128.

There are a number of addresses with special meaning in IPv6:

- ::/128 — The address with all zeros is an unspecified address, and is to be used only in software.
- ::1/128 — The loopback address is a localhost address. If an application in a host sends packets to this address, the IPv6 stack will loop these packets back to the same host (corresponding to 127.0.0.1 in IPv4).
- ::/96 — The zero prefix was used for IPv4-compatible addresses.
- ::ffff/96 — This prefix is used for IPv4 mapped addresses.
- 2001:db8::/32 — This prefix is used in documentation (RFC 3849). Anywhere where an example IPv6 address is given, addresses from this prefix should be used.
- fc00::/7 — Unique local IPv6 unicast addresses are routable only within a set of cooperating sites. They were defined in RFC 4193 as a replacement for site-local addresses (see below). The addresses include a 40-bit pseudorandom number that minimizes the risk of conflicts if sites merge or packets somehow leak out.
- fe80::/64 — The link-local prefix specifies that the address only is valid in the local physical link. This is analogous to the autoconfiguration IP address 169.254.x.x in IPv4.

- fec0::/10 — The site-local prefix specifies that the address is valid only inside the local organisation. Its use has been deprecated in September 2004 by RFC 3879 and systems must not support this special type of address.
- ff00::/8 — The multicast prefix is used for multicast addresses.

There are no address ranges reserved for broadcast in IPv6 — applications use multicast to the all-hosts group instead.

5.2.5. QoS in IPv6

There has been a long history of discussions in reference of QoS support in IPv6 environments. There is a debate on whether "IPv6 provides better QoS support than IPv4" or "IPv6 experiences worst performance than IPv4".

The IPv6 header is (re)designed to minimize header overhead and reduce the header process for the majority of the packets. This is achieved by moving less essential and optional fields to extension headers that are placed after the IPv6 header. Therefore, IPv6 and IPv4 headers are not interoperable. Furthermore, IPv6 header is not a superset – thus backward compatible – with IPv4 counterpart.

The IPv6 header has two fields that are related to QoS; the traffic class and flow label fields. The 8-bit traffic class field is used to distinguish packets from different classes or priorities. The same functionality is provided from the type of service (or precedence) field in the IPv4 header and, consequently, there is no essential difference among the packet headers of the two protocols.

By definition, a flow is a sequence of packets sent from a particular source to a particular unicast, anycast, or multicast destination. In the IPv4 world, flow classification is based on 5 fields; IP source and destination addresses, transport layer protocol type and ports. However, some of these fields may be unavailable due to fragmentation or encryption of packets in the network. In order to overcome such problems, flow classification in IPv6 world is based on the 3-tuple consisting of the flow label plus the source and destination address fields, which are in fixed predefined positions in the IPv6 header. The flow label field consists of 20 consecutive bits. Whenever the end host wants to identify the packets of a flow, it sets the flow label bits to the same non-zero value, which is unchanged throughout the network. Note that currently there is no application or service that takes advantage of the flow label field. It is easily concluded that IPv6 protocol, in terms of QoS functionality, is neither superior nor inferior to IPv4 counterpart. However, the available flow label field in the IPv6 header could be a valuable tool for the provision of services in the future.

5.2.6. Improving Performance with IPv6 Jumbograms

The packet size field of IPv4 and IPv6 has a size of 16 bits; hence, IP packets are limited to a maximum size of 64 KiB (= 2^{16} B). An optional feature of IPv6, the jumbo payload option, allows the exchange of packets larger than this size between cooperating hosts. Since both TCP and UDP include fields limited to 16 bits (length, urgent data pointer), support for IPv6 jumbograms requires slight tweaks to the transport layer. Both the jumbo payload options and the transport-layer tweaks are described in RFC 2675.

The IPv6 header [IPv6] has a 16-bit Payload Length field and, therefore, supports payloads up to 65,535 octets long. The Jumbo Payload option, carries a 32-bit length field in order to allow transmission of IPv6 packets with payloads between 65,536 and 4,294,967,295 octets in length. Packets with such long payloads are referred to as "jumbograms".

The jumbo payload option is relevant only for IPv6 nodes that may be attached to links with a link MTU greater than 65,575 octets (that is, 65,535 + 40, where 40 octets is the size of the

IPv6 header). The jumbo payload option need not be implemented or understood by IPv6 nodes that do not support attachment to links with MTU greater than 65,575.

On links with configurable MTUs, the MTU must not be configured to a value greater than 65,575 octets if there are nodes attached to that link that do not support the jumbo payload option and it can not be guaranteed that the jumbo payload option will not be sent to those nodes.

The UDP header [UDP] has a 16-bit Length field which prevents it from making use of jumbograms, and though the TCP header [TCP] does not have a length field, both the TCP MSS option and the TCP Urgent field are constrained to 16 bits. This document specifies some simple enhancements to TCP and UDP to enable them to make use of jumbograms.

5.2.7. IPv6 Security

IPv6 is not necessarily more secure than IPv4. In fact, IPv6 approach to security is only marginally better than IPv4 but not radically new. The following sub-sections summarize some IPv6's improvements that provide for better network security

Large address space. Port scanning is one of the best known reconnaissance techniques in use today. Port scanning allows "black-hats" to listen to specific services (ports) that could be associated to well-known vulnerabilities. In IPv4 networks, port scanning is a relatively simple task. Most IPv4 segments are of class C, with 8 bits allocated for host addressing. In IPv6 networks, the landscape is radically different. IPv6 subnets use 64 bits for allocating host addresses. Scanning such a large address space is almost an impossible task. However, it is not absolutely impossible.

IPSec. As mentioned above, IPv4 also offers IPSec support. However, IPv4's support for IPSec is optional. By contrast, the RFC 4301 mandates for IPv6 to use IPSec in all nodes. IPSec consists of a set of cryptographic protocols that provide for securing data communication and key exchange. IPSec uses two wire-level protocols, Authentication Header (AH) and Encapsulating Security Payload (ESP). The first protocol provides for authentication and data integrity. The second protocol provides for authentication, data integrity, and confidentiality. In IPv6 networks both the AH header and the ESP header are defined as extension headers. Additionally, IPSec provides for a third suite of protocols for protocol negotiation and key exchange management known as the Internet Key Exchange (IKE). This protocol suite provides the initial functionality needed to establish and negotiating security parameters between endpoints. Additionally, it keeps track of this information to guarantee that communication continues to be secure up to the end.

Neighbor discovery and address autoconfiguration. Neighbor discovery (ND) is the mechanism responsible for router and prefix discovery, duplicate address and network unreachability detection, parameter discovery, and link-layer address resolution. This protocol is entirely network-layer based. ND operates in tandem with auto-configuration, which is the mechanism used by IPv6 nodes to acquire either stateful or stateless configuration information. In the stateless mode, all nodes get what they need for global communication, including potential illegal ones. In stateful mode, configuration information can be provided selectively, reducing the possibility for rogue nodes. Both ND and address auto-configuration contribute to make IPv6 more secure than its predecessor. IPv6 provides for TTL values of up to 255; it prevents against outside sourcing of ND packets or duplicate addresses.

Unfortunately, IPv6 also has some disadvantages as far as security is concerned. For instance, attackers can utilize the automatic autoconfiguration of IPv6 to gain access to networks to which they would have not had access to otherwise. Additionally, the possibility to have several addresses for the same computer can be an administrator's nightmare.

Nevertheless, IPv6 represents an improvement if compared to the old IPv4 protocol stack. The new suite of protocols provides innumerable features that improve both the overall functionality as well as some specific security functions. However, it is far from being a panacea. Although IPv6 offers better security (larger address space and the use of encrypted communication), the protocol also raises new security challenges. Ultimately, the new protocol creates as many new security problems as it solves old ones. And if that is not enough, the transition from the old protocol stack to the new one may present even more challenges, something that will guarantee plenty of fun for security network professionals in the foreseeable future.

5.2.8. Combination of IPv6 and Grid Systems

Grid systems are normally considered as network middleware, since they lie between applications and network resources. The data of grid systems is currently transported using IPv4. The next generation Internet protocol IPv6 is improving IPv4 in this regard.

Since IPv6 is expected to become the standard for global networks, grid computing systems must track the migration of the lower-layer network protocols to IPv6. The period of transition from IPv4 to IPv6 will take time. It is important to make grid systems work on both IPv4 and IPv6, and to be able to communicate in heterogeneous IPv4/IPv6 networks.

While it is clear to network professionals that IPv6 is an important development, most of those concerned with grid computing are not interested in the network level at all. This has resulted in some problems in the way that the relevant software has been structured, which causes some problems in the migration to IPv6. There is a very substantial body of activity in the development of grid computing. It deals with the provision of networks, the provision of special middleware between grid applications and the network software, and the applications themselves. While it is intended that grid computing be carried out over the general Internet or enterprise intranet, the requirements made by the networks on either the applications or the middleware are largely ignored.

While it is important to take advantage of IPv6 features, we expect IPv4 environments to persist for a long time. This makes it vital to consider the heterogeneous IPv4/IPv6 networks. One effort to integrate IPv6 into grid systems takes an IP-protocol independent approach, i. e. it supports both IPv4 and IPv6. The IP-independent server has to be able to respond to client calls according to the IP family that the client uses. In other words, the client decides which version of IP is to be used. The grid server responds to the client calls according to which IP family the client uses. For instance, an IP-independent grid server on a dual-stack machine starts and listens on both its IPv4 and IPv6 interfaces. For communication in heterogeneous IPv4/IPv6 networks, there are a number of network transition aids, which essentially translate the packet headers between IPv4 and IPv6, leaving the payload untouched. These approaches may work in certain circumstances for grid applications.

To start any IPv6 experiments, the host must be IPv6 enabled. The IPv6 capable application API libraries are required in order to run the IPv6 enabled applications or IP independent applications over IPv6. All network associated applications, such as network-sharing database applications and web containers, need to be IPv6 enabled to run IPv6 tests. In order to run tests over a network rather than only on local hosts, IPv6 support on the network is essential. It requires IPv6-enabled routers, which provide forwarding and dynamic routing, and support from IPv6-enabled network services, such as IPv6 DNS, web services, etc. A number of the major router manufactures provide now basic IPv6 support and are beginning to provide more advanced support such as hardware forwarding. Support for IPv6 in the DNS – provides hostname and IPv6 address resolution which may be provided over IPv4 and/or IPv6 connection. For the communication in heterogeneous IPv4/IPv6 networks, there are many approaches to the provision of transition aids. They need to be considered when building an IPv6 environment within or around current global IPv4 networks. The integration of IPv6 into

the grid systems starts with finding IP-version dependencies in the network protocols. The implementation of network APIs within applications may involve a few IP-dependent functions.

IPv6 still has many compatibility problems, both with grid middleware as well with other applications:

- **IPv6 compatibility of grid middleware.** An official document of Globus Toolkit v3.2 (GT is one of the main components of grid middleware) states IPv6 compliance. However, Shen Jiang of University College of London has recommended GT v3.2.1 for full IPv6 support (see http://www.cs.ucl.ac.uk/staff/s.jiang/webpage/how-to-IPv6-Globus.htm).
  There are no official statements about the IPv6 compliance of Condor software, nor of its main components like Condor-C or Condor-G. Therefore it is our understanding that Condor software is not IPv6 compliant, but tests must be done in order to confirm this hypothesis.
  Furthermore, LSF by Platform Corporation (Platform LSF HPC is software for managing workloads of mission-critical High Performance Computing applications) is officially declared to be IPv6 compliant and to support a dual IPv4/IPv6 stack (see http://www.platform.com/NR/rdonlyres/B5CCBFDC-FCD4-458B-B183-357BB3DFC234/0/PltLSF7_HPC_DS_Oct06.pdf).
  Finally, there are no official statements about the IPv6 compliance of PBS both in the OpenPBS and PBSpro versions.
- **IPv6 compatibility of associated applications.** It must be pointed out that, in order to ensure IPv6 support, all network-related applications used within Globus need to be IPv6 enabled: Java Database Connectivity, which is used for reliable file transfer, needs an IPv6 patch. As recommended by the Globus Implementation Group, Jakarta Tomcat should be used as web container for grid services: The container environment needs to provide IPv6 web services for grid services. Tomcat version 5 has been tested for IPv6 capabilities. Finally, Java is IPv6 compliant already on the following operating systems: Solaris 8 and higher, Linux kernel 2.1.2 and higher (kernel 2.4.0 and higher recommended for better IPv6 support), Windows XP SP1 and Windows 2003.

6. Transport and Application Layer

6.1. Datagram Congestion Control Protocol

6.1.1. DCCP Protocol Overview

Historically, the great majority of Internet unicast traffic has used congestion-controlled TCP, with UDP making up most of the remainder. UDP has mainly been used for short, request-response transfers, like DNS and SNMP, that wish to avoid TCP's three-way handshake, retransmission, and/or state-full connections. UDP also avoids TCP's built-in end-to-end congestion control, and UDP applications tend not to implement their own congestion control. However, since UDP traffic volume was small relative to congestion-controlled TCP flows, the network didn't collapse. Recent years have seen the growth of applications that use UDP in a different way. These applications, including streaming audio, internet telephony, and multiplayer and massively multiplayer on-line games, share a preference for timeliness over reliability. TCP can introduce arbitrary delay because of its reliability and in-order delivery requirements; thus, the applications use UDP instead. This growth of long-lived non-congestion-controlled traffic, relative to congestion-controlled traffic, poses a threat to the overall health of the Internet: UDP traffic is able to saturate the connections and TCP cannot take countermeasures. Applications could implement their own congestion control mechanisms on a case-by-case basis, with encouragement from the IETF. Some already do this. However, experience shows that congestion control is difficult to get right, and many application writers would like to avoid reinventing this particular wheel. Datagram Congestion Control Protocol is a new protocol that combines unreliable datagram delivery with built-in congestion control. This protocol may be used in order to transfer timely data without destabilizing the Internet.

There are several reasons why protocds currently use UDP instead of TCP, among them:

- **Startup delay:** They wish to avoid the delay of a three-way handshake before initiating data transfer.
- **Statelessness:** They wish to avoid holding connection state, and the potential state-holding attacks that come with this.
- **Trading of reliability against timing:** The data being sent is timely in the sense that if it is not delivered by some deadline (typically a small number of RTTs), then the data will not be useful at the receiver.

Of these issues, applications that generate large or long-lived flows of datagrams, such as media transfer and games, mostly care about controlling the trade-off between timing and reliability. Such applications use UDP because when they send a datagram, they wish to send the most appropriate data in that datagram. If the datagram is lost, they may or may not resend the same data, depending on whether the data will still be useful at the receiver. Data may no longer be useful for many reasons:

- In a telephony or streaming video session, data in a packet comprises a time-slice of a continuous stream. Once a time-slice has been played out, the next time-slice is required immediately. If the data comprising that time-slice arrives at some later time, then it is no longer useful. Such applications can cope with masking the effects of missing packets to some extent, so when the sender transmits its next packet, it is important for it to only send data that has a good chance of arriving in time for its play-out.
- In an interactive application like grid computing, games or virtual-reality session, position information is transient. If a datagram containing position information is lost,

resending the old position does not usually make sense — rather, every position information datagram should contain the latest position information.

In a congestion-controlled flow, the allowed packet-sending rate depends on measured network congestion. Thus, some control is given up to the congestion control mechanism, which determines precisely when packets can be sent. However, applications could still decide, at transmission time, which information to put in a packet. TCP doesn't allow control over this; these applications demand it.

Often, these applications (especially games and telephony applications) work on very short play-out timescales. Whilst they are usually able to adjust their transmission rate based on congestion feedback, they do have constraints on how this adaptation can be performed so that it has minimal impact on the quality of the session. Thus, they tend to need some control over the short-term dynamics of the congestion control algorithm, while being fair to other traffic on medium timescales.

DCCP connections are congestion controlled, but unlike in TCP, DCCP applications have a choice of congestion control mechanism. In fact, the two half-connections can be governed by different mechanisms. Mechanisms are denoted by one-byte congestion control identifiers, or CCIDs. The endpoints negotiate their CCIDs during connection initiation. Each CCID describes how the HC-sender limits data packet rates, how the HC-receiver sends congestion feedback via acknowledgements, and so forth. CCIDs 2 and 3 are currently defined; CCIDs 0, 1, and 4-255 are reserved. Other CCIDs may be defined in the future. CCID 2 provides TCP-like congestion control, which is similar to that of TCP. The sender maintains a congestion window and sends packets until that window is full. Packets are acknowledged by the receiver. Dropped packets and ECN [RFC 3168] indicate congestion; the response to congestion is to halve the congestion window. Acknowledgements in CCID 2 contain the sequence numbers of all received packets within some window, similar to a selective acknowledgement (SACK) [RFC 2018]. CCID 3 provides TCP-Friendly Rate Control (TFRC), an equation-based form of congestion control intended to respond to congestion more smoothly than CCID 2. The sender maintains a transmit rate, which it updates using the receiver's estimate of the packet loss and mark rate. CCID 3 behaves somewhat differently than TCP in the short term, but is designed to operate fairly with TCP over the long term.

For simplicity, we say that senders send DCCP Data packets and receivers send DCCP Ack packets. Both of these categories are meant to include DCCP DataAck packets. The phrases "ECN-marked" and "marked" refer to packets marked ECN Congestion Experienced unless otherwise noted.

6.1.2. TCP-like Congestion Control

TCP-like Congestion Control is appropriate for DCCP flows that would like to receive as much bandwidth as possible over the long term, consistent with the use of end-to-end congestion control. CCID 2 flows must also tolerate the large sending rate variations characteristic of AIMD congestion control, including halving of the congestion window in response to a congestion event. Applications that simply need to transfer as much data as possible in as short a time as possible should use CCID 2. This contrasts with CCID 3, TCP-Friendly Rate Control (TFRC) [RFC 4342], which is appropriate for flows that would prefer to minimize abrupt changes in the sending rate. For example, CCID 2 is recommended over CCID 3 for streaming media applications that buffer a considerable amount of data at the application receiver before playback time, insulating the application somewhat from abrupt changes in the sending rate. Such applications could easily choose DCCP's CCID 2 over TCP itself, possibly adding some form of selective reliability at the application layer. CCID 2 is also recommended over CCID 3 for applications where halving the sending rate in response to congestion is not likely to interfere with application-level performance. An additional advantage of CCID 2 is that its

TCP-like congestion control mechanisms are reasonably well understood, with traffic dynamics quite similar to those of TCP.

Differences between CCID 2 and straight TCP congestion control include the following:

- CCID 2 applies congestion control to acknowledgements, a mechanism not currently standardized for use in TCP.
- DCCP is a datagram protocol, so several parameters whose units are specified in bytes in TCP, such as the congestion window cwnd, have units of packets in DCCP.
- As an unreliable protocol, DCCP never retransmits a packet, so congestion control mechanisms that distinguish retransmissions from new packets have been redesigned for the DCCP context.

6.1.3. TFRC Congestion Algorithm

TFRC is a congestion control mechanism designed for unicast flows operating in an internet environment and competing with TCP traffic. TFRC is designed to be reasonably fair when competing for bandwidth with TCP flows, where a flow is "reasonably fair" if its sending rate is generally within a factor of two of the sending rate of a TCP flow under the same conditions. However, TFRC has a much lower variation of throughput over time compared with TCP, which makes it more suitable for applications such as telephony or streaming media where a relatively smooth sending rate is of importance. The penalty of having smoother throughput than TCP while competing fairly for bandwidth is that TFRC responds slower than TCP to changes in available bandwidth. Thus TFRC should only be used when the application has a requirement for smooth throughput, in particular, avoiding TCP's halving of the sending rate in response to a single packet drop. For applications that simply need to transfer as much data as possible in as short a time as possible we recommend using TCP, or if reliability is not required, using an Additive-Increase, Multiplicative-Decrease (AIMD) congestion control scheme with similar parameters to those used by TCP. TFRC is designed for applications that use a fixed packet size, and vary their sending rate in packets per second in response to congestion. Some audio applications require a fixed interval of time between packets and vary their packet size instead of their packet rate in response to congestion. The congestion control mechanism should be used by those applications.

TFRC is a receiver-based mechanism, with the calculation of the congestion control information (i. e. the loss event rate) in the data receiver rather in the data sender. This is well suited to an application where the sender is a large server handling many concurrent connections, and the receiver has more memory and CPU cycles available for computation. In addition, a receiver-based mechanism is more suitable as a building block for multicast congestion control.

In order to compete fairly with TCP, TFRC uses a throughput estimation, which roughly describes TCP's sending rate as a function of the loss event rate, round-trip time, and packet size. A loss event occurs when one or more packets are lost or marked in a window of data, where a marked packet refers to a congestion indication from ECN. Generally speaking, TFRC's congestion control mechanism works as follows:

- The receiver measures the loss event rate and feeds this information back to the sender.
- The sender also uses these feedback messages to measure the round-trip time (RTT).
- The loss event rate and RTT are then fed into TFRC's throughput estimation, giving the acceptable transmit rate.
- The sender then adjusts its transmit rate to match the calculated rate.

The dynamics of TFRC are sensitive to how the measurements are performed and applied. Specific mechanisms are recommended in [RFC 3448] in order to perform and apply these

measurements. Other mechanisms are possible, but it is important to understand how the interactions between mechanisms affect the dynamics of TFRC.

