Project Number: **EP 29212**
Project Title: **QUANTUM**

Deliverable D6.1
Interim Report on the Results of the Quantum Test Programme

Deliverable Type:	PU - Public
Contractual Date:	31 October 1999
Actual Date:	03 November 1999
Work Package:	6 – Test Programme
Nature of Deliverable:	RE - Report

Authors:

Tiziana Ferrari	-	INFN/CNAF
Leon Gommans	-	University of Utrecht
Simon Leinen	-	SWITCH
Jan Novak –	-	DANTE
Simon Nybroe	-	Telebit Communications
Agnes Pouélé	-	DANTE
Roberto Sabatino	-	DANTE
Robert Stoy	-	University of Stuttgart
Jean Marc Uzé	-	RENATER

Abstract:

This deliverable describes the work carried out to date within the Quantum Test Programme. For each testing activity it reports on the objectives, the results obtained so far, and suggestions for future work together with implications for future operational services.

Keywords:
Differentiated Services, QoS, MPLS, IP Multicast, ATM, ATM signalling, IPv6

Table of Contents

1 Executive Summary

The goal of the Quantum Test Programme (QTP) is to evaluate emerging technologies with the aim of understanding how to implement operational services with them. Particular attention is paid to the provisioning of Quality of Service (QoS), but also to multicast (IP and ATM), IPv6 and ATM signalling. Other efforts are devoted to understanding and developing techniques in support of the above such as QoS monitoring, flow-based monitoring, route monitoring and policy control.

A joint DANTE/TERENA task force, TF-TANT, carries out the work.

Activity on the QTP started in November 1998 with a meeting in Cambridge, in which the areas of activity and the people responsible for them were identified. Subsequently, work was carried out to define in more detail the test proposal for each experiment, together with the finalisation of the participants in each activity and the definition of a test plan.

The activities in which substantial work has been performed are: MPLS, Differentiated Services and IP multicast.

The activities in which a limited amount of work has been done and some results are available are flow-based monitoring, route monitoring, IP over ATM, IPv6, QoS monitoring, Policy control and ATM Signalling.

The activities related to QoS monitoring are tightly related to those of Differentiated Services.

Activities that have been defined, but have not yet been developed are: RSVP to ATM signalling mapping, ATM multicasting, WDM and STM-4 concatenation issues.
The reasons for the diverse progress on activities are mainly due to prioritisation within the task-force: most of the members of TF-TANT take part in more than one activity, therefore there are simple time constraints on the possibility of performing all tasks at the same time. Other factors such as availability of test equipment also need to be taken into account.

Various equipment vendors have contributed to some of the work items. Cisco and Netcom Systems have contributed with the loan of equipment for the initial testing of MPLS. Cisco, IBM, Netcom Systems and Cabletron have contributed with the loan of equipment for the initial testing on diff-serv and QoS monitoring. Other vendors interested in the QTP are Torrent and Nortel. Telebit Communications is an Associated Partner of the Quantum project whose involvement is mainly for the QTP, with particular attention to IPv6 and RSVP.

Testing of MPLS was initially dedicated to a proprietary implementation by Cisco (tag switching), and has revealed that whilst the basic functionality is satisfactory, there are serious performance issues to be solved with regard to re-routing. The next phase of testing MPLS will involve implementations on other platforms and interoperability tests and the integration of MPLS with IP QoS.

Diff-serv testing has so far concentrated on functionality of the various mechanisms within routers that support IP QoS: CAR, WFQ, WRED, SCFQ. These mechanisms are being studied on Cisco and IBM platforms, whilst workstations and *smartbits* devices from Netcom Systems have been used as traffic generators and for performance measurements. So far the testing has revealed that the configuration of these features is far more complex than expected, and that minor changes to the configuration can have drastic effects on the overall performance.

Work on IP multicast has been very successful in that in October 1999 a pilot IP multicast service based on PIM-SM, MBGP and MSDP started on TEN-155 and there are plans to transform this into a fully supported operatiopnal service from January 2000. Future activities for IP multicast are related to the development of BGMP, and the mapping of IP multicast to ATM multicast.

The other activities do not yet have sufficient results to be able to draw conclusions.

2 Introduction

The principal activity within the QUANTUM project is the provision and support of the TEN-155 network, offering a service to the European research community which is state-of-the-art in terms of both performance and functionality. Because TEN-155 is an operational service which must maintain high levels of availability, it necessarily makes use of equipment and technologies which are tried and tested, even if the combination of functionality is in advance of that available in the market place for commercial services.

In order to maintain this leading edge position for their services, DANTE and the NRNs follow closely the new developments in the field. Within QUANTUM, the QUANTUM Test Programme (QTP) has the general objective of investigating and testing new technologies with a view to introducing them into the operational service as soon as it is feasible and effective to do so.

Overall responsibility for the QTP lies with the QUANTUM Policy Committee; day-to-day management is provided by DANTE in its role as Coordinating Partner. The work of the QTP is organised as a set of experiments carried out by a joint DANTE/TERENA Task Force, TF-TANT. In this way, the QTP is open to a wide range of organisations - individual universities, equipment suppliers, commercial service providers - in addition to the QUANTUM project participants themselves. The only qualification for participation in the QTP is that the participant - organisation or individual - must make a contribution in the form of manpower, expertise, equipment or service resource. The intention is that QTP results are published openly although it is also necessary to respect non-disclosure agreements that suppliers sometimes make a condition, for example when offering novel systems on loan.

The QTP is organised as a set of experiments which are essentially independent of each other although there are some inter-relationships between them. For example, the precise nature of tools developed to measure Quality of Service (QoS) parameters is closely related to the technologies which might be used to support QoS features in the network.