6.1.4. Quickstart

Quickstart (QS) is a new IETF experimental protocol that has been designed to provide lightweight signaling of the level of congestion (specifically available capacity) between routers and a pair of communicating end hosts. QS was originally conceived to improve the performance of TCP bulk transfers over lightly-loaded network paths. QS may also be useful for multimedia flows. In this case it can alleviate the effect of slowstarting to the encoding rate, and after periods of silence.

The sender sends a QS request for its required sending rate (measured in bytes per second) using a Quickstart option placed in the IP header. Figure 43 presents the header structure for a QS request packet. With multimedia streaming applications, the QS request uses the media encoding rate to select the requested sending rate.

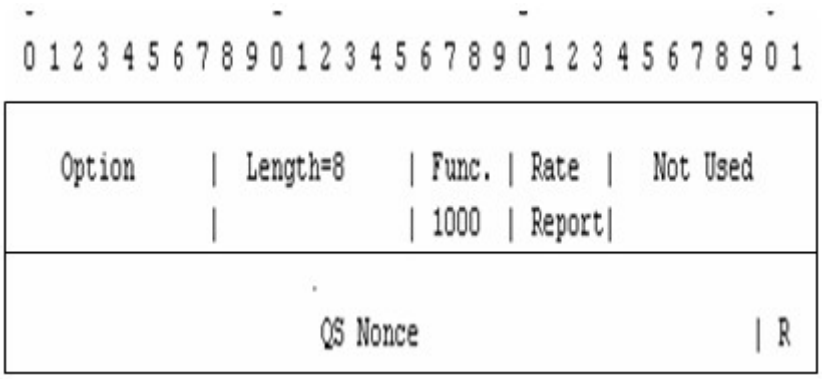


Figure 43: QS request header

Each router along the path should, in turn, either approve the requested rate, reduce the requested rate, or indicate that the Quickstart request is not approved (possibly by ignoring this new option). In approving a Quickstart request, a router does not give preferential treatment to subsequent packets from that connection; the router is only asserting that it is currently underutilized and believes there is sufficient available capacity to accommodate the sender's requested rate. The Quickstart mechanism can determine if there are routers along the path that do not understand the Quickstart option, or have not agreed to the Quickstart rate request. On receipt of the QS request, the receiver communicates the final rate request value (if the requested rate is approved) that it receives to the sender in a transport-level Quickstart response. Figure 44 presents the QS response header structure.

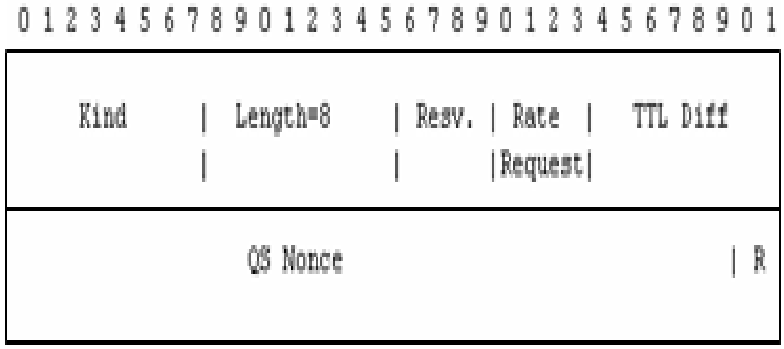


Figure 44: QS response header

If the Quickstart request is approved by all routers along the path, then the sender can send at up to the approved rate for a window of data. Subsequent transmissions will be governed by the default transport protocol's congestion control mechanisms of that connection. If the Quickstart request is not approved, then the sender would use the default congestion control mechanisms.

Even though QS was proposed with TCP in mind, QS could be used with any congestion control protocol that would prefer to inflate their sending rates without effectively slowstarting from a small initial rate. Using QS the standard Internet protocols can effectively and efficiently work over a wide range of links — including those with large propagation delay, as satellite is for example. Hence QS may be very useful for multimedia flows. In this case it can alleviate the effect of slowstart to the encoding rate, and after periods of silence.

6.1.5. Overhead

The applications we are concerned with often send compressed data, or send frequent small packets. For example, when internet telephony or streaming media are used over low-bandwidth modem links, highly compressing the payload data is essential. For internet telephony applications and for games, the requirement is for low delay, and hence small packets are sent frequently. For example, a telephony application sending a 5.6 Kbps data stream but wanting moderately low delay may send a packet every 20 ms, sending only 14 data bytes in each packet. In addition, the IP header takes up 20 bytes, with additional bytes for transport and/or application headers.

Clearly, it is desirable for such an application to have a low-overhead transport protocol header. In some cases, the correct solution would be to use link-based packet header compression to compress the packet headers, although we cannot guarantee the availability of such compression schemes on any particular link. The delay of data until after the completion of a handshake also represents potentially unnecessary overhead.

6.1.6. Firewall Traversal

This is one of the key factors that applications require. Typically, the firewall needs to parse RTSP, SIP, and H.323 to obtain the information necessary to open a hole in the firewall. Although, for bidirectional flows, the firewall can open a bidirectional hole if it receives a UDP packet from inside the firewall, in this case the firewall can't easily know when to close the hole again. Currently, streaming media players attempt UDP first, and then switch to TCP if UDP is not successful. Streaming media over TCP is undesirable and can result in the receiver needing to temporarily halt play-out while it "re-buffers" data. Telephony applications don't even have this option.

6.1.7. Parameter Negotiation

Different applications have different requirements for congestion control, which may map into different congestion feedback. Examples include ECN capability and desired congestion control dynamics (the choice of congestion control algorithm and, therefore, the form of feedback information required). Such parameters need to be reliably negotiated before congestion control can function correctly. While this negotiation could be performed using signaling protocols such as SIP, RTSP, and H.323, it would be desirable to have a single standard way of negotiating these transport parameters. This is of particular importance with ECN, where sending ECN-marked packets to a non-ECN-capable receiver can cause significant congestion problems to other flows.

6.1.8. Solution Space for Congestion Control of Unreliable Flows

There are a number of options for providing end-to-end congestion control for the unicast traffic that currently uses UDP, in terms of the layer that provides the congestion control mechanism:

- congestion control above UDP
- congestion control below UDP
- congestion control at the transport layer in an alternative to UDP

Providing Congestion Control Above UDP

One possibility would be to provide congestion control at the application layer, or at some other layer above UDP. This would allow the congestion control mechanism to be closely integrated with the application itself. A key disadvantage of providing congestion control above UDP is that it places an unnecessary burden on the application-level designer, who might be just as happy to use the congestion control provided by a lower layer. If the application can rely on a lower layer that gives a choice between TCP-like or TFRC-like congestion control, and that offers ECN, then this might be highly satisfactory to many application designers. There is a second problem with providing congestion control above UDP: it would require either giving up the use of ECN or giving the application direct control over setting and reading the ECN field in the IP header. A third problem of providing congestion control above UDP is that relying on congestion control at the application level makes it somewhat easier for some users to evade end-to-end congestion control.

However, we believe that putting the congestion control at the transport level rather than at the application level makes it just slightly less likely that users will go through the trouble of modifying the code in order to avoid using end-to-end congestion control.

Providing Congestion Control Below UDP

Instead of providing congestion control above UDP, a second possibility would be to provide congestion control for unreliable applications at a layer below UDP, with applications using UDP as their transport protocol. Given that UDP does not itself provide sequence numbers or congestion feedback, there are two possible forms for this congestion feedback:

1. Feedback at the application: The application above UDP could provide sequence numbers and feedback to the sender, which would then communicate loss information to the congestion control mechanism. This is the approach currently standardized by the Congestion Manager (CM) [RFC3124].
2. Feedback at the layer below UDP: The application could use UDP, and a protocol could be implemented using a shim header between IP and UDP to provide sequence number information for data packets and return feedback to the data sender.

6.2. Stream Control Transmission Protocol

6.2.1. Introduction

Over the past few decades, packet-switched networks have merged with new technologies to facilitate more efficient communication channels within network systems. The popularity of internet protocol (IP) based networks is attributed to the emergence of the Internet and a host of highly popular network applications. IP networks and telephony based networks are converging to support the inter-working of applications and services. In addition, integrated services digital networks (ISDN), asynchronous transfer mode (ATM) networks, and mobile networks are increasingly routing their signaling traffic via IP networks. SCTP provides flexible delivery and reliable transfer within IP networks.

Large-scale interexchange carriers (IXCs) are realizing that more and more of their network traffic is data rather than voice. Therefore, the scope of present networks must be extended to accommodate application signaling and data services. The primary goal of several IXCs is to carry data and voice using the same transport, thereby reducing additional infrastructure costs.

SCTP provides numerous advantages over user datagram protocol (UDP) and transmission control protocol (TCP). For instance, SCTP combines the datagram orientation of UDP with the sequencing and reliability of TCP. Additionally, SCTP uses multi-stream, message-oriented routing in multi-homed environments [IECSCTP].

6.2.2. Overview

SCTP is a reliable transport protocol operating on top of a potentially unreliable connectionless packet service such as IP. It offers acknowledged error-free non-duplicated transfer of datagrams (messages). Detection of data corruption, loss of data and duplication of data is achieved by using checksums and sequence numbers. A selective retransmission mechanism is applied to correct loss or corruption of data.

Originally, SCTP was designed to provide a general-purpose transport protocol for message-oriented applications, as is needed for the transportation of signalling data. It has been designed by the IETF SIGTRAN working group, which has released the SCTP standard draft document (RFC2960) in October 2000. Its design includes appropriate congestion avoidance behavior and resistance to flooding and masquerade attacks.

The decisive difference to TCP is multihoming and the concept of several streams within a connection (which will be referred to as association in the rest of these documents). Where in TCP a stream is referred to as a sequence of bytes, an SCTP stream represents a sequence of messages (and these may be very short or long).

SCTP can be used as the transport protocol for applications where monitoring and detection of loss of session is required. For such applications, the SCTP path/session failure detection mechanisms, especially the heartbeat, will actively monitor the connectivity of the session [ESSSCTP].

SCTP provides applications with enhanced performance, reliability, and control functions. This protocol is essential where detection of connection failure and associated monitoring is mandatory. Furthermore, SCTP could be implemented in network systems and applications that deliver voice/data and support quality real-time services (e. g. streaming video and multimedia) [IECSCTP].

6.2.3. Stream Control Transmission Protocol

As illustrated in figure 45, the SCTP transport service layer is positioned between the SCTP user application and the network service being used. Since SCTP is based on interfacing two

SCTP endpoints, there are certain application programming interfaces (APIs) that run in between the transport service layer and SCTP user layer. In addition, each endpoint hosts multiple IP addresses.

SCTP deploys multiple paths and streams to transport messages across two endpoints. In SCTP, data is transmitted between endpoints through a connection referred to as an "association". An association begins with an "initiation" and is maintained until all data has been successfully transmitted and received. Once all data is successfully received, the association is gracefully terminated through a "shutdown".

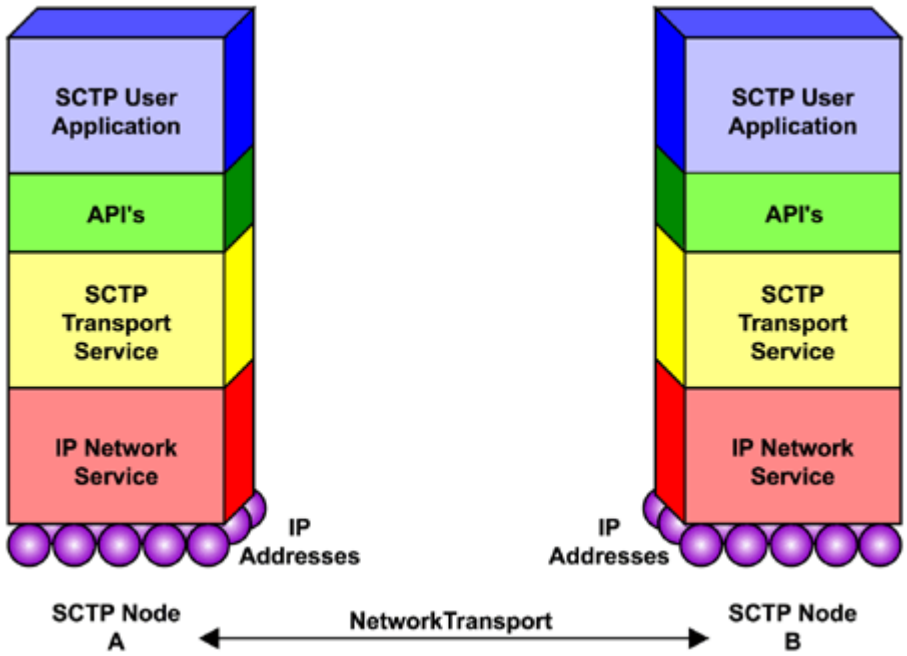


Figure 45: SCTP overview

Fundamental SCTP properties include the following:

1. **Validation and acknowledgment mechanisms:** Protects against flooding attacks and provides notification of duplicated or missing data chunks. During initiation, the validation mechanism bundles the data into a "cookie" that includes a secure hash of values and a secret key. Cookies are digitally signed with message authentication codes (MAC), which are used to prevent denial-of-service attacks.
2. **Path selection and monitoring:** Selects a "primary" data transmission path and tests the connectivity of the transmission path. SCTP packets are routed to the destination IP address of a peer endpoint through a "primary path". The primary path allows the user to determine the primary route for data flow. In addition, alternate paths exist for each IP address that the peer endpoint supports. In SCTP, a path is considered "active" when it has been acknowledged by the peer endpoint or has been used previously for SCTP packet transfer. A path is considered "inactive" if previous path transmissions have failed.
3. **Flow and congestion control:** While SCTP flow control is based on each association, congestion control is established within each transmission path. The peer endpoint assigns a receiver-window variable for flow control. The receiver-window variable alerts the endpoint of the amount of space available in the peer endpoint's inbound

buffer. SCTP deploys congestion control within each stream using a congestion-window variable. This variable limits the number of bytes that may be sent before an acknowledgement is received. A set of flow and congestion control parameters is subtly retained within the association and each transmission path.

These features give SCTP numerous advantages over TCP and UDP, as the following section elaborates [IECSCTP].

6.2.4. SCTP Advantages: Multi-Homing, Multi-Streaming and Other Features

SCTP gains advantage over TCP by the virtue of its unique features. This section explores how multi-homing, multi-streaming, and other SCTP features contribute to the SCTP advantage.

SCTP Multi-Homing

The multi-homing feature enables SCTP endpoints to support multiple IP addresses. Multi-homing protects an association from potential network failures by steering traffic to alternate IP addresses. During the initiation of an association, SCTP endpoints exchange lists of IP addresses. Therefore, each endpoint can send and receive messages from any of the IP addresses listed at the remote endpoint. For example, one of the listed IP addresses will be designated as the primary address during the initiation. If the primary address repeatedly drops chunks, however, all chunks will be transmitted to an alternate address until a connection to the primary address can be re-established.

Multi-homing is a step above conventional single-homed data exchange sessions (i. e. TCP). In single-homed environments, loss of session could be triggered by core network failures or by isolation of endpoints. Since multi-homing directs traffic on different paths to separate IP addresses, loss of session due to physical network failure is virtually non-existent in SCTP.

SCTP Multi-Streaming

The multi-streaming feature separates and transmits user data on multiple SCTP streams. These streams are capable of independent, sequenced delivery. Message loss in a particular stream will only hinder delivery within that stream. Therefore, other streams within an association are not affected.

Through multi-streaming, SCTP eliminates unnecessary blocking that often occurs in TCP transmissions. In TCP, a stream is defined as a sequence of bytes that conform to strict in-sequence delivery. In-sequence delivery results in a major drawback known as "head-of-the-queue blocking", where messages within a stream are not allowed to bypass each other. Since SCTP streams are independent, retransmitted and high-priority messages can bypass less significant messages.

SCTP Features

In the three stages of association, SCTP applies mechanisms that set it apart from TCP and UDP.

4. **Initiation features:** In contrast to the three-way handshake that occurs in TCP, SCTP uses a four-way handshake to initiate an association. This four-way handshake defends against denial-of-service attempts caused by attackers bombarding the SCTP nodes with counterfeit PDUs (protocol data units). In addition, SCTP packets that contain invalid verification tags are identified during initiation and removed from the transmission path.
5. **Data transmission features:** During data transmission, the chunk-bundling feature allows DATA chunks to be multiplexed with control chunks. The peer endpoint acknowledges the receipt of a data chunk by sending a SACK chunk. SACK chunks

contain transmission sequence numbers (TSN) that reveal any gaps in the sequence of data chunks. Within each stream, SCTP packets are also assigned stream sequence numbers (SSN). The SSN determines the sequence of data delivery within each independent stream. If the peer endpoint indicates gaps in the SSN, then the message will not be delivered until the gap is filled.

6. **Shutdown features:** The SCTP shutdown procedure has some significant advantages over TCP. For instance, a TCP connection is considered "half-open" when one endpoint continues to send data though the peer endpoint is no longer transmitting data. In contrast, SCTP implements a graceful close of an association by exchanging three messages. These messages acknowledge that both endpoints will cease in their transmissions of data [IECSCTP].

6.2.5. SCTP-based Middleware for MPI in Wide-Area Networks

MPI is a message passing library that is widely used to parallelize scientific and compute intensive programs. Why did we envision SCTP being a benefit for MPI (Message Passing Interface)? Several reasons jumped out at first. By nature, MPI passes messages, so TCP implementations require additional framing within the middleware. SCTP is message-based so the thought was that this framing could be off-loaded to the transport protocol. The same could be accomplished using UDP but unreliably [MPISCTP].

Some of SCTP's features make it attractive for use over WANs. First off, its multihoming and multistreaming features make it less susceptible to loss and latencies, traits often characteristic in WANs. Additionally, SCTP is overall more secure. For example, its four-step initiation sequence makes it avoid TCP SYN-like DoS attacks [MPISCTP].

Work in [DICKMPI] presents a user-level communication protocol where reliability was built on top of UDP and shows an improvement in performance over TCP in heavily loaded networks. One of the reasons for this improvement is the fact that their protocol does not attempt to modify its behaviour based on network congestion and maintains a constant flow regardless of the network conditions. Their protocol, in fact, captures bandwidth at the expense of other flows in the network. SCTP, on the other hand, is TCP-friendly and ensures fairness in the use of network resources.

In order to evaluate SCTP as the transport protocol for MPI, there is an iteratively extended open source implementation of MPI middleware [KAMALSCTP]. It has been shown that multistreaming and multihoming features of SCTP can be a good match for MPI and can solve many of the problems that current implementations face in WANs. This implementation is scalable and portable, it makes use of multiple streams to avoid head-of-line blocking and can achieve fast failover in case of network failure.

6.2.6. SCTP versus TCP for MPI

TCP is widely used as the underlying transport protocol in the implementation of parallel programs that use MPI. It was available in the first public domain versions of MPI (LAM [BURNSLAM] and MPICH [GROPPMPI]) for the execution of programs in local area network environments. More recently the use of MPI with TCP has been extended to computing grids [KARMPICH], wide area networks, the Internet and meta-computing environments that link together diverse, geographically distributed, computing resources. The main advantage to using an IP-based protocol (i. e. TCP/UDP) for MPI is portability and ease with which it can be used to execute MPI programs in diverse network environments [KAMVS].

One well-known problem with using TCP or UDP for MPI is the large latencies and difficulty in exploiting all of the available bandwidth. Although applications sensitive to latency suffer when run over TCP or UDP, there are latency tolerant programs such as those that are embarrassingly

parallel, or almost so, that can use an IP-based transport protocol to execute in environments like the Internet. In addition, the dynamics of TCP is an active area of research where there is interest in better models [CARTCP] and tools for instrumenting and tuning TCP connections [MATW100]. As well, TCP itself continues to evolve, especially for high performance links, with research into new variants like TCP Vegas [KAMVS]. Finally, latency hiding techniques and exploiting trade-offs between bandwidth and latency can further expand the range of MPI applications that may be suitable to execute over IP in both local and wide area networks. In the end, the ability for MPI programs to execute unchanged in almost any environment is a strong motivation for continued research in IP-based transport protocol support for MPI.

In SCTP, there is an ability to define streams that allow multiple independent message subflows inside a single association. This eliminates the head-of-line blocking that can occur in TCP-based middleware for MPI. In addition, SCTP associations and streams closely match the message-ordering semantics of MPI when messages with the same context, tag and source are used to define a stream (tag) within an association (source). SCTP includes several other mechanisms that make it an attractive target for MPI in open network environments where secure connection management and congestion control are important. It makes it possible to offload some MPI middleware functionality onto a standardized protocol that will hopefully become universally available. Although new, SCTP is currently available for all major operating systems and is part of the standard Linux kernel distribution.

Overview of Using SCTP for MPI. SCTP promises to be particularly well-suited for MPI due to its message-oriented nature and provision of multiple streams in an association. As shown in figure 46, there are some striking similarities between SCTP and MPI. Contexts in an MPI program identify a set of processes that communicate with each other, and this grouping of processes can be represented as a one-to-many socket in SCTP that establishes associations with that set of processes. SCTP can map each association to the unique rank of a process within a context and thus use an association number to determine the source of a message arriving on its socket. Each association can have multiple streams which are independently ordered and this property directly corresponds with message delivery order semantics in MPI. In MPI, messages sent with different tag/rank/context to the same receiver are allowed to overtake each other. This permits direct mapping of streams to message tags.

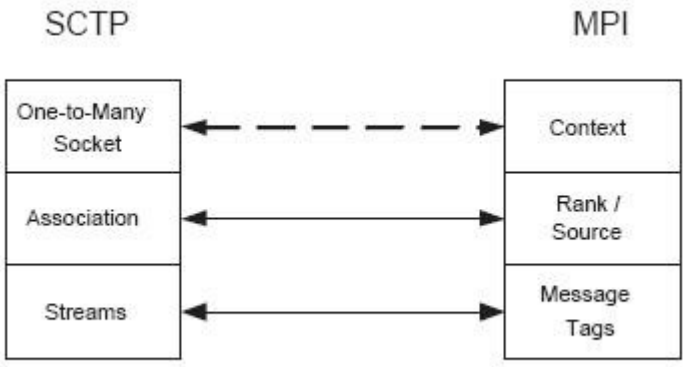


Figure 46: Similarities between the message protocol of MPI and SCTP

Related Work. The effectiveness of SCTP has been explored for several protocols in high latency, high-loss environments. Researchers have investigated the use of SCTP in FTP, HTTP,

and also over wireless and satellite networks. Using SCTP for MPI has not been investigated. There are a wide variety of projects that use TCP in an MPI environment. MPICH-G2 [KARMPICH] is a multi-protocol implementation of MPI for the Globus environment that was primarily designed to link together clusters over wide area networks. MPICH-G2 can use a custom transport for inside the cluster and a TCP connection between clusters. LAM as well has a TCP-based Globus component to support meta-computing. Work in [KAMVS] provides the opportunity to add SCTP for transport in a grid environment where we can take advantage of the improved performance in the case of loss. Furthermore, the fact that SCTP has been standardized and implementations have begun to emerge are all good indications of more wide-spread support.

6.2.7. Performance Improvement of Grid Web Services based on Multi-Homing Transport Layer

A data grid is one of the major grid applications that can be easily used in the future world. It spreads the necessary data over to one or more sites. Many web services including data grid may have a characteristic that it usually requires transferring the abundant data among many nodes in the world. To do this role well, fast transmission and self-rerouting against faults are important requirements. In [OTGMHOM] SCTP was adopted as a promising transport protocol which can support multi-homing physical connections between endpoints. Future network of SCTP/IP layered architecture can be expected to improve the transmission performance as well as the reliability. For the adaptation the Internet Web Services to the multi-homing transport environment, the description and publishing of Web Services should be changed in the interfaces and internal processing. Existing WSDL toolkits have been developed without considering the multi-homing transport environments. Work in [OTGMHOM] suggested the extended requirements of Web Service Descriptions Interfaces and extensions of UDDI function to provide the multi-homing accesses. It emulated the enhanced reliability and performance improvement of HTTP/SCTP/IP application. HTTP is widely used to exchange the messages in Web Services. When the performance of HTTP/SCTP/IP was compared with that of HTTP/TCP/IP, it was found that SCTP can transfer the data within shorter transmission time than traditional TCP in some erroneous environments by using the alternative stream paths actively. Regardless of complex reliable message exchange mechanism in Web Services, WS over SCTP can improve reliability. However, the load balancing function to speed up the transmission and reliability should be designed and developed further in the future. The tradeoffs of overhead and benefits of load balancing should be solved efficiently.

6.2.8. Conclusion

SCTP's appeal goes beyond being just a robust transport protocol. SCTP can be seamlessly introduced into present IP networks, simply as a higher layer user of IP services. The applicability and enhanced efficiency of SCTP over existing transport protocols and its conformity with existing systems may establish it as a protocol of choice within the future grid environment.

6.3. Reliable Multicast Transport with Forward Error Correction

Over the years many multicast protocols have been proposed; however, most protocols were application specific and a variety of different protocol techniques were used. This meant that no common standard protocol emerged. The lack of standardisation limited evolution of the mechanisms and denied the flexibility of re-using protocol components to design a protocol for a new application. The RMT WG [RMT WG] recently standardized a framework for the design of reliable multicast transport protocols. Their work was closely related to the reliable multicast work of the Internet Research Task Force (IRTF) [IRTF].

The RMT WG focused on the standardization of one-to-many transport of data (RFC 2887 explains the design-space). The diversity in the requirements exhibited by the various applications meant that a single protocol, which fulfilled all the requirements, was impractical. As a result a set of building blocks were developed that contain functions, which are general to different protocols. These building blocks could then be easily imported when designing protocols to do specific tasks. Figure47 depicts the general work structure of RMT WG.

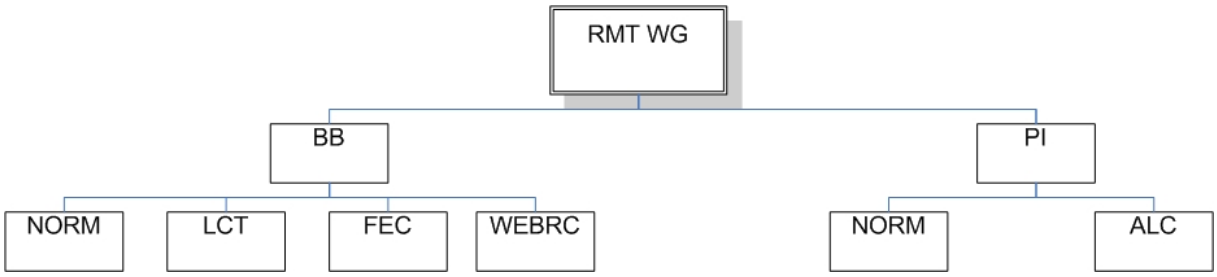


Figure 47: RMT WG work structure

Building Blocks (BB). A building block is a set of easily-separable coarse-grained modular components that are common to multiple protocols along with abstract APIs that define a building block's access methods and its arguments.

Protocol Instantiations (PI). Specifications that define the necessary gluing logic and minimal additional functionality required to realize a working protocol from one or more building blocks. These specifications will also include an abstract API that defines the interface between the protocol implementation and an application.

To provide reliable delivery, an RMT protocol is required to detect and repair packet losses. Reliability techniques can be generally divided into partial reliability and total reliability. In the case of partial reliability, the application can tolerate packet loss to a certain degree (e. g. some form of multimedia data). The opposite is true for total reliability, where the applications need to provide guaranteed delivery of data, (e. g. software distribution, stock information). Thus the underlying RMT protocol may be optimised for one specific case or may be consistent with diverse application requirements.

6.3.1. Data Carousel

Data Carousel [ACH95] is a technique by which a sender partitions the object into pieces of data and sends these as packets in continuous cycles. The receivers continue to receive the transmitted packets until all parts of the object are received.

An example should illustrate the concept: Source A is trying to send out a file by sending 4 packets. Receiver B wants to receive the file but it only received 3 packets for that file but missed the third one. He will keep on listening to source A who is sending the 4 packets again to the network right after it sent them last time. Eventually, receiver B receives the third packet. If this happens, B goes offline because the file can be assembled successfully.

An advantage of this method is that there is no need for a back channel, because there is no data flow from the receiver to the sender. However a limitation of this technique is that where a receiver undergoes a packet loss in one round of the transmission, it must wait an entire round before it has a chance to receive that packet again, which is also true for late joiners to the group. This technique makes inefficient use of capacity as the sender continually cycles through and transmits the packet until no receiver is missing a packet.

6.3.2. Forward Error Correction

Packet level Forward Error Correction (FEC) is a scheme where parity packets are transmitted along with the original stream in order to recover any corrupted/lost packets [LUB02]. FEC codes, in general, are able to overcome both erasures (packet loss) and bit-level corruption. Most link and transport protocols have a built-in CRC that checks for packet corruption and discards a packet if corrupted. Therefore the primary role of FEC in IP multicasting is to provide erasure correction. Since receivers are able to reconstruct lost packets without the sender needing to retransmit them, the use of packet level FEC is found to be scalable. An immediate benefit of using packet level FEC for multicast is that multicasting a single parity packet can repair loss of different packets at different receivers of the same session.

Packet level FEC techniques may be divided into two categories: Proactive FEC and reactive FEC [LI99]. In proactive FEC, the sender determines a priori the amount of redundant packets to be sent in order to repair lost/erased packets. Protocols that use proactive FEC are Digital Fountain (DF) protocol [BYE98] and FCast [GEM99]. In the case of reactive FEC, the sender uses feedback from the receivers to compute how many packets were lost in a particular round of transmission and then send the additional number of encoded retransmissions needed for complete recovery of the data. Protocols that use reactive FEC are MFTP/EC [ROB01] and Reliable Multicast data Distribution Protocol (RMDP)[RIZ98].

6.3.3. ACK/NACK Based Models

With this type of technique the receiver(s) uses a back channel, either via the terrestrial infrastructure (e. g. the Internet) or via satellite (e. g. DVB-RCS) to either acknowledge the packets (ACK packets) that it has received, or request the packets that was lost (NACK packets).

When considering the design of an RMT protocol over a satellite network, ACK based models do not scale well as the sender needs to keep track of the number of receivers and their state to provide flow and congestion control. The opposite is however true for a NACK based scheme. Furthermore, only a single NACK is required to indicate a missing packet for a number of receivers. Many NACK-based protocols provide NACK suppression techniques to prevent NACK implosion and improve scalability. Some protocols that implement NACK suppression techniques include Scalable Reliable Multicast (SRM) [FLO97], NACK Oriented Reliable Multicast Protocol (NORM) [RFC3940] etc. The use of such feedback suppression techniques is heavily dependent on the network infrastructure where the protocol is deployed. Intermediate systems such as routers (for terrestrial networks) or gateway nodes (for satellite networks), or dedicated receivers acting as local group controllers (LGC) can ideally provide such feedback suppression.

6.3.4. Combination of Reliability Techniques

Some reliable multicast protocols use a combination of basic reliable techniques, for example FCast uses FEC with carousel, SRM and Negative-acknowledgement Oriented Reliable Multicast protocol (NORM) uses FEC with NACK to provide reliability. An RMT protocol using FEC may either transmit multicast data to a single group address, or choose to spread the multicast data over several groups, e. g. Asynchronous Layered Coding (ALC) [RFC3450]. In this approach the sender uses a packet level FEC scheme to encode the original k data packets to generate n encoding packets (n>k), which are then striped across multiple multicast groups. The receiver(s) join(s)/leave(s) one or more of these multicast groups depending on their network capacity. The two main issues concerning this type of scheme are 1) creation of logical layers 2) choosing a transfer rate scheme for a layered multicast session. In order to provide efficient layering and transfer rates, some form of feedback about the current state of the network (e. g. congestion level, fade margin etc.) is required. An application (using

unidirectional protocol) or a bidirectional transport protocol can obtain this information from either the use of statistical data (e. g. network traffic analysis, meteorological data etc.) or various out-band techniques (e. g. negative acknowledgements NACK).

6.3.5. File Delivery Over Unidirectional Transport

FLUTE [RFC3926] is a unidirectional RMT protocol, designed using the components of the BB architecture, to deliver files over the Internet. The FLUTE protocol, as shown in figure 48, is build on ALC PI [RFC3450] of the Layered Coding Transport (LCT) BB [RFC3451]. FLUTE provides a mechanism for signalling and mapping the properties of files to the concepts of ALC in a way that allows receivers to assign those parameters for received objects [RFC3926]. It is designed to work with both ASM [RFC 1112] and SSM [RFC3569] multicast models. FLUTE transmits two types of objects:

- File Delivery Table (FDT) and
- the file(s).

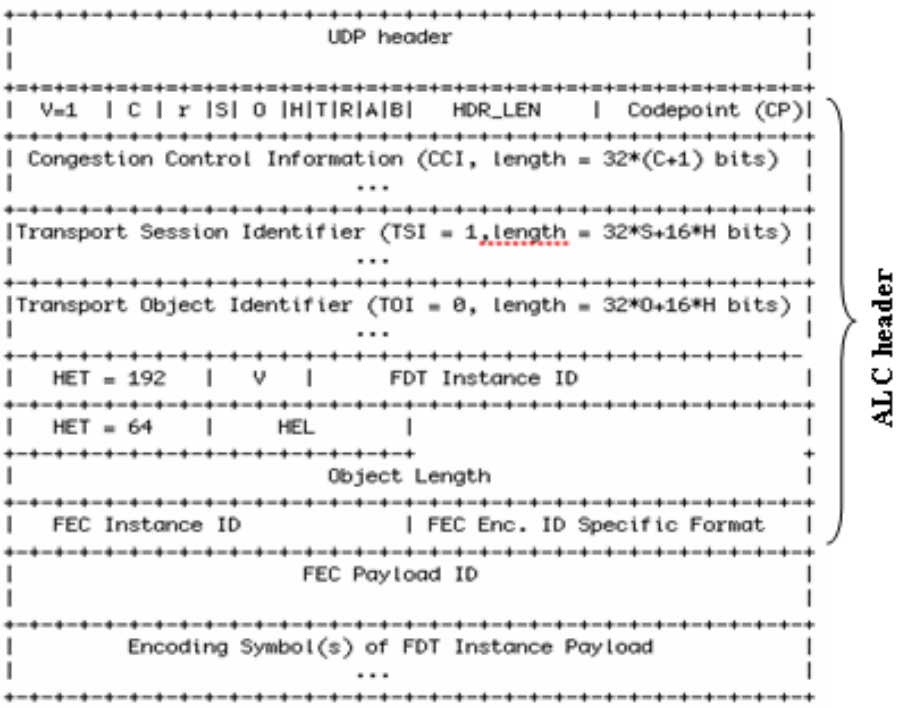


Figure 48: FLUTE header

The FDT provides the receiver with a means to describe the various attributes associated to the file that are to be delivered within a session. The FDT also provides attributes regarding the delivery of the file.

Attributes specific to the delivery of file:

- TOI value that represents the file
- FEC instance ID
- FEC object transmission information
- size of the transport object carrying the file
- aggregate rate of sending packets to all channels

Attributes specific to the file:

- name, identification and location of the file (specified by the URI)
- MIME media type of file
- size of file
- encoding of file
- message digest of file

During a FLUTE session an FDT is send as an FDT instance. A reserved TOI value of 0 is used to indicate that the packet is an FDT instance. A receiver must receive an FDT instance before it is able to recover the file. The FDT instance consists of two parts: FDT header and FDT instance payload.

The FDT instance header forms part of the LCT header extension (EXT_FDT). Any ALC/LCT packet carrying FDT instance must include EXT_FDT. The header extension type (HET) has a value of 192. The 4-bit field 'V' stands for the current version of FLUTE that is used, i. e. version 1. The header consists of a FDT instance ID that uniquely identifies FDT instances within a file delivery session. The FDT instance header is carried in each ALC packet carrying an FDT instance. The FDT instance payload follows the FEC Payload ID of the LCT header. The FDT instance payload consists of one or more file description entries, which are composed and structured in accordance to an XML scheme.

By use of the multiple rate congestion control BB, the encoded packets which are stripped on to different layers (each layer is a multicast group), can be transmitted at varying rates. A receiver joins/leaves a multicast group(s) depending on their network capacity. This makes FLUTE more scalable.

At the start of the session the sender transmits an FDT instance describing the various attributes associated with the delivery of the file (e. g. session id, object id, CC rates etc.) once the receiver has joined in a FLUTE multicast session, it should first receive an FDT to choose the parameters required to retrieve an object as well as the attributes associated to the file itself (e. g. name, location of file etc.). The FLUTE sender encodes the data and starts transmitting the object based on the session ID (TSI) and the object ID (TOI) values indicated in the FDT. The commonly used FEC schemes are Low Density Parity Check (LDPC) [VIN03] and Reed Solomon FEC [RIZ97]. FLUTE uses a layered-proactive FEC scheme to provide scalability and reliability. The receiver then allocates the resources based on some predefined techniques (e. g. blocking algorithm [RFC3926]). Once the receiver obtains enough packets to decode the object(s) it leaves the session. The type of decoder to choose depends on the FEC Encoding ID field specified by the sender either in the FDT or in the ALC extension header (EXT_FTI).

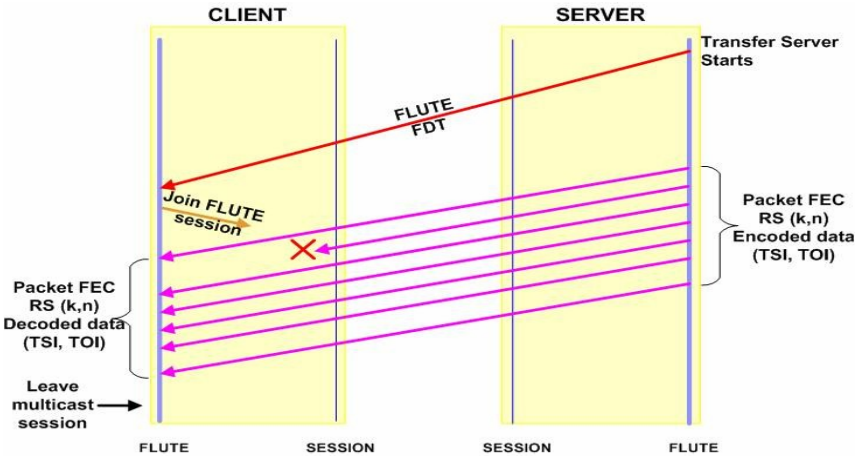


Figure 49: FLUTE transition diagram

6.4. User-Controlled Light Paths

6.4.1. General Protocol Information

User-Controlled LightPaths (UCLP) is a network configuration and provisioning tool built around grid technologies and web services. It would be even more appropriate to think of UCLP as a toolset, rather than a tool, used to concatenate cross-connects, routers and switches to produce a wide-area network which is under the users' control [FIG06]. UCLP was started by Canarie, the Canadian National Research and Education Network (NREN), but is now being developed by a larger consortium of partners which includes organizations from outside Canada as well.

UCLP has been built to enable end-users, be it humans or advanced applications, to set up Virtual Private Networks (VPNs) without the intervention of network managers. The basic idea behind UCLP is that networking links and equipment can be seen as objects and services, in the context of Service-Oriented Architectures (SOA). This way, they can easily interact with other entities and be used to construct multi-domain lightpaths by concatenating wavelengths on consecutive links. In the UCLP terminology, the various web services are assembled into Articulated Private Networks (APNs), that is, network virtualizations running their own protocols and services. Using these ideas, applications are extended into the network, establishing application-specific VPNs, and the network becomes a resource like any other, such as storage and CPUs. The users can control the resource (network) just like any other, as long as they are properly authorized for that.

On a software level, the project eventually wishes to build a set of services to expose different kinds of networking equipment, which can then be orchestrated into BPEL-powered workflows and integrated into other larger workflows/applications.

The software developed as part of UCLP (now in version 2), is written in Java using various established toolkits such as Axis, Eclipse, etc.

6.4.2. Architectural Details

A high-level architecture of UCLP is provided in figure 50 [FIG06].

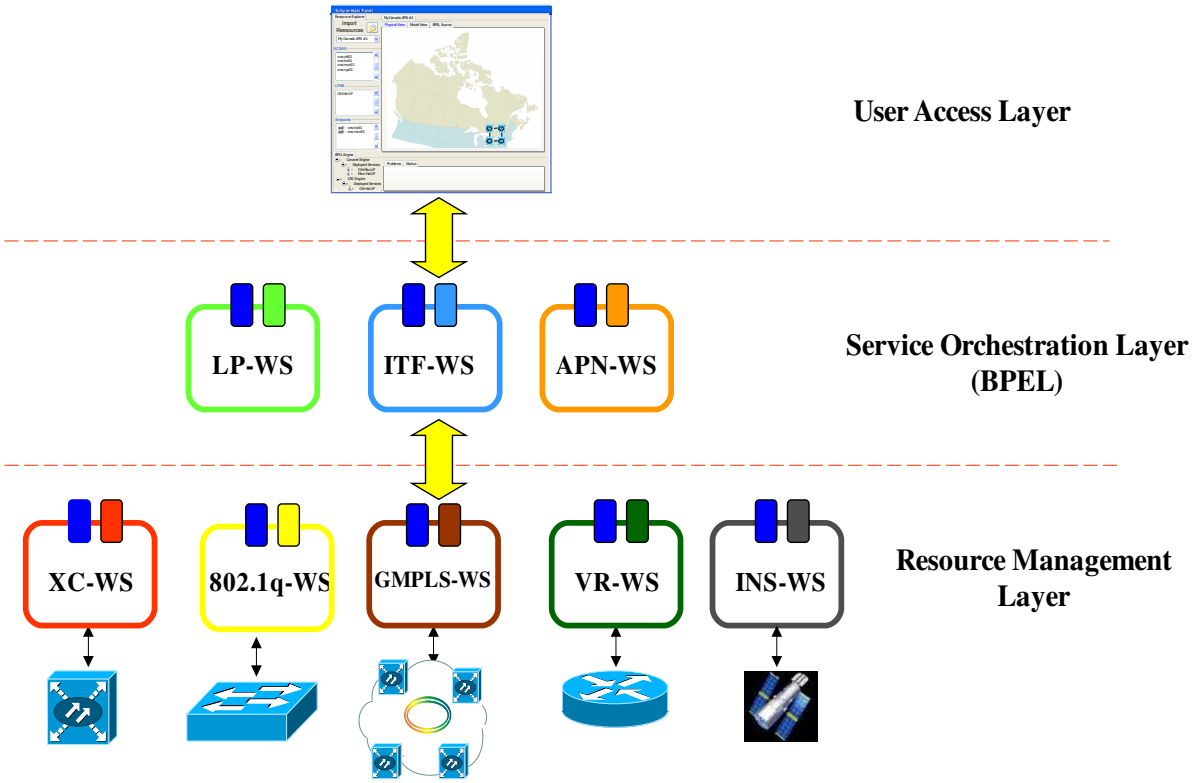


Figure 50: UCLP architecture overview

At the Resource Management Layer (RML) we can detect the Instrument Web Service (INS-WS) which is deployed at the consumer's end, and the Network Element Web Services (NE-WS) which are deployed on the network carrier's end. Depending on the type of equipment being controlled, different NE-WSs are being made available:

- XC-WS: cross-connect web service, to expose SONET, SDH, fibre and lambda cross connects
- 801.1q-WS: used to expose VLAN-enabled Ethernet switches
- MPLS-WS/GMPLS-WS: web services to expose the functionality of MPLS and GMPLS clouds
- VR-WS: the router WS is the abstraction of a L3 router

On the Service Orchestration Layer (SOL), the following services are made available:

- **LP-WS: LightPath Web Service.** This is an abstraction of a light path (Layer-1 VPN), consisting of one or more (interconnected) links in the form of a workflow script assembling a number of distinct web services. The workflow script exists on the same server where the WS resources are located.
- **ITF-WS: InterFace Web Service.** Represents a single resource on a Network Element.
- **APN resource list.** A pointer to a set of web services that may constitute all or part of an APN. This set has a common set of permissions.