Participation in the QTP is voluntary in the sense that each participant decides for himself whether he can justify to his own organisation the use of the resource that he commits. Given the heavy demands placed on everyone working in this field, this by itself gives a guarantee of the validity of the experiments that are carried out.

Each QTP experiment is led by an experiment leader who, in many cases, is the individual who proposed the experiment initially. Most of the experimental work is conducted in a distributed fashion with partners working in diverse locations throughout Europe. Meetings of TF-TANT are organised roughly every two or three months and allow QTP participants to meet face-to-face, to review progress of their experiments, and to discuss items of common interest. A special one day workshop was held on 1 October 1999 for members of both the QPC and TF-TANT at which the progress of the QTP was presented and reviewed and the relevance of the technical activity to strategic planning of service developments was discussed.

The set of experiments which make up the QTP cover a wide range of topics but many are oriented towards the investigation of the ways in which premium or guaranteed QoS can be provided over network services which cross multiple management domains.

In all cases, work remains to be done before each experiment is complete. The remainder of this Interim Report is a compilation of reports provided by each of the experiment leaders. Each section describes the current status and the results so far of one of the QTP experiments. A Final Report on the QTP will be produced when QUANTUM is completed at the end of May 2000.

3 MPLS (Author: Jean-Marc Uzé – RENATER)

3.1 Problem Statement

Label-based switching is a technique based on an integration of layer 2 switching and layer 3 routing. The engineering of this new technology is completely driven by the IETF. When this technique was designed the first objective was to use in a more efficient way the performance of switching for high-speed networks with the scalability and flexibility of routing.

Any protocol should benefit from MPLS to carry data over any layer 2 technology. However, according to the needs of the NRN community we are focusing our activity only on the problem of carrying IP traffic with MPLS. As the TEN-155 backbone and many NRN backbones are based on ATM technology, our work focuses also on the specific architecture of MPLS carrying IP over ATM networks.

Significant progress has been made on the switching capacity of the chips used by routers, therefore the switching capability of MPLS is no longer the main attraction of MPLS. In fact, there are other advantages proposed by this new technology that are much more interesting for our needs:

- First of all, MPLS is a technology that allows the deployment of an IP backbone independently of the level 2 technologies used, therefore it simplifies considerably the configuration of such networks. A good example is to compare the configuration of an IP over ATM backbone such as TEN-155 with an equivalent one using MPLS over ATM. Moreover, it permits designing an end to end *link work layer* in the backbone, even by mixing level 2 technologies (ATM PVPC, ATM PVCC, SDH, Gigabit Ethernet, DWDM,…). This point is very important for the upgrade ability of a backbone.
- A second advantage of MPLS is to propose a solution to the future internet backbone architecture based on the Optical Internet model by providing necessary level 2 services today supplied by ATM and SDH. If we look more in depth at the situation of IP backbones today (see Fig. 3.1), we see that many are based on the IP over ATM over SDH model. In this model, all applications are carried over IP and the network capabilities are distributed in the different technologies:
  - the data is carried over SDH;
  - the fast restoration, when a link failure occurs, is provided by SDH, technology based on rings (that means always 2 paths available between two points on this ring) that are automatically reconfigured in few milliseconds if a problem occurs. This technology has proven its efficiency and its necessity;
  - the asymmetry capability: SDH is a technology that has been designed for telephony. As telephony traffic is based on full duplex and symmetric flows SDH does not have the capability to provide asymmetric links. Data traffic could benefit from asymmetric links. A good example for this is the US link that is generally used 3-4 times more in one direction than the other. ATM technology is capable of establishing asymmetric circuits;
  - currently traffic engineering is mainly provided by ATM technology since basic IP routing protocols can offer this possibility only in a very limited way. The goal of traffic engineering is to optimise the distribution of the flows on a given physical infrastructure;
  - currently QoS is based essentially on ATM technology, even if emerging technologies such as DiffServ are arising at IP level. There is no significant deployment of IP based technologies in a production environment. A good example of QoS provided by ATM is the MBS[1] service deployed on TEN-155. Equivalent services (QoS provided by ATM) exist in most of the NRNs

[1] Management Bandwidth Service

connected to TEN-155, in either production or pilot phase. This allows the provisioning of end to end QoS between European sites.

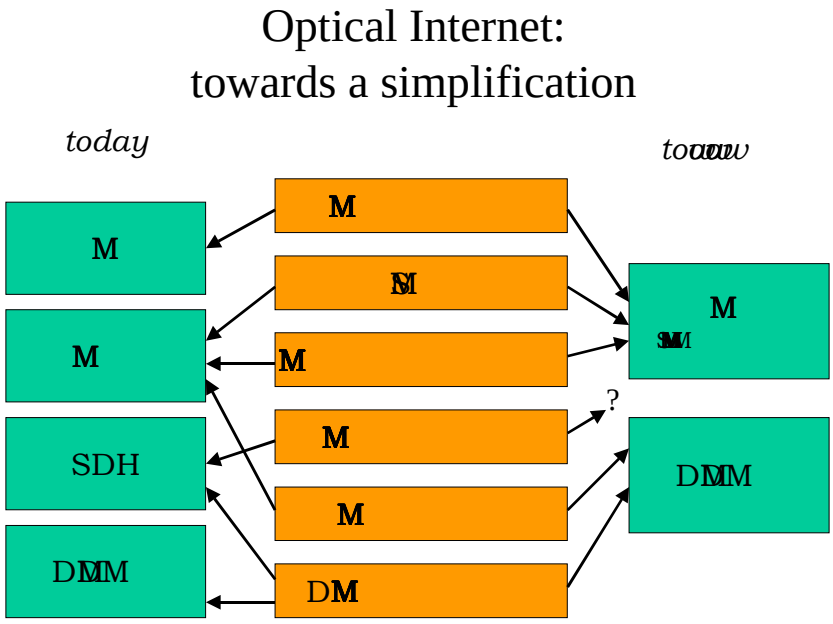


Fig. 3.1

- The third advantage of MPLS is to ensure the integration of new advanced internet services, such as:
  - IP QoS is today mainly based on diffserv technology and is probably the key to provide acceptable quality of service for specific applications;
  - The term VPN (Virtual Private Networks) is open to many interpretations, but in this context it consists of the isolation of a group of users on a backbone (and beyond) with QoS guarantees such as bandwidth, delay and jitters.
  - Multicast is currently performed at the IP level, but work is being done to take advantage of MPLS technology in order to optimise IP multicast.

In the future, the main model arising is the Optical Internet Model that goes towards a simplification by avoiding the ATM and SDH technologies. In this model IP packets will be carried by the light wavelengths inside fibre optics (DWDM technology). These coloured optical circuits will be switched by cross-connects and in a later stage by optical routers. In such a model the enhanced services and capabilities must be provided by only two layers: IP and DWDM. As asymmetry capability and data transport are achieved by DWDM, the challenge for IP in addition to carrying application traffic is to provide QoS and Traffic Engineering. These two goals correspond to the main objectives of MPLS. As far as the fast restoration feature is concerned it is today difficult to say how it will be supplied with acceptable performance.

MPLS has a lot of challenges before being deployed in production networks at a large scale:

- Stability and resilience;
- Scalability: MPLS is positioned at the backbone level. So it must be scalable in terms of number of the nodes and internet routes whilst maintaining enhanced services and capabilities;

- Management: MPLS must be manageable. Experience shows that it is difficult for ISPs to achieve Network Management in an easy way, specially for new technologies;
- Traffic Engineering: MPLS is sometimes presented as the new way to achieve traffic engineering. This attractive feature is not easy to carry out and represents one of the main challenges of MPLS;
- Fast re-routing: MPLS is designed to address fast re-routing (in the order of hundreds of milliseconds) but to date the performance observed is not satisfactory;
- "Ship in the night" mode support: MPLS must be deployed with high flexibility in operational networks by keeping available services provided by other underlying technologies, e.g. in the case of MPLS over ATM it may be desirable to keep the possibility to exploit native ATM features;
- Fast standardisation: the complexity associated with ATM has slowed down its development and deployment on a large scale. MPLS must learn from this experience and must be kept simple to be deployed rapidly: that could be the key of its success.

3.2 Objectives of the Experiment

The goal of the MPLS TF-TANT experiment is to gain practical experience of the new technology and study its applicability on a wide area backbone.

Therefore, the objectives of this experiment are to:

- Gain experience of the technology
- Survey existing implementations
- Evaluate advanced features, stability and performance
- Prove its applicability/scalability on a European ATM backbone
- Test the interoperability of available solutions

3.3 Outline Solution

The MPLS TF-TANT activity cannot be covered by one single experiment. The approach is to split the activity in different parts, depending mostly on the availability of products. In addition, joint activities with the Diffserv and VPN experiments are under study;

The various NRNs involved in the experiment provided the human resources. Participation is on a voluntary basis. Equipment vendors are expected to provide human expertise too.

As far as the hardware is concerned, all NRNs provide some resources (ATM switches, IP routers, test workstations…), but in each location more test equipment was needed to achieve a complete set-up. So far, the MPLS experiment has benefited from Cisco's loan (routers and ATM switches, or extension memory and cards) and from Netcom System's loan (Smartbit 200 Traffic generator/analyser).

As far as the network is concerned, the goal of the experiment is to perform tests on a wide area network with conditions as close as possible to an operational network. Laboratory tests will be performed only for those tests that cannot be performed on TEN-155, such as performance with high traffic loads. All other functionality tests are performed on TEN-155. To accomplish this, the TEN-155 MBS service is used in conjunction with similar services within the NRNs in order to establish dedicated connectivity between the participating sites. The MPLS experiment has been part of the beta phase of the TEN-155 MBS service.

The bandwidth needed for each link is between 1 and 2 Mbps, which is sufficient to test protocols and services.

3.4 Description of the experiment

A detailed description of the MPLS experiment and the results are available at the following URL: http://www.renater.fr/jmu/QTP/mpls-desc.html

The experiment has been divided into several parts.

The first part concerned the Cisco MPLS solution experiment, that was organised rapidly for several reasons:

- The software was available and had already interesting features (VPN). Also it could be used in the proposed architecture based on ATM tunnels (VPC) provided by TEN-155 and NRNs;
- Many Cisco devices were already available for experimentation on NRN sites. An extension loan was easy to set-up at a short notice;
- The task-force has good experience of Cisco products and already some MPLS experience on these products (from the TF-TEN experiment).

This first part was performed in April/May/June 1999 and is described in this report. More information on the next steps of the MPLS activity is given at the end of the report.

3.5 People/Organisations involved

The MPLS activity is involving the following participants:

- ACONET (AT)
- CERN (CH)
- CESNET (CZ)
- DANTE (UK)
- DFN (DE)
- GARR (IT)
- GRNET (GR)
- KPN (NL)
- REDIRIS (ES)
- RENATER (FR)
- SURFNET (NL)
- University of Namur (BE)

Mainly for technical reasons some organisations listed above were not able to participate in the first part described in this report.