An Articulated Private Network represents a user-built network, as it has been assembled by the various participating resources (from the APN resource list). In essence, this is materialized as a BPEL document which connects plenty of different resources in a workflow. The APN represented is a single network configuration with fixed links and bandwidth – these cannot be changed after the creation of the APN. The essential difference between a LP-WS and an APN

is that the latter (script) runs on the client's machine, while the former runs on the server where the web service(s) reside. The obvious direct implication is that the user can terminate the script at any time and re-assemble the same or other web services into a different APN [BSA].

The resources which are exposed over the previously mentioned web services are expected to be advertised/discovered in an automated way, which should also provide for regular updates. On the same time, an administrator should be able to insert rules and exceptions, in order to have full control over circuit advertisements and utilization.

Looking at how UCLP fits within the more general SOA context, figure 51 provides an example setup [BSA]:

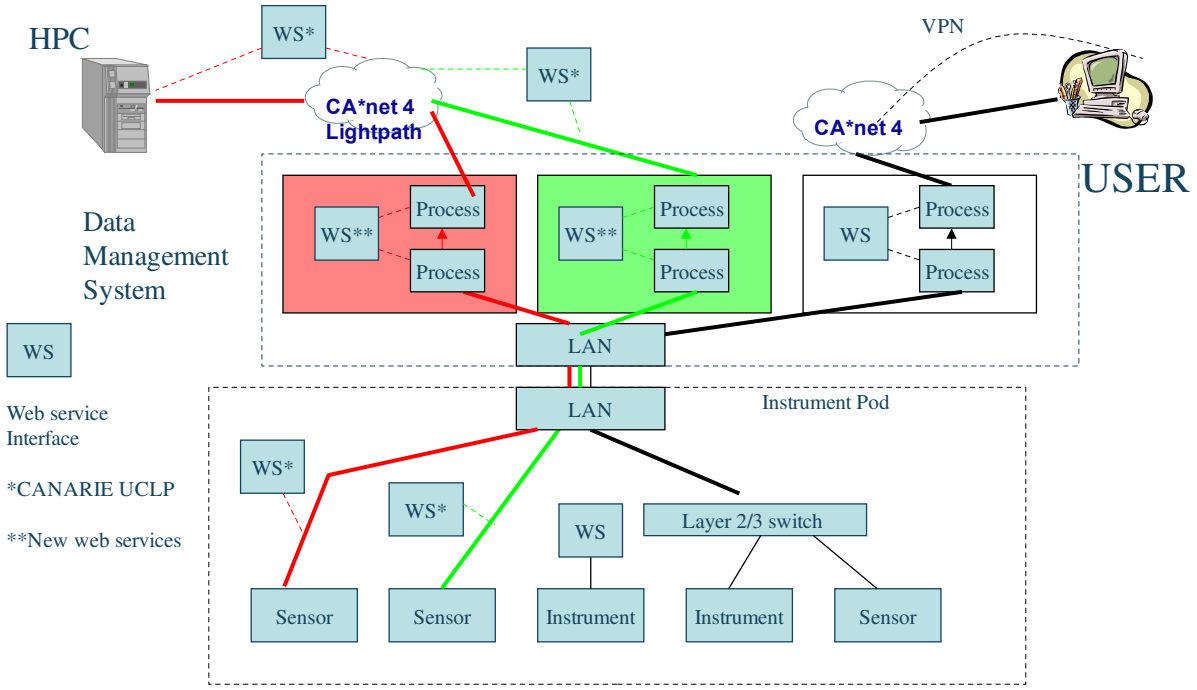


Figure 51: UCLP in a SOA environment

The general idea is that utilizing an instrument over the network has the same principles with controlling/utilizing a network device, with a control and a data plane to consider. Thus, the abstractions applied to networking equipment can be adapted to instruments and sensors, to make them part of APNs. As previously mentioned, this is ongoing work at the moment, without any concrete results so far.

From a security point of view, UCLP[v2] recommends SAML as the standard method for authorization, alongside x.509 certificates for authentication. Three basic roles are defined:

- The **Physical Network (PN) Admin** is essentially provisioning the network, by constructing Lightpaths and APN resources.
- The **APN Admin** who is receiving resources from PN or other APN admins, and is responsible for creating APN configurations for the users. Additionally, an APN admin can give or sublease his/her resources to other APN admins.
- The **User** who can use configurations (APNs) created by APN admins.

6.4.3. APN Setup Procedure

A graphical user interface is provided as part of UCLPv2. In it, a resource list can be found, alongside two composition windows for constructing the real workflows. Resources may be

either physical resources such as links, or APN resources such as LP-WS, XC-WS, ITF-WS etc. Initially, the provider (carrier) needs to construct all necessary LP-WSs, in the form of workflows as mentioned earlier. This allows the carrier to have control over the paths that the users may actually use to construct their application-specific VPNs. As soon as the LightPath objects are ready in the form of web services, they are populated to the resource list with all the other resources which are to be made available to the end user. The user would then be able to construct more complex workflows (aggregate services) using this primitives which have already been made available to her. Canarie provides detailed instructions on how to build an APN at the UCLP user guide [UCLPGUIDE].

6.4.4. UCLP and Scientific Instrumentation

By design, UCLP has been built with the scientific user in mind. As a matter of fact, the typical usage example is that of an instrument at one end of a path, sending data at very demanding rates over a WAN to an HPC unit which resides at the other end. Although UCLP is not about providing guarantees for latency, jitter or other similar characteristics per se, proper building of an APN and carefully thought-out mechanisms for sanity in reservations can ensure "by proxy" the desired thresholds for related QoS variables. The greatest strength of UCLP lies into the fact that it has been built with SOA and the grid in mind. The fact that resources are exposed as web services changes our view of the network, from a stack of layers (protocols) with restrictions on how each layer communicates with each other, to a pool of resources which can be assembled together in ad-hoc manners and create utility networks per application. The dynamic nature of this approach is important for users who are in need of bandwidth on demand (BOD), while on the same time the workflow-oriented mindset fits with the typical scientific application. Reuse and interoperability are in the core of SOA, and UCLP follows this path to minimize migration costs (in time and resources) from one network user (application) to another.

Additionally to provisioning the network though, UCLP considers instruments/devices to be part of its APNs, resources to be made available over web services and used in a SOA environment. In this regard, UCLP can be considered as a more generic middleware effort which encompasses the use of instruments in this manner. Although currently this functionality has not been implemented, it foreseen for UCLP to perform this kind of integration. In the UCLP roadmap document [UCLPRDM], examples for a proxy of an instrument/sensor are provided (figure 52).

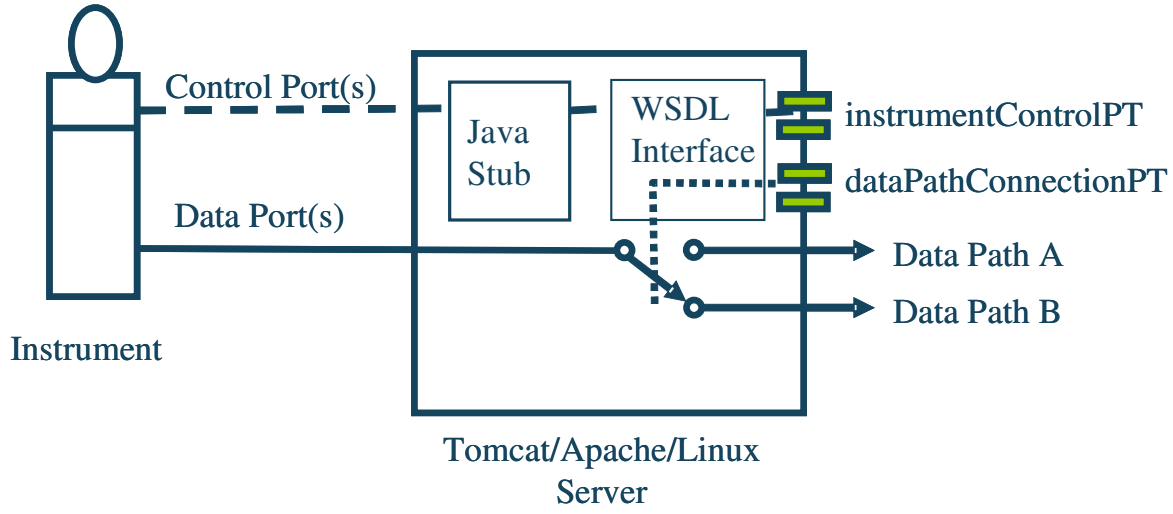


Figure 52: UCLP instrument proxy service

6.5. Globus Teleoperations Control Protocol

6.5.1. Introduction: NEESgrid and NTCP

The NTCP protocol has been designed and implemented in the Network for Earthquake Engineering Simulation (NEES) project, which aims to advance collaborative earthquake engineering research in the United States by improving facilities for physical and computational earthquake simulations and encouraging the sharing of data, facilities, and computational models.

Traditionally, earthquake engineers have simulated the effects of ground motion on structures using one of two basic approaches: using computational simulations or using physical simulations. Recently, earthquake engineers have begun doing hybrid experiments: Coupled computational and physical simulations in which one part of a structure is modeled computationally and another part is modeled as a physical experiment. The computational and physical simulations are run simultaneously; the results of both the computational and physical component for one time-step are used to determine the inputs to each for the next time-step. Hybrid experiments have generally been tightly coupled, with the computational system and the physical control system communicating via a shared-memory backbone. In other words tightly-coupled hybrid experiments combine the two approaches: One part of a structure is modeled computationally and another part as a physical experiment, and the computation and the physical experiment's control system communicate and influence each other's behavior over the course of the experiment. Hybrid experiments are relatively straightforward to perform when they involve a single physical experiment, as the computational simulation and physical apparatus can be co-located.

However, for some earthquake engineering problems, it would be desirable to construct a hybrid experiment that involves more than one physical experiment — for example, an experiment involving a large geotechnical centrifuge to model soil motion and a large shake table to model the motion of a structure above ground. Physical experiments (and the physical components of hybrid experiments) are often performed at a large scale — specimens weighing 50 tons are not uncommon — and require specialized facilities.

Thus, such multi-component hybrid experiments will typically require coupling over multiple geographically distributed sites. A large-scale distributed hybrid experiment is fundamentally about sharing heterogeneous resources (simulation, experimental apparatus), each owned and controlled by a different institution, and integrating them so as to enable a collaborative experiment to take place. Thus, a distributed hybrid experiment maps well into the concept of a virtual organization and would appear ideally suited to the application of grid technology. Recognizing this, we have created a grid based framework for conducting distributed hybrid experiments. Building on mechanisms provided by the Globus Toolkit's implementation of the Open Grid Services Infrastructure (OGSI) specification, domain-specific grid services, platform specific interfaces, and user interface tools have been created that make it possible to construct, perform and monitor distributed hybrid earthquake engineering experiments conducted across geographically and organizationally distributed sites with heterogeneous equipment and policy. This framework is called NEESgrid.

The NEES Teleoperations Control Protocol (NTCP), developed specifically for NEESgrid at the Information Science Institute at the University of Southern California, provides a common protocol that can be used to remotely control physical experiments or computational simulations. By giving users a standard interface to local equipment and simulation capabilities, it provides the transmission standards for the NEESgrid that builds on the TCP/IP protocol, which connects hosts on the Internet.

Some of the design goals of NTCP are: To support a common protocol for both physical experiments and computational simulations; to support fault recovery to the greatest extent possible; to allow for separate negotiation and execution phases; to allow a client application to verify that the actions proposed for a time-step are acceptable to all sites involved, before actually sending a request to take any physical action.

6.5.2. GTCP Overview

Globus Teleoperations Control Protocol (GTCP) is a service interface for telecontrol. It is the WSRF (Web Service Resource Framework) version of the NEESgrid Teleoperations Control Protocol (NTCP), which is used to control heterogeneous physical and computational simulations coupled in geographically-distributed earthquake engineering experiments. It has also been used to control data acquisition systems (triggering the collection of data) and high-resolution cameras (triggering image acquisition) during earthquake engineering experiments, and, to a lesser extent, to control the positioning, focal length, etc. of electron microscopes in a neuroscience application.

GTCP exposes two interfaces: A WSRF compliant service interface used by clients to control remote instruments and simulations, and a "plugin" interface to facilitate integrating new backends (physical or computational simulation platforms) to the GTCP server. The plugin interface is a Java interface definition that includes methods for each platform-specific action; a new platform is integrated by writing a class that implements this interface definition.

GTCP is a new component of Globus Toolkit and has been introduced in version 4.0. All configuring and running parameters are available on the Globus website.

6.6. Network Performance Measurement

IP network performance measurement provides the means to gain insight into the network operation state. It is useful for optimising the network because it can provide the feedback for the engineer to adaptively optimise network performance in response to events and stimuli originating within and outside the network. It is essential to determine the quality of network services and to evaluate the effectiveness of traffic engineering policies. And experience indicates that measurement is most effective when acquired and applied systematically [RFC 3272].

To deploy the measurement on a network, one has to address the following questions:

- Why is measurement needed in this particular context?
- What parameters are to be measured?
- How should the measurement be accomplished?
- Where and when should the measurement be perfomed?
- How frequently should the monitored variables bemeasured?
- What level of measurement accuracy and reliability is desirable?
- What level of measurement accuracy and reliability is realistically attainable?
- To what extent can the measurement system permissibly interfere with the monitored network components and variables?
- What is the acceptable cost of measurement?

The answers to these questions will determine what measurement tools and methodologies are suitable for the particular engineering context.

Measurement can occur at different levels of abstraction. For example, measurement can be used to derive packet level characteristics, flow level characteristics, user or customer level characteristics, traffic aggregate characteristics, component level characteristics, and network wide characteristics [RFC 3272]. Network wide characteristics give an overall view of the whole network. This includes information on network bottlenecks, overload of the network, and so on.

6.6.1. Performance Measurement Methodologies

Lots of research has been done to develop measurement methodologies, e. g. using log files and capturing packet form the Internet using software or hardware. The measurement methodologies can be divided into two main groups: passive approach and active approach. Both have their advantages and should be regarded as complementary, in fact they can be used together.

Passive Measurement

The passive measurement approach implies to use devices, to monitor the traffic when it passes by. These devices could be some specific tools such as sniffer hardware, or they can be pure software built into some network equipments such as routers, switches and end node hosts. Examples of such built-in techniques include Remote Monitoring (RMON) [RFC 2819], which enables various network monitors and consoles to exchange the monitored network data using a kind of database named Management Information Base (MIB), and Simple Network Management Protocol (SNMP) [RFC 3416, RFC 3418], which is one network management protocol by using specific messages and MIB, capable devices. The passive measurement will not create or modify traffic on the network. This is in contrast to active measurement, in which specific test packets are introduced into the network, and these packets are timed as they travel through the network being measured. The basic principle of the passive measurement is shown in figure 53.

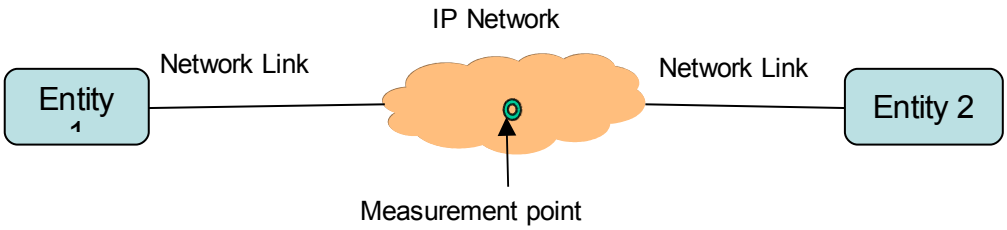


Figure 53: Basic principle of passive measurement

Figure 53 shows that two entities are connected via an IP network through network links. A measurement point for measuring traffic lies on the path between these two entities. The "entity" in figure 53 can represent various network situations: For example, the two entities could be two end users, or an internal organization network and the external Internet, or two routers, and so on. They could also be two parts of the Internet, e. g. the backbone and one edge network.

Passive measurement can provide a set of detailed information of the interested traffic at one point in the network that is being measured. Examples of the information passive measurements can provide are:

- traffic/protocol mixes
- accurate bit or packet rates
- packet timing/inter-arrival timing

It can also be deployed as a network application debugging method by capturing the entire packet contents. All of these advantages make the passive measurement valuable in network troubleshooting, single node behaviour study, source modelling and capacity management. It requires collecting the data and the traps and alarms all generated network traffic, which can be substantial. Moreover, the gathered data can be substantial especially if one is trying to capture information on all packets flowing in a network.

There are two major categories that passive measurement systems can fall into. The first is online processing that deals with realtime data. For instance, observing packet number and type, throughput in a period of time and so on. It is useful to monitor the instantaneous network status and bandwidth utilization situation. The second category is offline processing that is made possible by saving the captured packets as well as additional information (such as arrival time) into trace files. These trace files are processed and analysed after the measurement.

Online processing requires powerful devices to capture the packet while doing extra calculation when monitoring a high-speed network. For example, one wants to know the instantaneous throughput using a graph drawing based on the captured traffic flow. Normally, the monitor device should save captured packets in a buffer and update the throughput graph periodically. When the link is heavily loaded, say up to hundreds megabytes per second or even more, it is highly possible that the monitoring device might drop packets passing by while it has to calculate the throughput and update the graphic user interface. It could be worse when users want to know many parameters at once. Therefore, analyzing the traces offline is easier though it costs more storage space to save packets and relevant additional information. One can spend more time to derive more details from these saved traces, such as interarrival time, packet lossage rate, flow distributions and so on. People have tried to save partial packets instead of its entire contents in order to save both processing time and storage space. One very common subset of the data that is saved is the IP header and transport layer headers. One other common subsection of data captured is the data link layer headers. This is used primarily in ATM networks, but this type of capture has limited use for IP level analysis. The IP header provides information on the source of the datagram, the destination of the datagram, the length of the datagram and which transport protocol is carried in the payload. The transport layer can give an indication of what type of traffic was contained within the packet.

Header traces are commonly used for both of the online and offline passive measurement configurations. Full capture of all packet data on a link is normally restricted to the online process situation. The data rates created by a single computer are low when compared to backbones and gateways. Full capture allows complete analysis of the actual data passing on the network, which could be used for debugging purposes and also allow later "playback" of the entire data stream.

Active Measurements

The active approach relies on the capability to inject test packets into the network or send packets to servers and applications. This increases the network traffic. The volume and other parameters of the introduced traffic are fully adjustable, however small traffic volumes are enough to obtain meaningful measurements.

Active measurement provides very little information about a single point of a network. Instead they provide a representation of the characteristics of the entire network path between two hosts. Active systems can provide such indications of a networks performance as:

- packet round trip time (RTT),
- average packet loss and
- connection bandwidth.

Some active systems can also give indications of the following:

- asymmetric delay times and
- alterations in routing paths between hosts.

The active approach provides explicit control on the generation of packets for measurement scenarios. This includes control on the nature of traffic generation, the sampling techniques, the timing, frequency, scheduling, packet sizes and types (to emulate various applications), statistical quality, the path and function chosen to be monitored. Being active implies testing what you want, when you need it. Emulation of scenarios is easy and checking if QoS or Service Level Agreements (SLA) are met is relatively straightforward.

Passive measurement and active measurement can be used together. For example, the active measurement probe can schedule passive measurements of appropriate metrics at appropriate points along the path, while the active measurements are being made. When the active measurement is completed then the appropriate passive measurements can be paused thus reducing the gathering of unnecessary data. By comparing and contrasting the active and passive measurements, the co-validity of the different measurements can be verified, and much more detailed information on carefully specified/scheduled phenomena is made available. It is very common that one may need both measurement results to come to the final conclusion.

These measurement methodologies show how to measure the IP network traffic. Performance parameters will tell what to measure.

6.6.2. One-to-one Performance Parameters

To determine measurement parameters is the most important task before launching the measurement. It is important to decide on measurement tools, methodologies and accuracy. The Internet Engineering Task Force (IETF) IP Performance Metrics (IPPM) working group developed a set of standard metrics that can be applied to the quality, performance, and reliability of internet data delivery services. Another working group named Benchmarking Methodology (BMWG) made a series of recommendations concerning the measurement of the performance characteristics of various internetworking technologies, which includes terminology, identifying a set of metrics that aid in the description of traffic characteristics, and methodology, required to collect said metrics. Additionally, the ITU-T Working Group T1A1.3 made similar network performance parameter definition [ITUY1540].

The IPPM developed a set of measurement parameters as well as the corresponding measurement methodologies with the cooperation with other relative working group such as BMWG, TEWG, ITU-T SG 12 and SG 13. Those parameters include:

- **Connectivity:** If a packet transmitted from source A to destination B at time T can arrive B, it is said that A has the connectivity to B at time T.
- **One-way delay:** The difference between the time when the source sends out the first bit of the packet and the time when the destination receives the last bit of the packet.
- **One-way loss:** If a packet transmitted from source A to destination B cannot arrive B in a certain time threshold, it is said that this packet is lost.
- **Round-trip delay:** The sum of the times needed for a test packet travel from source A to destination B and from B back to A.
- **One-way delay variation:** The difference of the one-way delays of a selected pair of packets in the stream going from source A to destination B.
- **Loss patterns:** The packet loss distribution.

- **Bulk transport capacity:** The expected long term average data rate (bits per second) of a single ideal TCP implementation over the path in question.

The IPPM defined a general framework [RFC 2330] for particular parameter metrics that can be deployed to gain common understanding by internet users and internet providers of the performance and reliability both of end-to-end paths through the Internet and of specific "IP clouds" that comprise portions of those paths. The term "metric" is defined as a carefully specified quantity that is relative to the Internet performance and reliability one is interested in. It recommends defining particular metrics under some criteria and disciplines in order to allow people to speak clearly about Internet traffic performance. In several IETF meetings criteria for these metrics have been specified as follows [RFC 2330]:

- These metrics must be concrete and well defined.
- A methodology for a metric should have the property that it is repeatable: if the methodology is used multiple times under identical conditions, it should result in consistent measurements.
- The metrics must exhibit no bias for IP clouds implemented with identical technology.
- The metrics must exhibit understood and fair bias for IP clouds implemented with non-identical technology.
- The metrics must be useful to users and providers in understanding the performance they experience or provide.
- The metrics must avoid inducing artificial performance goals.

Each parameter metric will be defined in terms of standard units of measurement. The international metric system will be used, with the following points specifically noted [RFC 2330]:

- When a unit is expressed in simple meters (for distance/length) or seconds (for duration), appropriate related units based on thousands or thousandths of accepted units are acceptable. Thus, distances expressed in kilometres (km), durations expressed in milliseconds (ms), or microseconds (us) are allowed, but not centimetres (because the prefix is not in terms of thousands or thousandths).
- When a unit is expressed in a combination of units, appropriate related units based on thousands/thousandths of acceptable units are acceptable, but all such thousands/thousandths must be grouped at the beginning. Thus, kilometres per second (km/s) is allowed, but meters per millisecond is not.
- The unit of information is the bit.
- When metric prefixes are used with bits or with combinations including bits, those prefixes will have their metric meaning (related to decimal 1000), and not the meaning conventional with computer storage (related to decimal 1024). In any RFC that defines a metric whose units include bits, this convention will be followed and will be repeated to ensure clarity for the reader.

IPPM gives 6 sets of standardized metrics for the following parameters under the above criteria:

- metrics for measuring connectivity [RFC 2678]
- one-way delay metric [RFC 2679]
- one-way packet loss metric [RFC 2680]
- round-trip delay [RFC 2681]
- one-way loss pattern [RFC 3357]

- packet delay variation [RFC 3393]

Each of these metrics is normally defined with three sections including metric name, metric parameters and metric units. The metric name contains basic information of the measurement such as packet type (unidirectional or bidirectional) and parameter name. The metric parameter section defines what traffic parameters should be recorded in the metric that can be used for further analysis. The metric unit part describes the unit type of the metric. For instance, the one-way delay metric is named "Type-P-One-way-Delay" that means packets measured in this metric are all type P packet where P could be protocols such as TCP, UDP and ICMP. Its metric parameters are Src, the IP address of the packet source, Dst, the IP address of the packet destination, and T, the time the source sent out the first bit of the type P packet.

Corresponding to each metric, at least one measurement methodology is defined to acquire data from the network. These methodologies should have the property that it is repeatable: If the methodology is used multiple times under identical conditions, it should result in consistent measurements or continuity results with small variations.

The following traffic parameters and their measurement methodologies were defined by IETF for the purpose of network performance and reliability analysis. They are vital for the network evaluation, especially QoS evaluation:

One-way delay. The definition of one-way delay of a packet is the difference between the time when the source sends out the first bit of the packet and the time when the destination receives the last bit of the packet (whenever a time, i. e. a moment in history, is mentioned in this document, it is understood to be measured in seconds and fractions thereof) [RFC 2679].

Packet delay variation ("jitter"). The one-way delay variation of a pair of packets within a stream of packets is defined as the difference of the one-way delays of a selected pair of packets in the stream going from measurement point MP1 to measurement point MP2 [RFC 3393].

Round-trip delay. The round-trip delay is defined as the sum of the times needed for a test packet travel from the source to the destination and from the destination back to the source [RFC 2681].

One-way packet loss. If a test packet does not arrive at its destination in a time threshold, it is defined lost [RFC 2680].

7. Grid Infrastructures for Remote Instrumentation

7.1. Globus Toolkit

Emerging high-performance applications would be able to exploit diverse, geographically distributed resources. These applications typically use high-speed networks to integrate supercomputers, large databases, archival storage devices, advanced visualization devices, and/or scientific instruments thus building up networked virtual supercomputers or meta-computers. While the physical infrastructure to build such systems is becoming widespread, the heterogeneous and dynamic nature of the meta-computing environment poses new challenges for developers of system software, parallel tools, and applications. Globus toolkit [FOSTER1] is an open source software toolkit used for building grids and it is being developed by the Globus Alliance and many others all over the world [FOSTER2]. A growing number of projects and companies are using the Globus Toolkit to unlock the potential of grids for their cause. The Globus system is intended to achieve a vertically integrated treatment of application, middleware, and network. A low-level toolkit provides basic mechanisms such as communication, authentication, network information, and data access. These mechanisms are used to construct various higher-level meta-computing services, such as parallel programming tools and schedulers. The I-WAY experiment [DEFANTI] identified four significant application classes.

1. **Desktop supercomputing.** These applications couple high-end graphics capabilities with remote supercomputers and/or databases. This coupling connects remote users with computing capabilities, while at the same time achieving distance independence among resources, developers, and users.
2. **Smart instruments.** These applications connect users to instruments such as microscopes, telescopes, or satellite downlinks that are themselves coupled with remote supercomputers. This computational enhancement can enable both quasi real-time processing of instrument output and interactive steering.
3. **Collaborative environments.** A third set of applications couple multiple virtual environments so that users at different locations can interact with each other and with supercomputer simulations.
4. **Distributed supercomputing.** These applications couple multiple computers to tackle problems that are too large for a single computer or that can benefit from executing different problem components on different computer architectures.

We can distinguish scheduled and unscheduled modes of operation. In scheduled mode, resources, once acquired, are dedicated to an application. In unscheduled mode, applications use otherwise idle resources that may be reclaimed if needed; Condor [LITZKOW] is one system that supports this mode of operation. In general, scheduled mode is required for tightly coupled simulations, particularly those with time constraints, while unscheduled mode is appropriate for loosely coupled applications that can adapt to time-varying resources. Nevertheless, we can make general observations about their characteristics:

Scale and the need for selection. In the future we can expect to deal with larger and larger collections of test beds, from which resources will be selected for particular applications according to criteria such as connectivity, cost, security, and reliability.

Heterogeneity at multiple levels. Both the computing resources used to construct virtual supercomputers and the networks that connect these resources are often highly heterogeneous. Heterogeneity can arise at multiple levels, ranging from physical devices, through system software, to scheduling and usage policies.

Unpredictable structure. Traditionally, high-performance applications have been developed for a single class of system with well-known characteristics or even for one particular computer. In contrast, meta-computing applications can be required to execute in a wide range of environments, constructed dynamically from available resources. Geographical distribution and complexity are other factors that make it difficult to determine system characteristics such as network bandwidth and latency a priori.

Dynamic and unpredictable behavior. Traditional high-performance systems use scheduling disciplines such as space sharing or gang-scheduling to provide exclusive and hence predictable access to processors and networks. In meta-computing environments, resources especially networks are more likely to be shared. One consequence of sharing is that behavior and performance can vary over time. For example, in wide area networks built using the Internet Protocol suite, network characteristics such as latency, bandwidth and jitter may change, as traffic is re-routed. Large-scale meta-systems may also suffer from network and resource failures. In general, it is not possible to guarantee even minimum quality of service requirements.

Multiple administrative domains. The resources used by meta-computing applications often are not owned or administered by a single entity. The need to deal with multiple administrative entities complicates the already challenging network security problem, as different entities may use different authentication mechanisms, authorization schemes, and access policies. The need to execute user-supplied code at different sites introduces additional concerns. Mechanisms that allow applications to obtain real-time information about system structure and state are fundamental to all these issues. The Globus toolkit comprises a set of modules. Each module defines an interface, which higher-level services use to invoke that module's mechanisms, and provides an implementation, which uses appropriate low-level operations to implement these mechanisms in different environments.

Currently identified toolkit modules are as follows.

Resource location and allocation. This component provides mechanisms for expressing application resource requirements, for identifying resources that meet these requirements, and for scheduling resources once they have been located. Resource location mechanisms are required because applications cannot, in general, be expected to know the exact location of required resources, particularly when load and resource availability can vary. Resource allocation involves scheduling the resource and performing any initialization required for subsequent process creation, data access, etc. In some situations, for example, on some supercomputers location and allocation must be performed in a single step.

Communications. This component provides basic communication mechanisms. These mechanisms must permit the efficient implementation of a wide range of communication methods, including message passing, remote procedure call, distributed shared memory, stream-based, and multicast. Mechanisms must be cognizant of network quality of service parameters such as jitter, reliability, latency, and bandwidth.

Unified resource information service. This component provides a uniform mechanism for obtaining real-time information about meta-system structure and status. The mechanism must allow components to post as well as receive information. Support for scoping and access control is also required.

Authentication interface. This component provides basic authentication mechanisms that can be used to validate the identity of both users and resources. These mechanisms provide building blocks for other security services such as authorization and data security that need to know the identity of parties engaged in an operation.

Process creation. This component is used to initiate computation on a resource once it has been located and allocated. This task includes setting up executables, creating an execution environment, starting an executable, passing arguments, integrating the new process into the rest of the computation, and managing termination and process shutdown.

Data access. This component is responsible for providing high-speed remote access to persistent storage such as files. Some data resources such as databases may be accessed via distributed database technology or the Common Object Request Broker Architecture (CORBA). The Globus data access module addresses the problem of achieving high performance when accessing parallel file systems and network-enabled I/O devices such as the High Performance Storage System (HPSS).

The various Globus toolkit modules can be thought of together as defining a meta-computing virtual machine. The definition of this virtual machine simplifies application development and enhances portability by allowing programmers to think of geographically distributed, heterogeneous collections of resources as unified entities.

7.2. Gridge Toolkit

7.2.1. Overview

Gridge Toolkit [GRIDGE] is an internal PSNC open source software initiative aimed to help users to deploy ready-to-use grid middleware services and create productive grid infrastructures. All Gridge Toolkit software components have been integrated together and form a consistent distributed system following the same interface specification rules, license, quality assurance and testing.

Gridge Toolkit components have been successfully tested with different versions of Globus Toolkit TM as well as other core grid middleware solutions. The Gridge Toolkit software is available for free with full commercial support. Additionally to the services described in the next section PSNC offers for its partners and users:

1. technical support, consulting, training and development for Gridge Toolkit and Globus Toolkit TM,
2. assistance in design, deployment and configuration of grid middleware,
3. on-site installation and integration of Gridge Toolkit and Globus Toolkit TM key components,
4. workshop and hands-on training on "grid-enabled" technologies.

7.2.2. Tools and Services

Gridge Toolkit consists of the following tools and services:

1. **Grid Authorization Service (GAS)** is an authorization system, which can be the standard decision point for all components of a system. Security policies for all system components are stored in GAS. Using this policies GAS can return an authorization decision upon the client request. GAS has been designed such a way that it is easy to perform integration with external components and it is easy to manage security policies for complex systems. Possibility to integrate with many Globus Toolkit and operating system components makes GAS an attractive solution for grid applications.
2. **Grid Data Management System** is one of the main components of Gridge Data Management Suite (GDMSuite) — a middleware platform providing an uniform interface for connecting heterogeneous data sources over a network. GDMSuite stands for the backbone of the Gridge environment, on which computational services would

perform its operations. Gridge Data Management Suite constitutes a bundle of packages, designed for the creation of a complete and robust data management environment. It is intended to fulfill even the enterprise requirements of grid environments in terms of reliability, security and performance.

3. **Grid Mobile Services.** Mobile software development in our approach is focused on providing a set of applications, that would enable communication between mobile devices, such as cell phones, Personal Digital Assistants (PDA) or laptops and grid services on the other side.
4. **Toth Logging System.** This component was prepared to handle the problem of collecting events generated by distributed services of the Gridge environment. Some of them take advantage of a common approach for logging using LOG4J library, so Toth had to be LOG4J compatible.
5. **Grid Monitoring System.** Mercury Grid Monitoring System has been developed within the GridLab project and provides a general and extensible grid monitoring infrastructure. It is designed to satisfy specific requirements of grid performance monitoring: providing monitoring data represented as metrics via both pull and push model data access semantics and also supporting steering by controls.
6. **Grid Portals** are built depending on the target environment usage scenarios and requirements of the end user communities. The Gridge portals are composed of the following tools and packages: GridSphere, which is a standards-based, reliable Java portal framework and Grid Service Provider, which provides support for fast and easy deployment and usage of applications on the grid, and allows to easily scale the access environment to multiple independent portals.
7. **Grid Resource Management System (GRMS).** This component is an open source meta-scheduling system, which allows developers to build and deploy resource management systems for large scale distributed computing infrastructures.

All the pieces are integrated with each other and follow the same interface specification rules, license, quality assurance and testing, distribution etc.

Gridge tools and services enable applications to take advantage of dynamically changing grid environment. These tools have ability to deliver dynamic or utility computing to both, the application users and developers and resource owners. Through supporting the shared, pooled and dynamically allocated resources and services, managed by automated, policy-based GRMS that interfaces with such services as Mercury monitoring system, adaptive component services, data and replica management services and others, the Gridge offers the state of the art dynamic grid features to applications. Gridge technology can be used by various kinds of businesses, including vendors, but also financial companies or service organizations.

7.3. Akogrimo

Akogrimo is a project sponsored by the EU, which lasts for 3 years. It will be finished in September 2007. Akogrimo is an abbreviation for "Access to KnOwledge through the GRId in a MObile world" [WESAKO].

7.3.1. What Is This Project About?

As the acronym points out, the project is about bringing two diverse worlds, mobility and the grid, together. Basically, grids are now expected to be static, with static machines and static assignments of users, computing power and network connectivity. The vision of the project is that everything should be mobile and nevertheless provide the computing power to the users one would expect from a conventional grid.

7.3.2. What Are The Problems?

The grid resources must be managed and shared in a reliable and secure way, even if the grid computers spread among several organisations. While this is a problem on its own, Akogrimo adds another dimension: The users and computing elements may move from one place to another without any notice in advance. This should not be only possible by signing off and on again but also by changing the access network provider during a session.

The reason for such an idea comes from the fact that it should not be only possible to use laptops as grid elements, but also mobile phones and other similarly small electronic devices. This way the grid will become pervasive and much more useful than it is now.

Mobility brings many new problems such as personalization (moving your personal data to the new location), privacy (not leaving trails which could be used by people working with the same resources to spy on you and your work), security (encrypt communication channels and data storage facilities so that strangers cannot access your data) and trust (one must be able to evaluate whether information, for example given by central servers, are trustworthy). The algorithms used must be carefully chosen so that they will be effective even for low-bandwidth connections or fast-moving targets.

7.3.3. How Will The Problems Be Tackled?

Akogrimo consists of four layers:

1. **Network Services Layer** is the bottom-most layer. It is responsible for network management, Quality of Service (QoS) and provides an interface to the IPv6 and MIPv6 infrastructure.
2. **Network Middleware Services Layer** is providing Authentication, Authorization and Accounting (AAA).
3. **Grid Infrastructure Services Layer** is the top-most layer. It provides services like policy management, execution management or data management. Because this layer is also needed from other projects there already exists a good infrastructure in the form of Open Grid Service Infrastructure (OGSI) and Web Services Resource Framework (WSRF). It is expected that Akogrimo will make use of these two frameworks.
4. **Application Support Layer** is orthogonal to the previous three layers and provides low-level services as well as high-level services. Among these services are the "VO Manager" or "Workflow Enactment".

7.4. Virtual Laboratory

7.4.1. Overview

The Virtual Laboratory research project [VLABWEB] has been developed in Poznań Supercomputing and Networking Center since the beginning of the year 2002. VLab was a part of the "High Performance Computations and Visualization for Virtual Laboratory Purposes with the Usage of SGI Cluster" project, which was funded by the State Committee for Scientific Research. Furthermore, a parallel research grant has been launched by the Ministry of Scientific Research and Information Technology, more oriented towards the research approach.

The main goal of the VLab project was to design and develop a universal system which should provide a remote access to the wide variety of different laboratory equipment. Additionally, the concept of the VLab tasks, scientific instruments and computational tasks and resources in the Grid environment was generalized. The idea of "dynamic measurement scenarios" [LAWDYNM, LAWDYNW] was introduced, and the users are capable of creating a dynamic workflow, describing their research process. On those dynamic scenarios (or workflow

diagrams) experiments with remote access to laboratory equipment are treated as any other "regular" tasks. The user creates the scenario using the graphical tool, by placing appropriate boxes representing tasks and connecting them with arrows, defining the data flow between different experiments on scientific instruments and "regular" computational tasks, distributed across the grid.

A lot of effort was also put on the aspect of sharing and reusing the experiments' results. The Digital Science Library [REKDSL] — a distributed database system, was introduced, allowing users to store their large data files with experiments results, pictures, papers and any other kinds of material. The data can be described using the flexible meta-data system. Each element from the database can have a custom defined meta-data information, used for categorizing and effective search. The owners of the information stored in the Digital Science Library have full control over access rights to their material.

The pilot installation of this system was done for the domain of the NMR spectroscopy, with cooperation with Institute of Bioorganic Chemistry of the Polish Academy of Science from Poznań, Poland [PASWEB]. The system was installed and configured for two NMR spectrometers, allowing its users to access Varian Unity 300MHz spectrometer and more advanced Bruker Avance 600MHz. After its successful deployment, the work has started on the remote access to the 32m radio telescope located in Piwnice, Poland. This work is currently under progress, and it is done in close cooperation with the Radio Astronomy Department of the Nicolaus Copernicus University in Toruń, Poland [RADWEB]. Furthermore, the work has just started to include the Freeze Atmospheric Dryer — a custom built device from the Faculty of Process and Environmental Engineering of the Technical University of Łódź, Poland [PEEWEB].