The participants having successfully participated to the first tests are ACONET (AT), CERN (CH), CESNET (CZ), DFN (DE), GARR (IT), GRNET (GR) and RENATER (FR).

3.6 Test Roadmap

The tests were scheduled in two phases:

- The first phase consisted of setting-up the test bed, evaluating the stability of the code, performing redundancy tests and measuring the recovery time when a link failure occurs.
- The second phase consisted of testing the new VPN feature available in the Early Field Test (EFT) Software Release.

3.7 Technical set-up

The test bed was composed of:

- The MPLS backbone, together with equipment loaded with standard software (and later with Early Field Test Software Release during the VPN phase).
- The directly connected devices (physically or through a LAN) to the MPLS backbone which are:
  - external switches and routers
  - test workstations (Sun and PC)
  - Smartbits 200 (ATM generator/analyser)
- The external routes announced to the MPLS backbone by the directly connected routers.

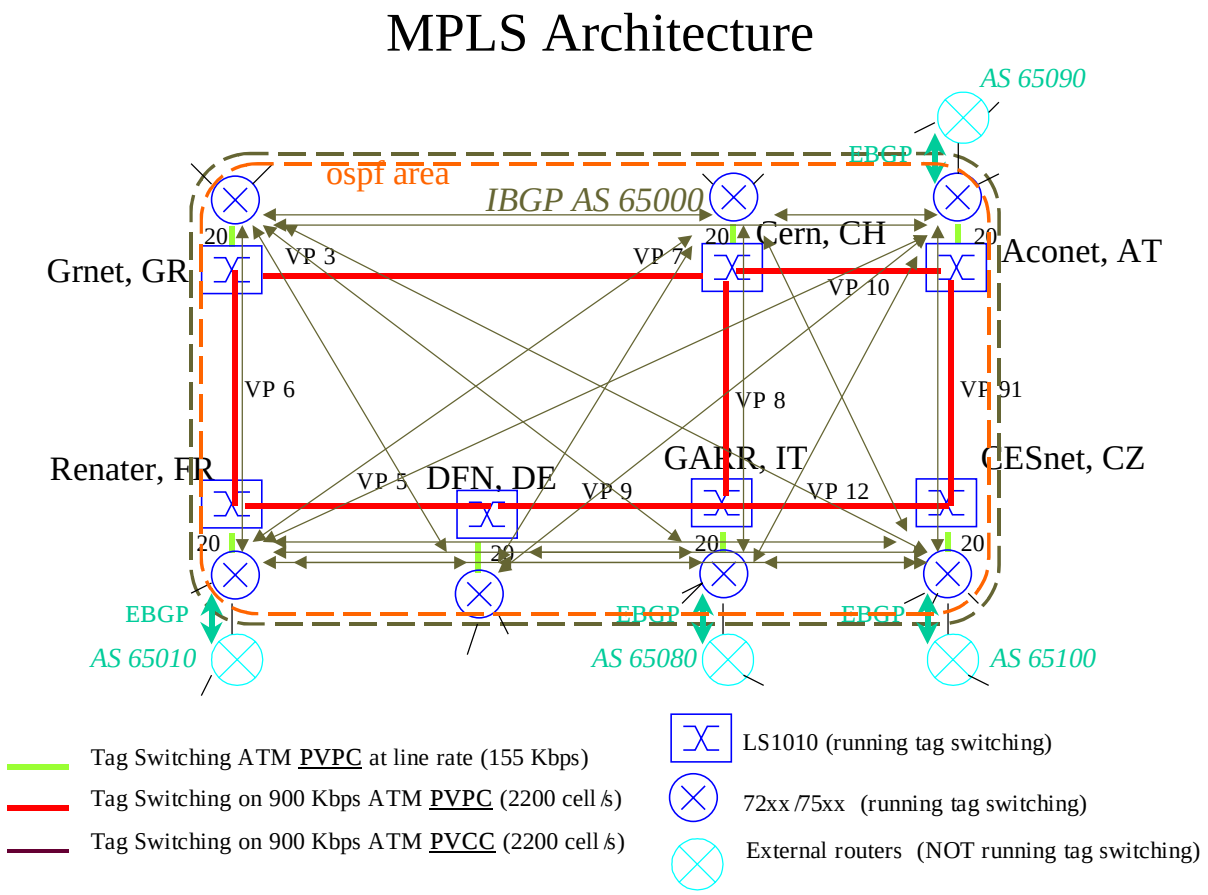


Fig. 3.2

The network infrastructure used to perform the tests was a portion of the European TEN-155 Network by means of its MBS service. In order to provide end to end connectivity, ATM resources from the NRNs were required. The overlay network (described in figure 8.2) consisted of ATM CBR permanent virtual circuits configured with PCR=2200 cells/sec (approximately 1 Mbps) on all links. The infrastructure consisted of a MPLS core network (corresponding to the area inside the dashed line) and a set of peripheral non-MPLS local networks in each country (the figure shows only the routers that are directly

connected to the MPLS network, not other devices such as workstations or Smartbits). Redundant links have been set-up to evaluate the re-routing time when a link failure occurs and when it is restored.

The MPLS cloud was composed of LightStream1010 switches and Cisco routers of the 7500 and 7200 series. All switches and routers were running commercial MPLS software recommended by Cisco. The ATM switches constitute the core of the MPLS domain (backbone), because they provide very high performance switching, while routers are deployed on the periphery, because they provide the high level routing features necessary to interconnect external networks. In each country a MPLS core node (switch) and a MPLS edge node (router) were set up. The routers were connected to their adjacent switch with a STM-1 link, whilst switches were connected through the NRNs and TEN-155 infrastructure. The MPLS protocol is entirely tunnelled in the ATM VP infrastructure therefore it was completely transparent to the ATM equipment on the operators side (NRNs + TEN-155).

MPLS uses an IP routing protocol (OSPF) to exchange all routes through the backbone, in order to set-up the corresponding LSP. Therefore on each MPLS node an IP loopback address was used by the LDP and an OSPF routing process was configured. LDP uses a dedicated control PVC that is automatically configured between adjacent MPLS nodes (during the initialisation phase), to exchange all IP routes and establish a full mesh of LSPs.

Outside the MPLS domain, in each country external networks were connected, i.e. routers (C750x and C720x), workstations, and Smartbit 200, which were connected to the MPLS domain through Fast Ethernet, Ethernet or ATM network interface cards.

To achieve scalability, the BGP sessions were set-up to exchanges external routes only between edge nodes, without redistributing them into the MPLS routing domain (OSPF area).

Switches in the core and routers at the edge of the MPLS cloud run the interior routing protocol, OSPF, that is used by the MPLS protocol.

All MPLS nodes constitute one unique backbone AS (AS65000) and edge nodes run exterior BGP sessions with external routers, whose networks are associated to their own AS number.

In addition, a complete mesh of internal BGP sessions is set-up between MPLS edge nodes. These iBGP sessions rely on LSPs established by MPLS. These iBGP sessions permit to exchange external routes between MPLS edge nodes.

In this architecture, all IP datagrams to destinations reachable through a given MPLS edge node, are forwarded through a unique LSP. For instance all the traffic between France (AS65010) and Italy (AS65080) is forwarded through a single LSP set-up between the French MPLS edge node (router) and the Italian one. The total number of TVCs doesn't depend on the routing table size, but on the number of physical nodes and on the links architecture if VC merging feature is used. The number of LSPs is proportional to n^2, where n is the number of physical nodes in the MPLS backbone.

3.8 Stability evaluation

The stability of the network was measured in different ways:

- stability of the code (crashes…)
- availability of the network and delay measurement
- QoS performance with the Smartbits 200

- basic performance tests with the workstations (ftp, mgen, netperf)

3.9 Redundancy test

The redundancy test consisted of switching off some VPs on the core backbone (interface shutdown on ATM switches) and measuring the re-routing time. The re-routing time was also measured when the links were restored.

3.10 VPN test

The VPN phase consisted of installing Cisco Early Field Test Software Release on the routers in order to use the new advanced VPN feature. This feature, based on BGP extensions and communities associated to the MPLS forwarding mechanism, permits to isolate sets of external networks connected to the backbone from other networks. It means that:

- Only external networks belonging to the same VPN exchange their routes.
- The MPLS edge node will forward packets only if the destination corresponds to networks belonging to the same VPN.

The current implementation of VPNs does not offer integrated QoS.

The test consisted of:
- setting-up different VPNs and evaluate the configuration complexity
- testing isolation between the VPN
- allocation of external networks to several VPNs (multi-homing)

3.11 Results of the experiment

3.11.1 Set-up observations

The full set-up of the network took more time than expected for various reasons:

- In some countries the installation of the equipment was delayed.
- The setting up of some end-to-end ATM VPCs took considerably more time than expected, due to liasion between different operators and/or end-equipment problems. For example, several problems were encountered when attempting to set up connectivity between SURFnet and GRnet, CERN and DFN. After several attempts, it was decided to continue the tests without SURFnet's participation. Investigation into this problem is being pursued by SURFnet.
- The test bed was big and involved many people working in different locations.

In all, three weeks were required to complete the setting up of the network.

The tests proved that MPLS configuration is very simple, even more if we compare it to IP over ATM VCC configuration.

As far as code stability is concerned, a very small number of crashes on the equipment (AT and CH) was observed during a four week trial period. In some cases it was necessary to clear the BGP sessions in order to establish connectivity after configuration changes.

3.11.2 Stability and performance results

The ping tests showed good results in terms of delay measurement. Sometimes the delay variation was of a few milliseconds, which is probably due the SDH reconfiguration somewhere in the network. It is quite difficult to investigate where a network problem occurs in such tests because:

- many carriers and ATM operators are involved in the end-to-end set-up,
- some ATM operators in the NRNs provide best effort service (ATM pilot service).

The stability measurement showed good results that confirm the stability of the code.

The performance tests are very difficult to interpret:

- The Smartbit 200 tests (full-meshed TCP sessions between all Smartbits) showed very bad results such as high packet drops, retransmission and sessions delay.
- Basic mgen tests (UDP) have confirmed bad results
- Basic pings, ftp and netperf tests showed acceptable results.

These contradictory results make it difficult to draw conclusions about the performance of MPLS. Further investigation has not yet been carried out because:

- Performance tests are not relevant with very low speed links. Intensive performance tests must be performed in laboratory with high capacity links between equipment.
- Performance tests are probably not evaluating the MPLS protocol itself but the switching capacity of the vendor's equipment. So such tests should be performed during an evaluation process that is not within the current scope of our activity.

3.11.3 Redundancy and re-routing time measurement results

This test was repeated many times in France by switching off the VPCs to DE and GR, and in Italy with the VPCs to CH, DE and CZ. These tests have also been repeated after loading the VPN software to compare the results.

We have measured the availability of the network at different levels:

- connectivity between core MPLS nodes (switches);
- connectivity between MPLS edge nodes (routers);
- connectivity between directly connected networks;
- connectivity with external routes.