7.4.2. General Virtual Laboratory Architecture

The Virtual Laboratory system has a modular architecture. The modules can be grouped into three main layers. The general diagram is presented in figure 54 below [LAWGENC]:

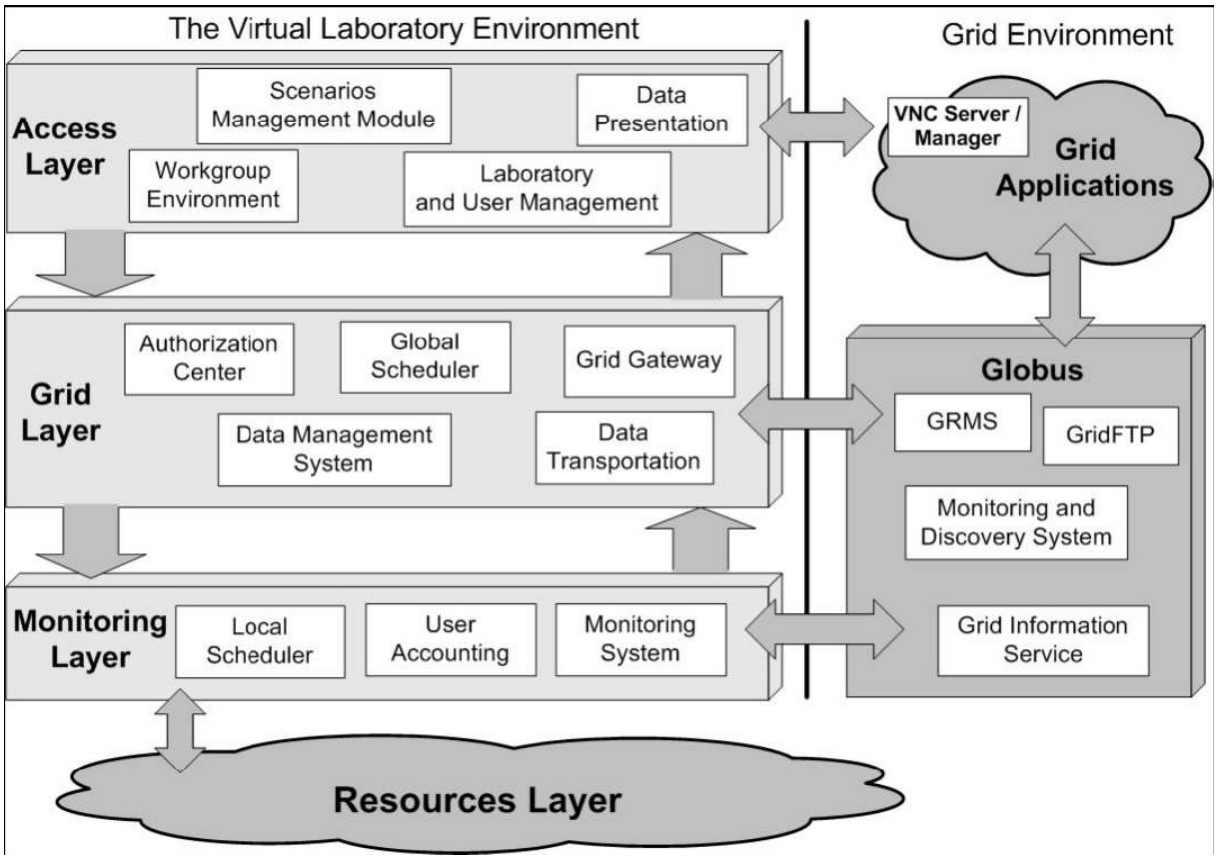


Figure 54: Virtual Laboratory system architecture

An access layer is the top layer of this structure. Basically, it contains all modules and components responsible for a VLab user access and graphical interface to the system (including a web portal), and data input interface. Below there is a grid layer, which communicates with the external grid environment. It is responsible for user authorization and authentication, data management, general task scheduling. Modules from this layer also handle the transfer of the computational tasks to the Globus [GLOBWEB] system, and gather feedback data and the computed results. The monitoring layer consists of lower-level modules, such as hardware dedicated schedulers, system monitoring, gathering accounting data, etc. All the physical scientific devices are located in the resources layer, as well as modules responsible for their direct control. On the grid environment side the most important element is the Globus system, with all its components (GRMS [GRMSWEB], GridFTP etc). Globus also allows to execute computational tasks (batch and interactive) on a wide variety of grid applications.

8. Projects and Testbeds

8.1. GridCC

The Grid-enabled Remote Instrumentation with Distributed Control and Computation (GRIDCC) EU project aims at integrating scientific instrumentation with the grid. More specifically, GRIDCC builds software which will be integrated with gLite and will support the interactions between instruments and currently available grid components, as well as satisfy the requirements of instrument control and monitoring.

The project's main contribution is the Instrument Element (IE), an abstraction similar to the Computing Element and the Storage Element of the classical grid. Additionally to that, the project studies methods to provide QoS guarantees to service consumers of IEs (or other resources, when that is possible) and to enable building and executing complex scientific workflows which are QoS-enabled, using the Workflow Management System, the Agreement Service and the Performance Repository. It also comes with a novel Multipurpose Collaboration Environment (MCE) which includes an integrated Virtual Control Room (VCR) for the control of remote instrumentation. A host of applications are being developed based on the GRIDCC middleware (or ported to it), as part of the project. Security-wise, GRIDCC is using a novel approach based on Kerberos tickets, thus reducing the security-related load significantly. A problem solver component based on artificial intelligence methods is also included, to allow for efficient tracking and solving of operational problems.

GRIDCC includes a host of diverse use cases, applications for which are being built or ported to the produced middleware.The main applications that the project is looking at are:

- Compact Muon Solenoid (CMS) experiment
- far remote operation of particle accelerator
- power grid monitoring and control
- meteorology/weather forecasting

Other applications of the project include a network intrusion detection system, a geology application, a neurophysiology one, and a "device farm" of telecommunication measurement instruments.

8.2. EXPReS

Over the last decade, enormous progress has been made in the area of Information & Communication Technologies (ICT). Like many other research communities, radio astronomy has greatly benefited from the ready availability of cheap, large-capacity storage media, affordable high-performance desktop computers and local multi-processor PC clusters. Radio astronomers have also proven themselves to be adept at taking advantage of the emergence of high-speed, optical-fiber based communication networks. Such networks are of prime interest to radio astronomers as the sensitivity of a radio telescope array is roughly proportional to the square-root of the radio telescope data output rate (i. e. the digitally sampled observing bandwidth). Motivated by this fact, and the possibility of creating a distributed real-time Very Long Baseline Interferometer (VLBI) radio telescope array, an international consortium has come together over the last two years and successfully demonstrated the feasibility of conducting real-time VLBI observations. The consortium includes the leading radio astronomy institutes in Europe, together with DANTE (operators of the pan-European Research Network, GÉANT), and the major National Research Networks (NRENs) in Europe.

VLBI is a technique invented by astronomers, in which physically independent and widely separated radio telescopes observe the same region of sky simultaneously in order to generate very high-resolution continuum and spectral line images of cosmic radio sources. The technique also has practical applications in geodesy (measuring secular and periodic motions of the earth's surface and variations in the earth's rotation rate) and in precision spacecraft navigation. Since VLBI telescopes are usually separated by many thousands of kilometers, data from each telescope are digitally sampled and stored locally, using high-capacity magnetic tape systems (or more recently magnetic disk-array systems). These tapes and disks are physically transported to a central data processor (a purpose-built supercomputer) where the data from each telescope are decoded, accurately aligned, and then correlated together in an exhaustive pair-wise fashion for every possible telescope combination. The total flow of data into the central processor is approximately 10-100 Terabytes per observation, after processing this is reduced to 10-100 Gbytes. The output from the data processor may be pipelined — a process in which the data are automatically calibrated and preliminary images are made. These data (including the raw processed data) are presented to the astronomer for further detailed and careful analysis. In Europe, VLBI observations are organised and conducted by the European VLBI Network (EVN). The central EVN data processor was developed (and is now operated) by JIVE — the Joint Institute for VLBI in Europe, located in Dwingeloo, the Netherlands.

The overall objective of EXPReS, is to create a production-level "electronic" VLBI (e-VLBI) service, in which the radio telescopes are reliably connected to the central data processor at JIVE via a high-speed optical-fibre communication network. With an aggregate data flow of up to 16 Gbps into the central processor, we aim to create a robust and distributed e-VLBI infrastructure of continental and indeed trans-continental dimensions, capable of generating high-resolution images of cosmic radio sources in real-time. EXPReS also seeks to design and prototype elements of the hardware, software and data transport services required to support future e-VLBI facilities in which the net VLBI data flows will be hundreds of Gbps, with a central data processing environment possibly based on distributed (grid based) computing resources.

Explicitly, the principal objectives of EXPReS are:

- Develop an operational, production-grade e-VLBI network that is capable of simultaneously transferring data at aggregate data rates of up to 16 Gbps, from telescopes located across the globe, through to the central data processor at JIVE via public networks operated by the NRENs & DANTE (GÉANT2). To expand the number of e-VLBI capable telescopes in Europe by either supporting the provision of additional last-mile (local-loop) connections or upgrading existing fibre connections to Gigabit Ethernet standard. In total, at least 12 European telescopes will be part of the e-VLBI network offered to astronomers.
- Transparently include the UK e-MERLIN telescope array within the e-VLBI facility.
- Support the connection of radio telescopes located outside of Europe, recognising that VLBI is very much a global pursuit, requiring network provision to telescopes in the US, Asia, South Africa, Australia & South America. By including telescopes across the globe, we aim to offer a 16 telescope e-VLBI array to astronomers.
- Reliably service and robustly process in real time, e-VLBI data streams of up to 16 Gbps (net) at the EVN data processor at JIVE.
- Ensure that this state-of-the-art, real-time e-VLBI network is able to conduct "Target of Opportunity" and "Rapid Response" science, reacting reliably and flexibly to unexpected astronomical events, such as supernovae explosions, giant magnetar and x-ray binary flare stars, (orphan) radio gamma-ray bursts and other transient phenomena.

- Assess the suitability of advanced networking and computing technology to support the creation of a next-generation e-VLBI network in which the aggregate data flows will be many hundred of Gbps with a data processing environment possibly based on distributed grid-based computing resources.
- Investigate how the new radio astronomy facilities now under development (e. g. e-MERLIN in the UK, and LOFAR in the Netherlands, Germany and Sweden), can further expand using e-VLBI as a model for their own use of public communication networks.
- Promote and demonstrate the way in which communication research networks can be used to create enhanced, large-scale distributed scientific facilities and strengthen the links between radio astronomers, engineers, network operators and grid computing experts via a comprehensive programme of networking activities.

8.3. CRIMSON

CRIMSON is an acronym for "Cooperative Remote Interconnected Measurement Systems Over Network", an Italian research program which aimed at investigating a number of aspects that pertain to the realization of a system of interconnected collaborative distributed nodes, such as sensors for the acquisition of information, signal processing platforms, telecommunication networks and measurement devices, concurring in the characterization tele-measurement tasks of systems and processes.

The nodes in the measurement network may include simple low cost sensors (e. g. for monitoring geographically distributed processes), digital signal processing platforms (e. g. software radio), high performance measurement instrumentation, and even whole laboratories, within a complex hierarchical structure for tele-measurement. The key element underlining the entire architecture is the collaboration among nodes, with the aim of improving the measurement capabilities of the whole system.

The very nature of measurement systems, aimed at providing results that must be i) characterized by a precisely quantifiable degree of uncertainty, ii) repeatable, and iii) comparable with reference samples, poses quite specific requirements on their utilization within distributed systems; this fact adds complexity to the well-known and largely experimented models for the network interconnection and sharing of pure computational resources.

The research units who participate in this program are:

- Department of Communications, Computer and Systems Science (DIST)
- Department of Biophysical and Electronic Engineering (DIBE), University of Genoa
- Department of Electronics, Turin Technical University
- Department of Information Engineering, University of Padova
- Department of Electronics, Computer and Systems Science, University of Bologna
- Department of Information Engineering, University of Parma

In this framework CRIMSON has pursued multiple objectives, reflecting different aspects of a complex system, with heterogeneous components, both in terms of tasks and functionality. The objectives (essentially the workpackages WP) of the program, have been the following:

- WP0: Coordination and dissemination of results
- WP1: Analysis of scenarios and platforms for tele-measurement
- WP2: Requirements and architectural specificatons for interoperability

- WP3: Reconfigurable user interfaces for access to di tele-measurement platforms
- WP4: Study of techniques for the acquisition, transport and pre-processing of signals
- WP5: Computational GRIDs for the sharing of distributed instrumentation and distributed signal processing
- WP6: Demonstration and experimentation

Among the main results achieved within these objectives, we can mention:

- Study and realization of architectural solutions for interoperability [DAVDCL, BENATE, BERARM, BENSCN, BERMSA, DAVLN, ANDTM]
- Definition of requirements of applications for creation and rendering of user interfaces aimed at the remote access to instrumentation and measurement test benches [BAGSLI]
- Metrological characterization of distributed measurement systems [CARMMM, CARTAT], and the development of new lossy compression algorithms, specifically oriented to measurement systems [CACUEC]
- Characterization of sensor networks, as distributed large-scale acquisition devices [QUEEEW, DARMEW, FRADLPC, FERDDCL, FERSDB]
- Definition and analysis of grid architectures for tele-measurements [CAVIRV, BAGEMI]

8.4. UCRAV

The Chilean organization — Red Universitaria Nacional (REUNA) — which is a member of CLARA, has launched a project based on the concept of a virtual laboratory called UCRAV (http://www.ucrav.cl/). UCRAV will eliminate geographical barriers, expanding its services from a national level (Chile) to an international one.

UCRAV is a pilot platform built with a set of tools designed for the Internet whose purpose is to offer remote instrumentation services. UCRAV uses resources available in the universities participating in the project in order to benefit researchers and academics from universities, research centres, and private and public enterprises. It also provides a new range of possibilities and benefits for both the providers and the users of the service.

The UCRAV solution uses the grid concept — application of distributed computing — for the research and development of the scientific-technological activity, allowing the remote manipulation of instruments available within an environment of collaboration among researchers and users. All the UCRAV applications are built with open code standards. An outstanding feature is the use of Globus Toolkit tools for the construction of the services grid.

8.4.1. Confocal Microscopy, Unit of Integral Cellular Analysis CESAT-ICBM

This instrument is located at the Faculty of Medicine, Universidad de Chile. Types of analyses available:

1. observation and microscopic analyses by means of transmitted light DIC and in epifluorescence in different materials prepared for microscopy
2. detection of auto-fluorescent compounds or molecules marked with fluorescent antibodies (immunocytochemistry)
3. determine the distributions of substances in different planes of cells by obtaining series of images in different optical planes

4. dynamic determinations of molecules in live cells by obtaining images across time (temporal series)
5. temporal quantification of the calcium levels present if live cells by using Fluo 3
6. quantitative measurements of changes of volume in live cells charged with calceine in response to specific stimuli. Characteristics of the samples: they can be composed of live cells and histological cuts of normal and pathological preparations

Areas of application: Medicine, Chemistry, Pharmacy, Veterinary Science, Agronomy, etc.

8.4.2. Nuclear Magnetic Resonance Spectrometer

This instrument is located at the Universidad de Concepción. Types of analyses available:

1. 13C-NMR. Determination of molecules'structure
2. 13C-NMR makes it possible to characterise the carbon atoms of a molecule (methyl, methylene, methyne and quaternary carbons).
3. 13C-1H Bidimensional spectrum. It specifically shows correlations to a linkage (1JC-H) between carbon and hydrogen atoms.
4. 13C-1H Bidimensional spectrum. It specifically shows correlations to more than one linkage (2JC-H and 3JC-H) between carbon and hydrogen atoms. Particularly useful for the assignation of quaternary carbons.
5. 1H-NMR. Determination of the structure of molecules and the composition of mixtures.
6. 1H-1H Bidimensional spectrum. It specifically shows H-H correlations (nJH-H).
7. 1H-1H Bidimensional spectrum. It specifically shows dipolar couplings.
8. X nucleus NMR, whose resonance frequency is within the range of 129 Ag and 31P. For instance, 11B, 15N, 27 A1, 31P

Characteristics of the samples: In general the analysis can be applied to all kinds of soluble compounds in deuterated sdvents, except natural polymers such as proteins.

Areas of application: Chemical Sciences, Biological Sciences, and Pharmacy.

8.4.3. X-Rays Diffractometer

This instrument is located at the Universidad Católica del Norte.

Types of analyses available: The X-rays diffraction analysis, in powder samples, is used for the structural identification of all kinds of crystalline compounds of inorganic, organic and mineral nature.

Characteristics of the samples: a minimum quantity of 1 or 2 grams, free of humidity and with a granulometry below 65 microns,is required for the analysis.

Areas of application: Geology, Mining Industry, Cement Industry, and Chemical Industry.

8.4.4. Semprobe, Analytical Electron Microscope

This instrument is located at the Department of Geology at the Faculty of Physical and Mathematical Sciences, Universidad de Chile.

Types of analyses available:

1. Qualitative and quantitative microanalysis by x-rays longitude dispersion, for elements with an atomic number superior to that of Boron.
2. Scanning Electron Microscopy (SEM).

3. Conventional and digital microphotography.
4. Analysis of digital, photographic and video images where reactions and events at micrometric scale and occurring in real time, can be observed.

Characteristics of the sample: In general, samples are of an inorganic nature, and they must not have a bigger size than 3×2×2 cm (length, width, height).

Areas of application: Mineralogy, petrology, geochemistry, mine geology, mineral deposits, edaphology, archaeology, gemmology, numismatics, chemistry and biology, among others. Quality control in metallurgy, ceramics and glass, inorganic chemistry, odontology, solid materials alteration, and jewellery are some of the possible applications.

8.5. SatNEx II: Satellite Network of Excellence

SatNEx II (and its predecessor, SatNEx) is an FP6 research Network of Excellence (NoE), funded by the European Commission, which combines the research excellence of 24 major players in the field of satellite communications [SNXWEB, EVASNXO, WERSNXO, SHETRAI]. The primary goal of SatNEx II is to achieve a long-lasting integration of European research in satellite communications, and to develop a common knowledge base. This collected expertise will support the European satellite industry through standardization, collaboration/consultancy and training. Through co-operation of outstanding universities and research organizations with excellent expertise in satellite communications, SatNEx II is building a European virtual centre of excellence in satellite communications and will contribute to the realization of the European Research Area (ERA). A dedicated satellite platform links partners in a broadcast, multicast or unicast configuration, providing training and video-conferencing capabilities, and promoting the simplicity and cost-effectiveness of using satellites for this purpose. SatNEx II has established an advisory board incorporating key representatives of the European space industry, satellite service providers, and standardization and regulation organizations. SatNEx II is steered by these players in providing a critical mass of resources and expertise, to make Europe a world force in the field of satellite communications. Part of the SatNEx II mission is to disseminate internal research and expertise.

Among SatNEx II workpackages, WP 2400 is dedicated to test beds and trials. Within the proposed experiments to be carried out on various satellite platforms, the interconnection of remote measurement instrumentation, with or without the presence of grid middleware, is one of the foreseen possibilities. Some preliminary efforts in this direction have been made over the CNIT satellite network [BERIESS].

8.6. RedCLARA

RedCLARA (see figure 55) is the Latin American (LA) regional research and education (R&E) network, interconnecting LA NRENs, and providing them with intercontinental access to advanced R&E networks in Europe and the United States, and hence to the rest of the world (see www.redclara.net/en/index.htm).

RedCLARA became operational in 2004 and is based on a backbone ring with 155 Mbps capacity, with nodes in Argentina, Brazil, Chile, Mexico and Panama. By February 2006, in addition to these five countries, RedCLARA also connected a further 9 LA NRENS in the following countries: Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Nicaragua, Peru, Uruguay and Venezuela, at access rates between 10 and 45 Mbps. Since 2004, the RedCLARA backbone is connected at 622 Mbps from the its node in Brazil to the GÉANT node in Madrid. This pioneering regional R&E network was built by the ALICE (Latin America Connected to Europe) project, partially financed by DG EuropeAid and coordinated by DANTE (see http://alice.dante.net/). The ALICE project is expected to continue at least until March, 2008.

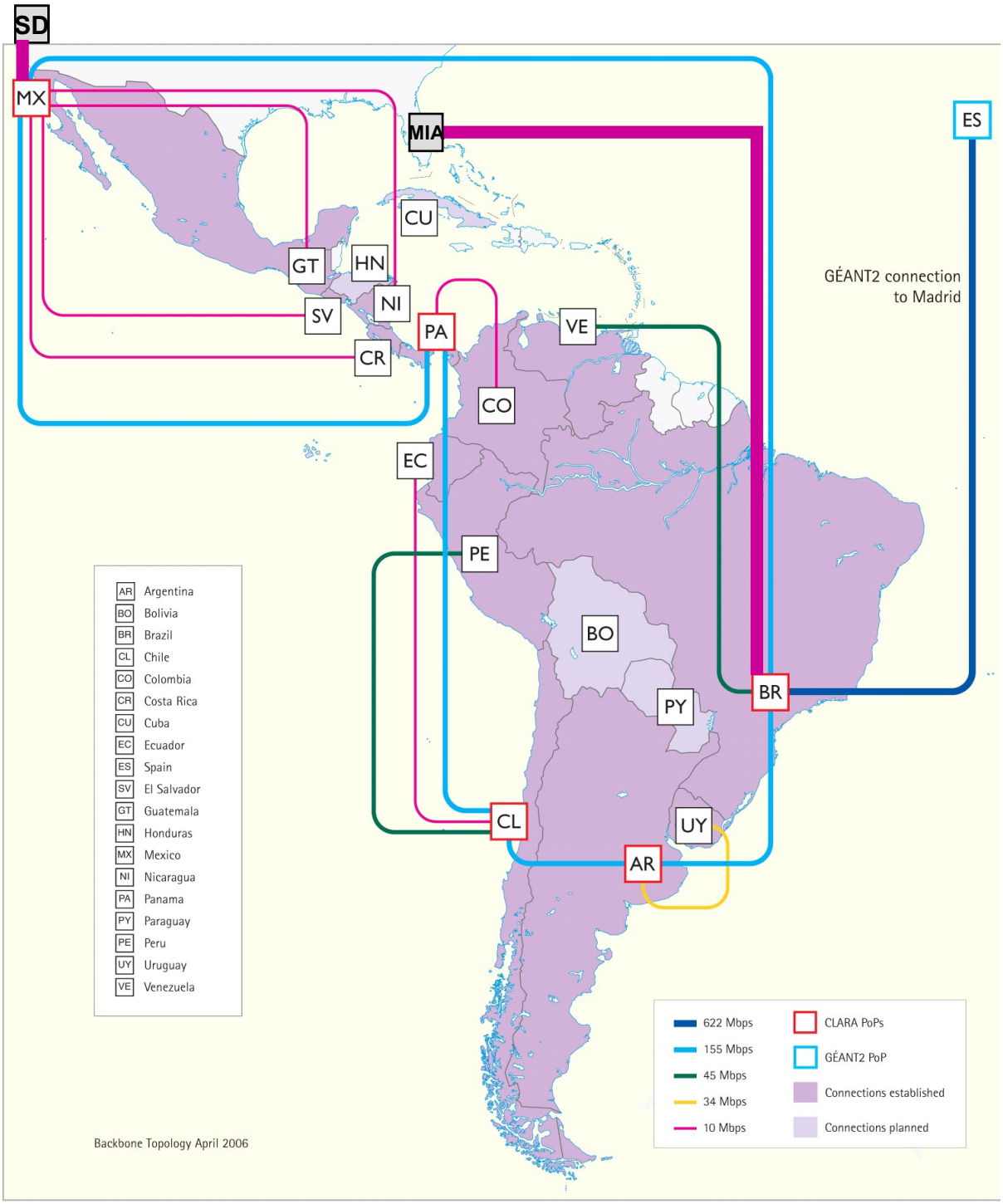


Figure 55: RedCLARA network in 2006

In addition to the EU investment in LA R&E connectivity, the US government, through the NSF, has also invested in intercontinental connectivity to Latin America, through the International Research Network Connexions programme (IRNC), which is supporting the WHREN-LILA project for the period 2005-2009. This project currently provides two shared multigigabit connections between the US and RedCLARA nodes in Brazil and Mexico (see http://whren.ampath.net/). Between San Diego, California, and Tijuana, Mexico, a cross-border

dark fibre connection is shared by 2 GigE connections, one used by RedCLARA and the other by the Mexican NREN, CUDI. Between Miami, Florida, and São Paulo, Brazil, a 2.5 Gbps SDH link is shared between RedCLARA, the Brazilian NREN, RNP, and the Brazilian state of São Paulo's R&E network, ANSP.