The conclusions are:

- The re-routing time, when a link failure occurs, is about 4min 30sec. Traces have been taken at different times during the recovery process. This long delay does not seem due to the routing update of MPLS (performed by OSPF), but on LDP that took a long time to establish a new LSP. Better results were observed in very few cases (minimum about 40 seconds) just after the initialisation of the equipment. These results are unacceptable.
- The time needed by MPLS to re-route the traffic towards the initial path when the link is restored was also measured. The same re-routing time of 4min 30sec was observed. This is of course worse because the network is expected to recover its initial configuration without interruption.
- Note that these results were observed on the following specific architecture:
    - ATM switches in the core and IP routers on the edges (edge nodes).
    - IP addressing: unnumbered interfaces were used on the MPLS nodes.

- LS1010[2] + 75xx/72xx[3] platforms were used.
- The transport protocol used was ATM in a VP tunnelling architecture (with the labels corresponding to VC numbers).

It would be interesting to compare these results with an different architecture, e.g. MPLS with IP over SDH links.

3.11.4 VPN tests results

Four different VPNs in the backbone called Earth, Mercury, Jupiter and Mars (illustrated by figure 8.3) were set-up.

External networks were allocated to these different VPNs and the isolation was verified.

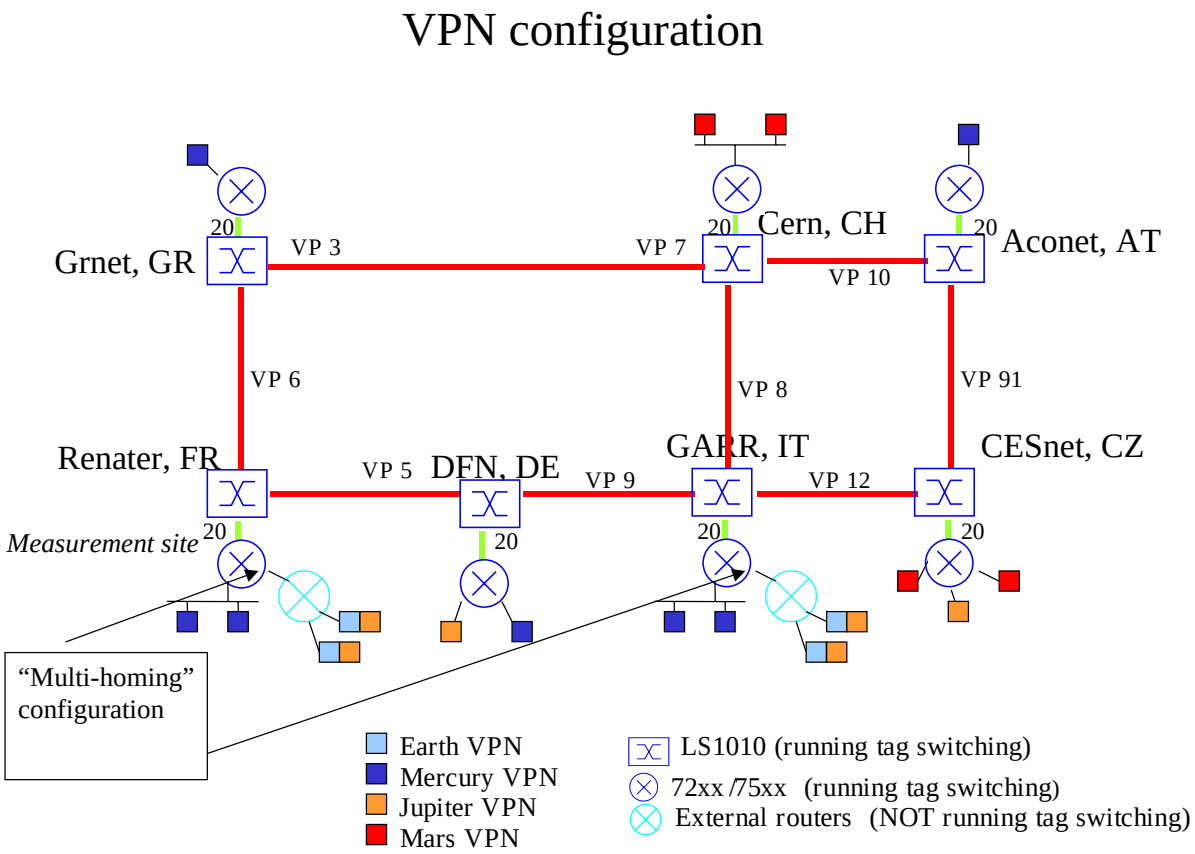


Fig. 3.3

The VPN functionality is based on BGP and extended communities.
The configuration steps on the MPLS edge nodes are:

- Configure a VRF for each VPN:

[2] with 12.0(1a)W5(5b) Commercial Software Release
[3] with 12.0(3)T Commercial Software Release

- allocate an extended community RD
- define import/export rules: for a strict Intranet VPN, you only import and export route from and to the same RD, e.g.:

```
ip vrf Mercury
rd 65000:2
route-target import 65000:2
route-target export 65000:2
```

- activate PE[4] peers for exchange of VPN information (BGP):

```
address-family vpnv4
```

  A full mesh of BGP sessions must be set-up between PE.

- activate on each PE a BGP session to each CE[5] within those VRF. You can also configure static routes:

```
address-family ipv4 vrf Mercury
neighbor 192.168.14.2 remote-as 65010
neighbor 192.168.14.2 activate
network 192.168.12.0
```

- set-up each interface (or sub-interface) as a VRF link to a CE router or Customer directly connected networks

```
interface Fastethernet 2/0.3
ip vrf forwarding Mercury
```

Note that the allocation of an external network to a VPN is not directly based on the IP address itself but on the interface (or sub-interface). This means that all IP networks connected to the PE (MPLS edge node) via the same interface (or sub-interface) belong to the same VPN.