RedCLARA is maintained and operated by CLARA, an association of existing and recently created LA NRENs to provide advanced international connectivity to member networks (see www.redclara.net/en/01.htm). Services are provided by the staff of member organisations. In particular, network engineering is carried out by Brazil's RNP, and network operation's by Mexico's CUDI.

RINGrid partners include UNAM (Mexico), REUNA (Chile) and RNP (Brazil), and so a short description will be provided of the NRENs of these three countries.

CUDI is the Mexican NREN, and currently operates a backbone network at 155 Mbps with access speeds up to 34 Mbps. The topology of the CUDI network is shown in figure 56, and indicates the international connections through the nodes of Tijuana (RedCLARA and US networks) and Ciudad Juárez (US networks). The network interconnects 22 members and a further 49 affiliates, including universities, colleges and research centres. The CUDI website can be found at www.cudi.edu.mx.

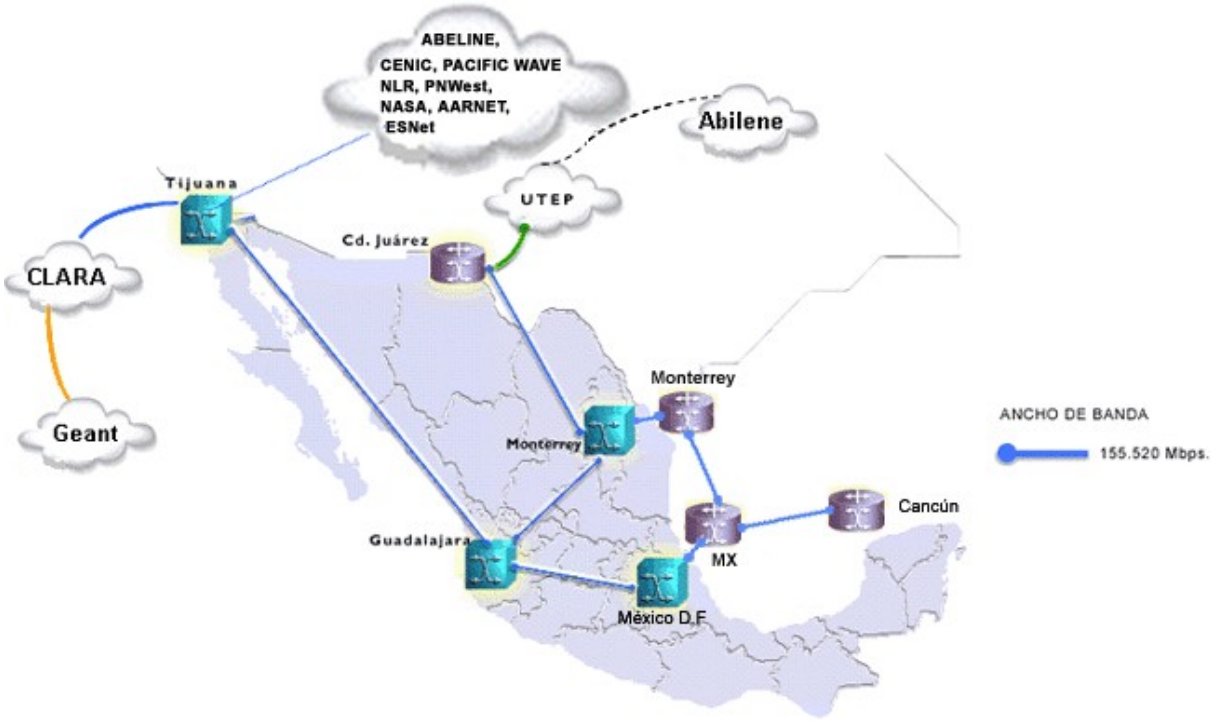


Figure 56: Topology of the CUDI backbone in Mexico

REUNA is the Chilean NREN, and since 2006 operates the GREUNA backbone network with up to 310 Mbps capacity. Access speeds from particular members varies considerably. Currently 17 universities and research institutions are served by GREUNA, whose topology is shown in figure 57. International R&E connectivity is provided by RedCLARA, through the RedCLARA node in the Chilean capital, Santiago. The REUNA website may be found at www.reuna.cl.

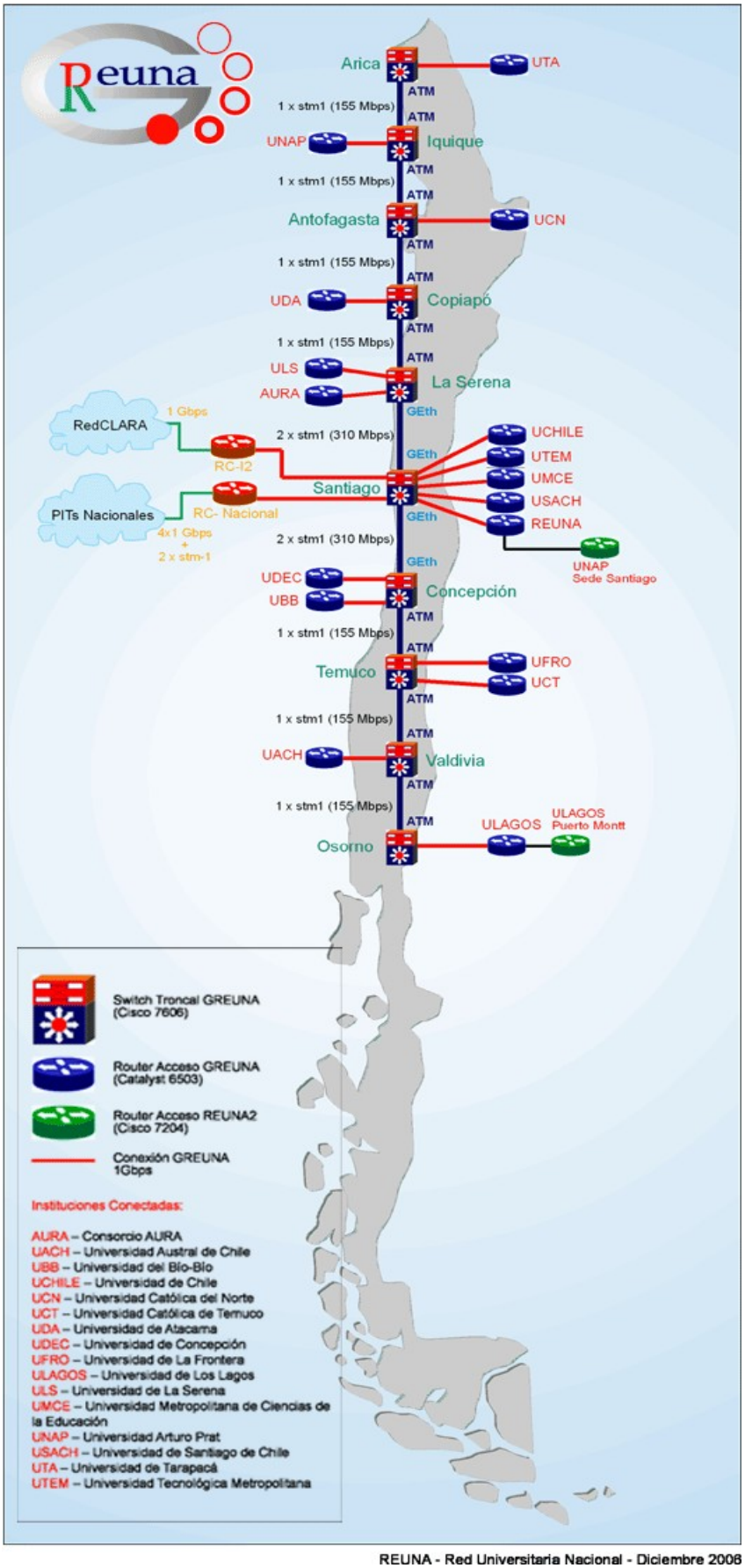


Figure 57: Topology of the GREUNA network in Chile

RNP is the Brazilian NREN, which is supported by the federal government of Brazil, and provides national and international connectivity to the Brazilian higher education and research community through its national backbone IPÊ network, with points of presence in the 26 state capitals and in the national capital, Brasília. Currently 4 capitals are connected with 10 Gbps links, 6 with 2.5 Gbps links and the remainder mostly at 34 Mbps, although some lower rate connections are used in the extreme north and northwest (see figure 58). Current international R&E connectivity is provided by RedCLARA, with its 622 Mbps link to GÉANT and by the WHREN-LILA 2.5 Gbps link to the US, which is shared with RedCLARA and ANSP. RNP also provides commodity internet access for its clients, through peering agreements with Brazilian telcos and the purchase of international access from international suppliers. See www.rnp.br/en.

Access to the IPÊ backbone network is through the points of presence in the capital cities. By the end of 2007 it is expected that over 200 institutions located in the metropolitan districts of the capitals will be connected to the local point of presence through

at least a 1 Gbps GigE connection, over a permanent fibre-optic infrastructure currenly being built by RNP, in collaboration with local institutions. RNP also maintains direct connections between the points of presence and federal government maintained higher education and research institutions located in non-capital cities at transmission rates up to 155 Mbps. In all, it is estimated that between 300 and 400 institutions are currently served, directly or indirectly, by the RNP infrastructure.

Several of the 26 state governments support stateside R&E networks, which are connected to RNP's IPÊ backbone. In the state of São Paulo, the government finances partially the WHREN-LILA link to Miami (US), and shares this between its own network, ANSP, RNP and RedCLARA. The Rio de Janeiro state network purchases national and international commodity internet access for its clients.

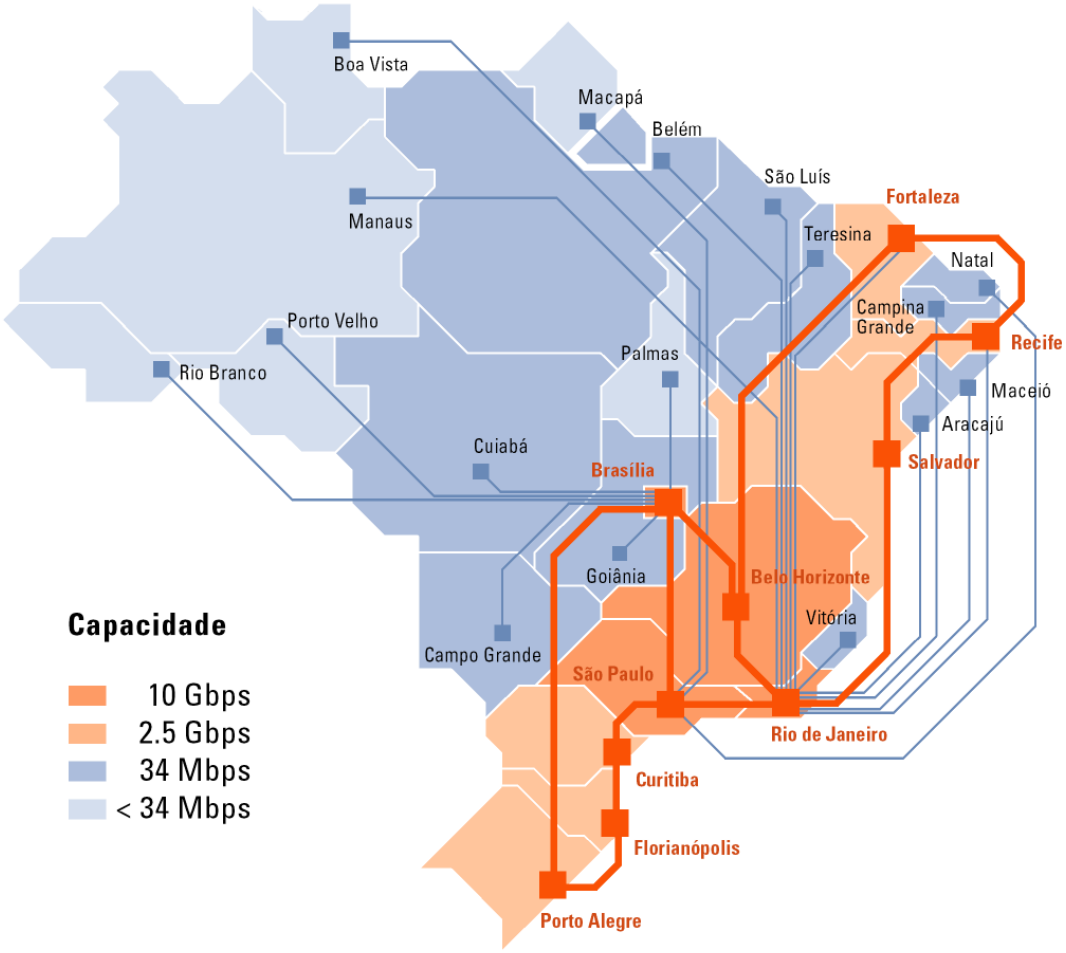


Figure 58: Topology of the IPÊ backbone network operated by RNP

The LA networks (RedCLARA, CUDI, REUNA and RNP) provide traditional best-effort IP service to their clients. All of these networks also support an increasing set advanced services, such as IP multicast, VoIP, and other multimedia services. RedCLARA also implements MPLS services in its backbone. However, until now, no support except overprovisioning has been given for QoS. Recent interest in improving end-to-end service provision has led to increased monitoring of network performance, particularly in support of the grid services provided by the

EU's EELA project (www.eu-eela.org), and the significant participation of Brazil's RNP in the PerfSONAR project (see www.perfsonar.net/partners.html).

8.7. GÉANT2

GÉANT2 (see figure 59) is a multi-gigabit research and education network, pan-European in scale and reach. It provides the European research and education community with a state-of-the-art data communications backbone network. The network provides the most advanced services and widest geographical reach of any network of its kind in the world, boasting leading-edge standards of reliability and innovation.

GÉANT2 connects 30 European national research and education networks (NRENs) which serve 34 countries. The NRENs connect research and educational institutions within their respective countries (though the exact structure for doing so varies between nations). More than 30 million research and education end users in over 3,500 institutions across Europe are connected to GÉANT2.

The initial topology of the new network was announced at the project's official launch on 14-15 June 2005. The first links of the network came into service in Q4 2005. The network will operate until September 2008.

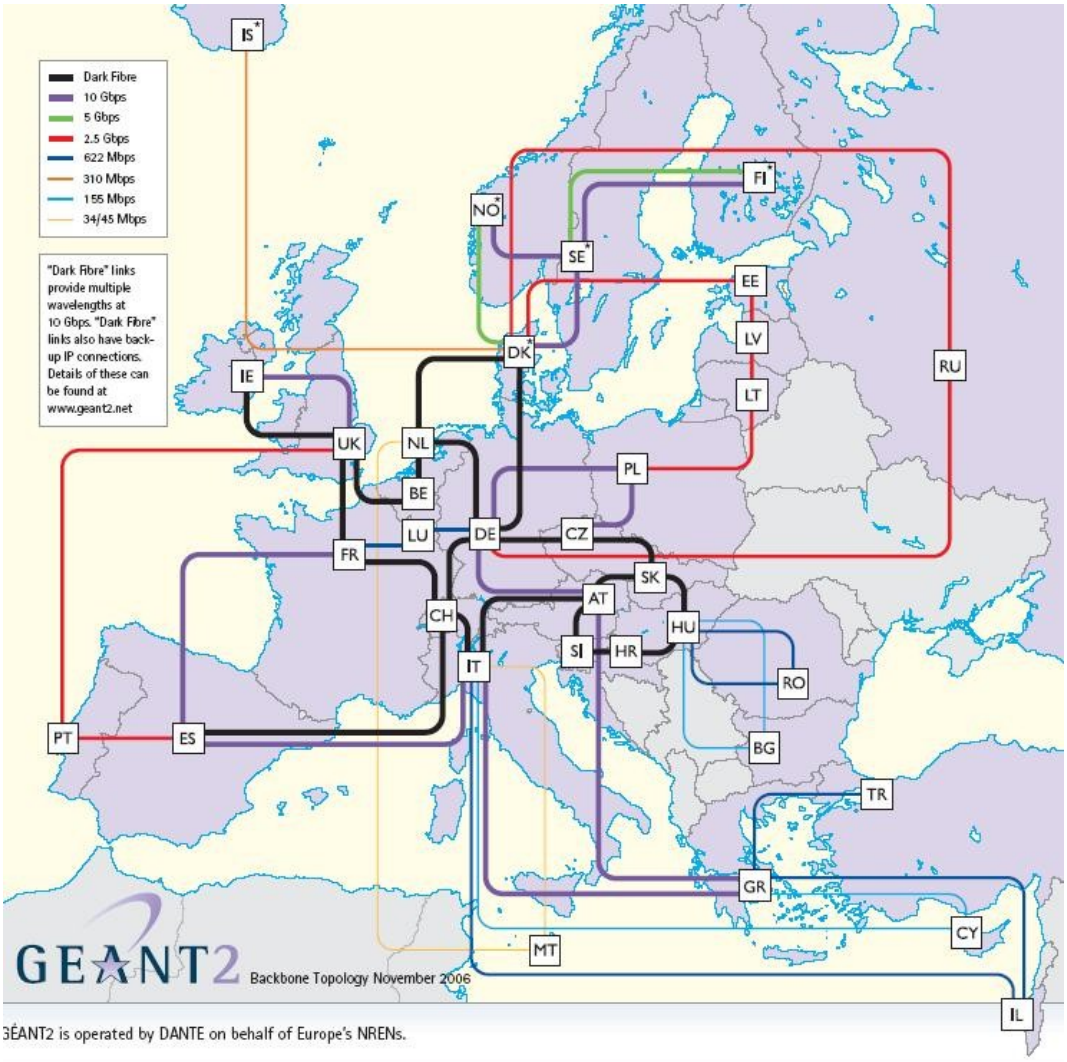


Figure 59: GEANT2 topology as of November 2006

The pan-European backbone approach has been definitively proven by GÉANT2's predecessor networks, and can offer, amongst other benefits:

- Maximum efficiency in the centralisation of network management,
- high concentration of networking expertise in support of European research and education,
- easy access for users in other regions and countries around the world to resources and equipment which would otherwise be out of reach.

Research and education networking in Europe is organised in a hierarchical fashion. GÉANT2 provides the pan-European backbone to interconnect Europe's national research and education networks. Together, GÉANT2 and the NRENs provide advanced communications services to Europe's research and education community.

The network architecture is evolving to a more flexible structure based on a combination of routed IP and switched components. The objective is to create a hybrid infrastructure that meets the needs of different types of user with the most appropriate technology.

Within GÉANT2, a best-of-breed approach to end-to-end service provision is being developed. Representatives of the NRENs that will be connected to GÉANT2 are working with DANTE to develop a number of production-quality tools for enabling users to receive a premium service, and to provide a helpdesk designed to identify and resolve network performance issues.

The end-to-end service provision activity is divided into two areas of work.

8.7.1. Provisioning At The IP Layer

The first end-to-end guaranteed service planned for GÉANT2 is Premium IP (PIP). To have Europe-wide, true PIP, all connected NRENs would have to commit to and implement the DiffServ model. However, a network which does not implement DiffServ can still take part in PIP QoS if it meets the following criteria:

- Its capacity is significantly over-provisioned.
- It forwards IP packets without changing the DSCP field in the packet header.
- It correctly deploys the provisioning system designed by the end-to-end service initiative.

This area of activity is developing the Advance Multi-domain Provisioning System (AMPS). This system, operating in a distributed manner, will manage the whole provisioning process, from user request through to the configuration of the appropriate network elements. AMPS version 1 was released at the end of 2005. In its initial form, actual router configuration must be performed manually, with the lead-time to reserve bandwidth measured in days. In time, however, AMPS will be able to automatically configure network elements, with a potential reduction in lead-time to minutes.

8.7.2. Provisioning At Lower Layers

IP technology as it is commonly used does not cater for those with exacting traffic requirements, such as low packet loss or low jitter. The Bandwidth on Demand activity is investigating methods of providing a guaranteed-bandwidth service across more than one network. Increasingly, and in particular with the advent of more widespread grid activity, research projects and initiatives are beginning that have potential requirements for high-capacity BoD services. These users require reserved bandwidth, predictable performance and logical separation from other network traffic.

The multi-domain nature of the networks on which the BoD service will be implemented makes it probable that more than one technology will be used to provide the service. Potential technologies that fall within the scope of this activity include:

- MPLS label-switched paths (LSPs), possibly enhanced with packet-based QoS standards
- native (and emulated) layer 2 channels (particularly Ethernet)
- time-division multiplexing (TDM) channels (based on SONET or SDH transmission)
- native layer 1 wavelengths on fibre ("lambdas")

8.8. EGEE

EGEE is short for Enabling Grids for E-sciencE. The EGEE project brings together scientists and engineers from more than 90 institutions in 32 countries world-wide to provide a seamless grid infrastructure for e-science that is available to scientists 24 hours a day. Conceived from the start as a four-year project, the second two-year phase started on 1 April 2006, and is funded by the European Commission.

Expanding from originally two scientific fields, high energy physics and life sciences, EGEE now integrates applications from many other scientific fields, ranging from geology to computational chemistry. Generally, the EGEE grid infrastructure is ideal for any scientific research especially where the time and resources needed for running the applications are considered impractical when using traditional IT infrastructures.

The EGEE grid consists of over 20 000 CPUs available to users 24 hours a day, 7 days a week, in addition to about 5 Petabytes (5 million Gigabytes) of storage, and maintains 20 000 concurrent jobs on average. Having such resources available changes the way scientific research takes place. The end use depends on the users' needs: large storage capacity, the bandwidth that the infrastructure provides, or the sheer computing power available.

9. Summary

This deliverable sheds light onto current and prospective networking technologies and grid infrastructures, which can be used to interconnect scientific instruments. The deliverable discusses relevant technologies in depth, starting from base technologies like wave division multiplexing, lambda networks and other optical or fibre topics. While these technologies cannot be deployed within the scope of the RINGrid project, the relevant sections give detailed insight into the to-be-used networks if they happen to utilize these technologies. RINGrid partners (especially WP6 partners) will be able to understand network characteristics better with the help of this knowledge.

The sections "Switching in Networks" as well as "Transport and Application Layer" give several algorithms, describe several protocols which can be implemented in the scope of WP6. It has to be determined which protocols are worth deploying and which have a too narrow scope in order to be useful.

The chapter on "Network Layer Protocols" describes IPv4 and IPv6, whereas nobody will deny the advantages of IPv6. However, it should help the implementation in the scope of WP6 to understand the differences between the old and the new internet protocol.

Lastly, middleware-related information as well as testbeds are discussed. This information will be extended in the next deliverable, giving a profound overview on what middleware should be chosen in order to reach the goal of remote instrumentation.

References

[FOKT01] I. Foster, C. Kesselman, S. Tuecke: "The Aantomy of the Grid: Enabling Scalable Virtual Organisations", Intl. Journal of Supercomputing Applications, Vol. 15, No. 3, 2001.

[FIG06] Sergi Figuerola, UCLPv2 Update, 16th Global Grid Forum meeting, http://www.ggf.org/GGF16/materials/plenary/Tuesday/GHPN_Networking_enhancements_for_grids/Sergi_Figuerola.ppt

[BSA] Bill St. Arnaud, UCLP Roadmap Figures, http://www.uclp.ca/files/uclpv2/UCLP%20Roadmap%20Figures-v4-BilStArnaud-2005_05_11.ppt

[UCLPGUIDE] UCLPv2 User Guide, http://www.uclp.ca/uclpv2/documents/help/uclpv2.0.2/

[UCLPRDM] Bill St. Arnaud, UCLP Roadmap for creating User Controlled and Architected Networks using Service Oriented Architecture, http://www.uclp.ca/files/uclpv2/UCLP_Roadmap_v2.doc

[IECSCTP] http://www.iec.org/online/tutorials/sctp/topic01.html

[ESSSCTP] http://tdrwww.exp-math.uni-essen.de/inhalt/forschung/sctp_fb/

[MPISCTP] http://www.cs.ubc.ca/labs/dsg/mpi-sctp/

[DICKMPI] P. Dickens and W. Gropp, "Efficient communication across the internet in wide-area MPI". In PDPTA, Las Vegas, 2001

[KAMALSCTP] H. Kamal, B. Penoff, A. Wagner, "SCTP-based Middleware for MPI in Wide-Area Networks", Department of Computer Science, University of British Columbia

[BURNSLAM] G. Burns, R. Daoud and J. Vaigl. LAM: An Open Cluster Environment for MPI. In Supercomputing Symposium 1994 (Toronto, Canada, June 1994)

[GROPPMPI] W. Gropp, E. Lusk, N. Doss and A. Skjellum. High-performance, portable implementation of the MPI message passing interface standard. Parallel Computing 22, 6 (Sept. 1996), 789–828

[KARMPICH] Karonis, N. T., Toonen, B. R., and Foster, I. T. MPICH-G2: A grid-enabled implementation of the message passing interface. CoRR cs.DC/0206040 (2002)

[KAMVS] Humaira Kamal, Brad Penoff, Alan Wagner, "SCTP versus TCP for MPI", Department of Computer Science, University of British Columbia, Vancouver, BC

[CARTCP] Cardwell, N., Savage, S., and Anderson, T. Modeling TCP latency. In INFOCOM (2000), pp. 1742–1751

[MATW100] Mathis, M., Heffner, J., and Reddy, R. Web100:Extended TCP instrumentation for research, education and diagnosis. SIGCOMM Comput. Commun. Rev. 33, 3 (2003), 69–79

[OTGMHOM] Otgonchimeg.B Jin Tae Kim Seung Yong Lee Youngmi Kwon, "Performance Improvement of Grid Web Services based on Multi Homing Transport Layer", Dept. of InfoComm Engineering, Chungnam National University

[UEDPON] H. Ueda et al., "Deployment Status and Common Technical Specifications for a B-PON System," IEEE Commun. Mag., vol.39, Dec. 2001, pp. 134–41.

[ITUPSV] ITU-T Recs. G.983.1–G.983.10, SG 15, "Broadband Passive Optical Network."

[FSANWEB] http://www.fsanweb.org

[ITUGPON] ITU-T G.984.1, SG 15, "Gigabit-Capable Passive Optical Networks (GPON): General Characteristics," Mar. 2003.

[ITUTCONV] ITU-T G.984.3, SG 15, "Gigabit-Capable Passive Optical Networks (G-PON): Transmission Convergence Layer Specification," July 2005.

[ITUPMD] ITU-T G.984.2, SG 15, "Gigabit-Capable Passive Optical Networks (GPON): Physical Media Dependent (PMD) Layer Specification," Mar. 2003.

[ITUONT] ITU-T G.984.4, SG 15, "Gigabit-Capable Passive Optical Networks (G-PON): ONT Management and Control Interface Specification," June 2005.

[MAESTDT] Y. Maeda, "Standardization Trends of Next Generation Optical Access Systems", IEEE Communications Magazine, vol. 44, no. 10, Oct. 2006.

[CAUGPON] A. Cauvin, A. Tofanelli, J. Lorentzen, et al. "Common technical specification of the G-PON system among major worldwide access carriers", IEEE Communications Magazine, vol. 44, no. 10, Oct. 2006

[PAPSUR] D. Papadimitrou, "Survey on GMPLS implementations", IETF – ISOC, available on-line http://www.isoc-gfsi.org/gfsi/reunions/docs/IETF2002/Papadimitriou/sld032.htm

[CHEWEB] CHEETAH: http://www-ee.engr.ccny.cuny.edu/wwwb/web/ibrahim/cheetah.htm

[DRAWEB] DRAGON: http://dragon.east.isi.edu/twiki/bin/view/Main/WebHome

[WESAKO] S. Wesner, J. Jähnert, M. A. T. Escudero, "Mobile Collaborative Business Grids - A short overview of the Akogrimo Project", Akogrimo White Paper, available from http://www.akogrimo.org/download/White_Papers_and_Publications/Akogrimo_WhitePaper_Overview.pdf

[OBS1] http://www.ikr.uni-stuttgart.de/~gauger/BurstSwitching/HTML/

[OBS2] Optical Burst Switching (OBS) - A New Paradigm for an Optical Internet C. Qiao (Dept of CSE), M. Yoo (Dept of EE) Lab for Advanced Network Design, Evaluation and Research (LANDER) University at Buffalo

[OBS3] Time Sliced Optical Burst Switching Jeyashankher Ramamirtham, Jonathan Turner, Computer Science and Engineering Department, Washington University in St. Louis, St. Louis, MO-63130

[OBS4] JumpStart: A Just-in-Time Signaling Architecture for WDM Burst-Switched Networks, I. Baldine, Dan Stevenson - MCNC ANR, Research Triangle Park, NC, USA; Harry G. Perros, George N. Rouskas - NCSU Department of Computer Science, Raleigh NC, USA

[GTCP1] http://www.globus.org/toolkit/docs/4.0/techpreview/gtcp/GTCPFacts.html

[GTCP2] http://www-fp.grids-center.org/news/pdf/nees-hpdc-final.pdf

[GTCP3] NTCP: A Grid Service for Remote Control Systems L. Pearlman, M. D'Arcy, C. Kesselman - USC Information Sciences Institute, Marina del Rey, CA, P. Plaszczak - Argonne National Laboratory, Argonne, IL

[VLABWEB] Virtual Laboratory Website http://vlab.psnc.pl

[LAWDYNM] Lawenda, M., Meyer, N., Rajtar, T., Okon, M.et al.: Workflow with Dynamic Measurement Scenarios in the Virtual Laboratory. 6th CARNet Users Conference, ISBN 953-6802-04-X, Zagreb, Croatia, 2004

[LAWDYNW] Lawenda, M., Meyer, N., Rajtar, T., Okon,et al.: Dynamic Measurement Scenarios in the Virtual Laboratory system. 5th IEEE/ACM International Workshop on Grid

Computing, IEEE Computer Society Order Number P2256, ISBN 0-7695-2256-4, ISSN 1550-5510, pp. 355-359, Pittsburgh, USA, 2004

[REKDSL] Rek, P., Kopec, M., Gdaniec, Z., Popenda, L., et al.: Digital Science Library for Nuclear Magnetic Resonance Spectroscopy. The 4th Cracow Grid Workshop, ISBN 83-915141-4-5, pp. 404-411, Cracow, Poland, December 12-15, 2004

[PASWEB] Institute of Bioorganic Chemistry PAS, Poznań: http://www.ibch.poznan.pl

[RADWEB] Radio Astronomy Departament NCU Toruń: http://www.astro.uni.torun.pl

[PEEWEB] Faculty of Process and Environmental Engineering of the Technical University of Łódź http://wipos.p.lodz.pl/index.php?lang=en

[LAWGENC] Lawenda, M., Meyer, N., Rajtar, T., Okon, M., et.al: General Conception of the Virtual Laboratory. International Conference on Computational Science 2004, LNCS 3038, pp. 1013-1016, Cracow, Poland, June 6-9, 2004

[GLOBWEB] Globus website http://www.globus.org

[GRMSWEB] GRMS - WP9 Resource Management - GridLab Project http://www.gridlab.org/WorkPackages/wp-9/

[GRIDGE] http://www.gridge.org

[G709A] Interfaces for the Optical Transport Network (OTN), Recommendation G.709/Y.1331 (03/03): http://www.itu.int/rec/T-REC-G.709-200303-I/en/

[G709B] The G.709 Optical Transport Network - An Overview: http://documents.exfo.com/appnotes/anote153-ang.pdf

[G709C] A G.709 Optical Transport Network Tutorial: http://www.innocor.com/pdf_files/g709_tutorial.pdf

[DAVDCL] F. Davoli, S. Palazzo, S. Zappatore, Eds., Distributed Cooperative Laboratories: Networking, Instrumentation, and Measurements, Springer, New York, NY, 2006; ISBN: 978-0-387-29811-5.

[BENATE] L. Benetazzo, M. Bertocco, C. Narduzzi, "Networking automatic test equipment environments," IEEE Instrumentation and Measurement Magazine, vol.8, no.1, pp.16-21, March 2005.

[BERARM] M. Bertocco, "Architectures for remote measurements," in F. Davoli, S. Palazzo, S. Zappatore, Eds., Distributed Cooperative Laboratories: Networking, Instrumentation, and Measurements, Springer, New York, NY, 2006, pp. 349-362.

[BENSCN] L. Benetazzo, M. Bertocco, C. Narduzzi, "Self-configuring measurement networks," Metrology and Measurement Systems, 2007, in press.

[BERMSA] M. Bertocco, A. Sona, "On the measurement via a superheterodyne spectrum analyzer," IEEE Transactions on Instrumentation and Measurement, vol. 55, no. 5, pp. 1494-1501, Oct. 2006.

[DAVLN] F. Davoli, G. Spanò, S. Vignola, S. Zappatore, "LABNET: towards remote laboratories with unified access," IEEE Transactions on Instrumentation and Measurement, vol. 55, no. 5, pp. 1551-1558, Oct. 2006.

[ANDTM] O. Andrisano, A. Conti, D. Dardari, A. Roversi, "Telemeasurements and circuits remote configuration through heterogeneous networks: Characterization of communications systems," IEEE Transactions on Instrumentation and Measurement, vol. 55, no. 3, pp. 744-753, June 2006.

[BAGSLI] A. Bagnasco, M. Chirico, A.M. Scapolla, "A new and open model to share laboratories in the Internet," IEEE Transactions on Instrumentation and Measurement, vol. 54, no. 3, pp.: 1111-1117, June 2005.

[CARMMM] A. Carullo, Metrological management of large-scale measuring systems," IEEE Transactions on Instrumentation and Measurement, vol. 55, no. 2, pp. 471-476, April 2006

[CARTAT] A. Carullo, M. Parvis, S. Corbellini, "Traceability assurance for tele-measurements," in F. Davoli, S. Palazzo, S. Zappatore, Eds., Distributed Cooperative Laboratories: Networking, Instrumentation, and Measurements, Springer, New York, NY, 2006, pp. 367-372.

[CACUEC] V. Cacciatore, A. Carullo, M. Parvis, A. Vallan, "Uncertainty effects of data compression in measurement applications," Proc. IEEE Instrumentation and Measurement Technology Conf. (IMTC 2006), Sorrento, Italy, DVD, pp. 1462-1467.

[QUEEEW] T. Q. S. Quek, D. Dardari, M. Z. Win, "Energy efficiency of dense wireless sensor networks: To cooperate or not to cooperate," IEEE Journal on Selected Areas in Communications, vol. 25, pp. 1-12, 2007, in press.

[DARMEW] D. Dardari, A. Conti, C. Buratti, R. Verdone, "Mathematical evaluation of environmental monitoring estimation error through energy-efficient wireless sensor networks," IEEE Transactions on Mobile Computing, 2007, in press.

[FRALDPC] M. Franceschini, G. Ferrari, R. Raheli, "Does the performance of LDPC codes depend on the channel?," IEEE Transactions on Communications, vol. 54, no. 12, pp. 2129-2132, Dec. 2006.

[FERDDCL] G. Ferrari, R. Pagliari, "Decentralized detection in sensor networks with noisy communication links," in F. Davoli, S. Palazzo, S. Zappatore, Eds., Distributed Cooperative Laboratories: Networking, Instrumentation, and Measurements, Springer, New York, NY, 2006, pp. 233-249.

[FERSDB] G. Ferrari, M. Martalò, "Sensor networks with decentralized binary detection: clustering and lifetime," Proc. Int. Workshop on Wireless Ad-hoc Networks 2006 (IWWAN'06), New York, NY, July 2006.

[CAVIRV] L. Caviglione, F. Davoli, P. Molini, S. Zappatore, "Integrating real and virtual instruments in the Grid: architectural design choices and performance," International Journal of Communication Systems (to appear).

[BAGEMI] A. Bagnasco, A. Poggi, G. Parodi, A. M. Scapolla, "Exposing measurement instruments as Grid Services," in F. Davoli, S. Palazzo, S. Zappatore, Eds., Distributed Cooperative Laboratories: Networking, Instrumentation, and Measurements, Springer, New York, NY, 2006, pp.: 321-329.

[SNXWEB] http://www.satnex.org

[EVASNXO] B. G. Evans, "SatNEx - A European Network of Excellence in satellite communications", International Journal of Satellite Communications and Networking, vol. 23, no. 5, p. 263, Sept./Oct. 2005.

[WERSNXO] M. Werner, A. Donner, E. Lutz, R. Sheriff, F. Hu, R. Rumeau, H. Brandt, G. Maral, M. Bousquet, B.G. Evans, G. Corazza, "SatNEx - the European Satellite Communications Network of Excellence", Proc. 59th IEEE Vehicular Technology Conference (VTC 2004-Spring), vol. 5, p. 2842, Milan, Italy, May 2004.

[SHETRAI] R. E. Sheriff, Y. F. Hu, P. M. L. Chan, M. Bousquet, G. E. Corazza, A. Donner, A. Vanelli-Coralli, M. Werner, "SatNEx: A Network of Excellence providing training in satellite communications", Proc. 61st Vehicular Technology Conference (VTC 2005-Spring), Stockholm, Sweden, vol. 4, pp. 2668–2672, May 2005.

[BERIESS] L. Berruti, F. Davoli, S. Vignola, S. Zappatore, "Interconnection of laboratory equipment via satellite and space links: Investigating the performance of software platforms for the management of measurement instrumentation", Proc. 2006 International Tyrrhenian Workshop on Digital Communications, Ponza, Italy, Sept. 2006, in press.

[DOB02] Dobbelaere, P. et al., "Digital MEMS for Optical Switching". IEEE Communications Magazine. March, 2002.

[GUI98] GUILEMOT, C. et al. "Transparent Optical Packet Switching: The European ACTS KEOPS Project Approach", IEEE Journal Lightwave Tech., vol 16, No. 12. Deecember, 1998.

[KAO66] K. C. Kao and G. A. Hockham, "Dielectric surface waveguides for optical frequencies", Proc. IEE, vol. 113, pp. 1151–1158, 1966

[MAS93] MASETTI, F. et al., "Fiber Delay Lines Optical Buffer for ATM Photonic Switching Applications", Proc. IEEE INFOCOM, vol. 3, pp. 935–42. March, 1993

[MID00] J. E. Midwinter, "The Start of Optical Fiber Communications as Seen from a U.K. Perspective", IEEE Journal on Selected Topics in Quantum Electronics, vol. 6, no. 6, p. 1307 a 1311, November/December, 2000

[MUK00] B. Mukherjee, "WDM Optical Communication Networks: Progress and Challenges", IEEE Journal on SelectedAreas in Communications, vol. 18, no. 10, October, 2000

[MUR02] C. S. R. Murthy, M. Gurusamy, "WDM Optical Networks: Concepts, Design and Algorithms", Prentice-Hall PTR, 2002

[NEG04] Kees Neggers, "Next Generation research networking in the Netherlands", SC2004, Pittsburgh, USA, 12 November 2004, available at www.glif.is/publications/presentations/20041112KN_SC04_glifpanel.ppt

[NET06] About Netherlight, available at www.netherlight.net/info/about/introduction.jsp

[RAM02] R. Ramaswami, K. Sivarajan, "Optical Networks: A Practical Perspective", 2a ed., Morgan Kaufman, 2002

[SUM06] Rick Summerhill, "The HOPI Testbed and the new Internet2 Network", presentation at ONT3 Meeting, Tokyo, Japan, 8 September 2006, available at www.nren.nasa.gov/workshops/pdfs9/PanelC_HOPI-Summerhill.pdf

Contact Information

All authors affiliation:
Poznań Supercomputing and Networking Center
ul. Noskowskiego 10
61-704 Poznań, Poland

URL: http://www.man.poznan.pl
Tel. (+48 61) 858-20-00
Fax (+48 61) 852-59-54

Thomas Prokosch	thomas.prokosch @ gup.jku.at
Dieter Kranzlmüller	dk @ gup.jku.at
Constantinos Kotsokalis	ckotso @ admin.grnet.gr
Tasos Zafeiropoulos	tzafeir @ grnet.gr
Afrodite Sevasti	sevasti @ admin.grnet.gr
Michael Stanton	michael @ rnp.br
Luca Caviglione	luca.caviglione @ cnit.it
Davide Adami	davide.adami @ cnit.it
Davide Dardari	ddardari @ deis.unibo.it
Franco Davoli	franco.davoli @ cnit.it
Antonio-Blasco Bonito	blasco.bonito @ isti.cnr.it
Alberto Gotta	alberto.gotta @ isti.cnr.it
Damian Kaliszan	damian @ man.poznan.pl
Tomasz Rajtar	ritter @ man.poznan.pl
Romeo Ciobanu	rciobanu @ ee.tuiasi.ro
Cristina Schreiner	cschrein @ ee.tuiasi.ro
Lei Liang	l.liang @ surrey.ac.uk
Marcela Larenas Clerc	mlarenas @ reuna.cl
Zhili Sun	z.sun @ surrey.ac.uk