The global configuration is easy to set-up.

Isolation of the networks belonging to different VPNs was observed.

The possibility of external networks to belong to several VPNs (multi-homing) was also tested. This configuration is very simple to achieve and completely transparent to the external routers/networks: it consists of one configuration command-line on each PE managing the external networks concerned.

It has been observed that, in order to achieve scalability, each PE keeps in its forwarding table only the routes needed by the customers connected to it. This means that in this architecture, the whole Internet routing table is not necessary in all PEs.

3.12 Difficulties Encountered - Lessons learned

The set-up time has been under-estimated for the following reasons:

- The debugging of ATM connections is very difficult because of the involvement of different management domains
- Hardware installation took a long time and the tests were not possible until the whole network was ready.

[4] Provider Edge (=MPLS edge node)
[5] Customer Edge (=external routers directly)

The first part of this activity has been successful overall. A laboratory platform set-up in parallel would be very helpful to debug some problems and investigate some points in more details, e.g.:

- investigate various configuration difficulties,
- repeat some tests by tuning the configuration (software, architecture),
- carry out performance tests at very high speeds,
- test MPLS with different level 2 technologies.

3.13 Future Activities

As far as the Cisco's MPLS solution is concerned, the following points could be investigated in more depth:

- RRR is an implementation of traffic engineering associated to constrained routing. Unfortunately it was not possible to test this new feature because it was not available on ATM interfaces at the time of the tests.
- MPLS QoS: the interaction between diffserv and MPLS is potentially a very interesting feature to evaluate. It was decided to await the conclusion of the initial diff-serv tests before undertaking this test activity
- The re-routing time problem has to be investigated in more depth.

One of the main goals is to perform interoperability tests between different vendor's implementations. Depending on availability of hardware and software, this is scheduled for 1Q2000 or 2Q2000. Traffic engineering, QoS and re-routing time can be investigated.

Network Management has not yet been part of the experiment. However, the simplicity of the configuration may increase the difficulty of the management. E.g., with IP over VCC, it is possible to manage the circuits that have been configured manually; with MPLS, it is more difficult because the network is automatically configured. Therefore Network Management is a real challenge for MPLS and must be investigated.

3.13.1 Implications for Future Services

From the beginning, the Task Force (TF-TEN and then TF-TANT) has followed the development of MPLS and performed experiments with this new and promising technology. The last tests have showed that MPLS is now a reality. Of course many technical points (performance, re-routing time and network management) still have to be investigated before a deployment in an operational network such as TEN-155 is possible.

MPLS is a technology that could be the strategic key for the next generation backbones, and this for two main reasons:

- MPLS is the technology that unifies in the best way all layer 2 technologies. It is important because the economic key of the next generation networks is probably on the layer 2 links (SDH, DWDM, Gigabit ethernet…). MPLS could be the technology that allows to modify the physical infrastructure with the highest flexibility and without changing the higher-level services.
- Traffic engineering associated to QoS is a way to optimise the cost of the network and to provide an equivalent service that is currently provided only by ATM technology. It potentially forms the basis of the future development of the TEN-155 MBS service.

4 Differentiated Services (Author: Tiziana Ferrari – INFN)

4.1 Problem Statement

During the last few years the international research community has devoted a lot of effort to develop and standardise a new IP oriented Quality of Service (QoS) architecture called differentiated services (diffserv). Differentiated services was born as a simple, pragmatic and scalable QoS solution as opposed to existing QoS protocols and architectures, such as the resource ReSerVation Protocol (RSVP) - an IP reservation protocol developed at the IETF - and ATM (Asynchronous Transfer Mode).

4.1.1 Diffserv and intserv

Diffserv scalability stems from the absence of signalling: Resources are provisioned statically through network dimensioning [6] and QoS guarantees apply to traffic aggregates rather than to micro flows.
Diffserv moves complexity from the core of the network to the edge, where several functions such as packet classification, marking and policing are placed.

In addition, unlike the Integrated Services architecture, in which a set of QoS classes is pre-defined, diffserv does not define services. Diffserv focuses on the standardisation of models for packet treatment and of the corresponding packet identification codes. Packet treatments are called *Per Hop Behaviours* (PHB). Only a small set of well-understood PHBs is under standardisation, while a large range of experimental *code-points* will be left undefined.

Another element, which contributes to the flexibility of diffserv, is interoperability. Diffserv networks can be built on top of a set of independent diffserv *domains*, each deploying an independent set of PHBs. Interoperability is achieved through specific functions at the boundaries between different domains like PHB mapping, traffic shaping and policing, and through *Service Level Specifications* (SLS).

The problem of end-to-end QoS support to the application can be solved through the combination of diffserv and intserv, which are complementary architectures to be deployed respectively at the edge and in the core.

4.1.2 Diffserv and ATM

Diffserv has some advantages compared to ATM.

First of all, it is an IP based architecture independent of layer 2 technologies, but still highly interoperable[7] since diffserv does not standardise PHB implementations.

In addition, it can be deployed to provide end-to-end services, while in ATM QoS cannot be supported to the micro flow unless both the sender and receiver have ATM native connections. Secondly, ATM needs signalling to dynamically establish connections, but end-to-end signalling protocols are subject to poor scalability, a problem that limits the deployment in production.

[6] Dynamic resource provisioning in diffserv capable networks is subject of current research at IETF. Bandwidth brokerage is the name of the architecture, which addresses this problem, and BB testing is part of the test programme.
[7] Diffserv PHBs can be mapped into ATM or, generally speaking, layer 2 QoS classes. This is an implementation issue and different diffserv domains can choose independent solutions.

Generally speaking, the deployment of IP QoS mechanisms in the national research networks is of great importance in particular for the resource allocation mechanisms they provide, a feature that is often a requirement even in high speed networks.

Through IP QoS based techniques a wide range of IP services can be deployed such as:

- fair deployment of expensive network resources (such as international network trunks) via traffic prioritisation
- virtual leased lines
- better than best effort services
- low delay and/or delay jitters - for research or mission critical applications -

4.2 Objectives of the Experiment

The main goals of the diffserv test programme are the following:

- To gain experience in diffserv network design and related issues such as network dimensioning and service level agreements
- To gain familiarity with QoS features available on different router platforms and with their implementation details
- To produce guidelines for the deployment of QoS features
- parameter tuning, QoS performance measurement, the analysis of end-to-end performance and the validation of the diffserv architecture
- the study of interoperability issues
- the definition, implementation and analysis of services relevant in the design of production networks
- the study of interoperability between IP QoS and ATM classes of services
- the study of the integration between the diffserv and the intserv architectures
- experimentation with Bandwidth Brokerage according to the ongoing developments in this area

4.3 Outline Solution

The first semester of 1999 was devoted to the definition of the test programme, which required:

- the study of the more recent developments in the diffserv working group at IETF;
- the survey of diffserv capable platforms and the analysis of interoperability issues;
- the design of the diffserv test network.

Contacts have been established with several vendors: Cisco, IBM, Netcom Systems, NORTEL and TORRENT, and three different loans from Cisco, IBM and Netcom Systems have been organised. Meetings with engineers from two of the vendors mentioned above were held to get technical information and help with the development of the test programme.

The test programme is divided into three areas: IP precedence testing, diffserv testing and interoperability testing between the intserv and the diffserv architecture.

The period from June 21st 1999 to August 31st was devoted to the following activities:

- diffserv network configuration;
- baseline performance testing through best-effort traffic;

- study of precedence QoS based mechanisms on Cisco routers when connected through a wide area network;
- analysis of diffserv features on IBM equipment;
- test of a research application based on object oriented distributed databases, when deployed in a QoS capable network. Testing of this application was carried out in the framework of the MONARC project at CERN (http://www.cern.ch/MONARC)
- configuration and deployment of dedicated equipment for traffic generation and performance measurement through the support of GPS synchronisation

4.4 Resources

4.4.1 Loans

Three different loans were made available to several test sites:

- CISCO loan: 1 C7200, 2 C7500, 1 LS1010 (the loan also deployed for MPLS testing) Hardware distributed to: GRNET, INFN and RedIRIS
- IBM loan: 5 IBM 2216, 5 IBM 2212.
  Hardware distributed to: CERN, GRNET, INFN, Uni. of Stuttgart and Uni. of Utrecht
- Netcom Systems: 3 SmartBits 200 with GPS kit (GPS antenna and GPS receiver).
  Hardware distributed to: INFN, Uni. of Twente and Uni. of Utrecht

4.4.2 Hardware available on site

In each test site dedicated test equipment was made available as listed below:

- ***test workstations***

Platforms: HP, Linux, Sun Solaris. Workstations with several types of network interface were available: ATM, Ethernet, FastEthernet and GigaEthernet.

- ***traffic generators***

1 SmartBits equipped with 2 10/100 Ethernet interfaces (for traffic generation) and 1 ctrl Ethernet interface (for the configuration of the apparatus), 1 GPS receiver and 1 GPS antenna.
The SmartBits are configured through the Windows application called SmartApplications (v. 2.22).

- ***ATM switches***
  1 per site
- 1 Cabletron Smart Switch Router (INFN, Uni. of Utrecht)

- ***routers***
- 1 C7500 or C7200 router per site
- IBM 2214 and 1 IBM 2216

4.4.3 3.1.3 Test partners

- ***Test partners***:

CS-FUNDP (BE), CERN, EPFL (CH), Switch (CH), DFN & GMD Fokus (DE), Uni. Stuttgart (DE), RENATER (FR), GRNET (GR), HUNGARNET (HU), GARR/ INFN (IT), IAT (IT), CTIT (NL), SURFnet (NL), Uni. of Utrecht (NL), ARNES (SI), RedIRIS (SP), Dante (UK)

- ***Test sites being part of the diffserv test network***:

CERN, DANTE, GRNET, INFN/GARR, RedIRIS, SWITCH, University of Stuttgart, University of Twente, University of Utrecht

4.5 Description of the Experiment

A detailed description of the test programme is available at the following URL: http://www.cnaf.infn.it/~ferrari/tfng/diffserv.html whilst a detailed description of the activities and results related to the testing conducted so far are available at the following URL: http://www.cnaf.infn.it/~ferrari/tfng/ds-test.html. For detailed information please refer to those pointers. The following paragraphs provide an overview of the activities and a summary of the first test results.

4.6 Technical set-up

The diffserv test-bed interconnects nine test sites as illustrated in figure 4.1. The wide area network is partially meshed and is based on CBR ATM VPs at 2 Mbps (ATM overhead included). On each VP a single PVC at full bandwidth is configured. The PVC is deployed as a point-to-point connection between two diffserv capable routers[8].

The testing activity conducted by the group did not require any ATM specific feature: ATM PVCs are simply deployed as point-to-point connections.

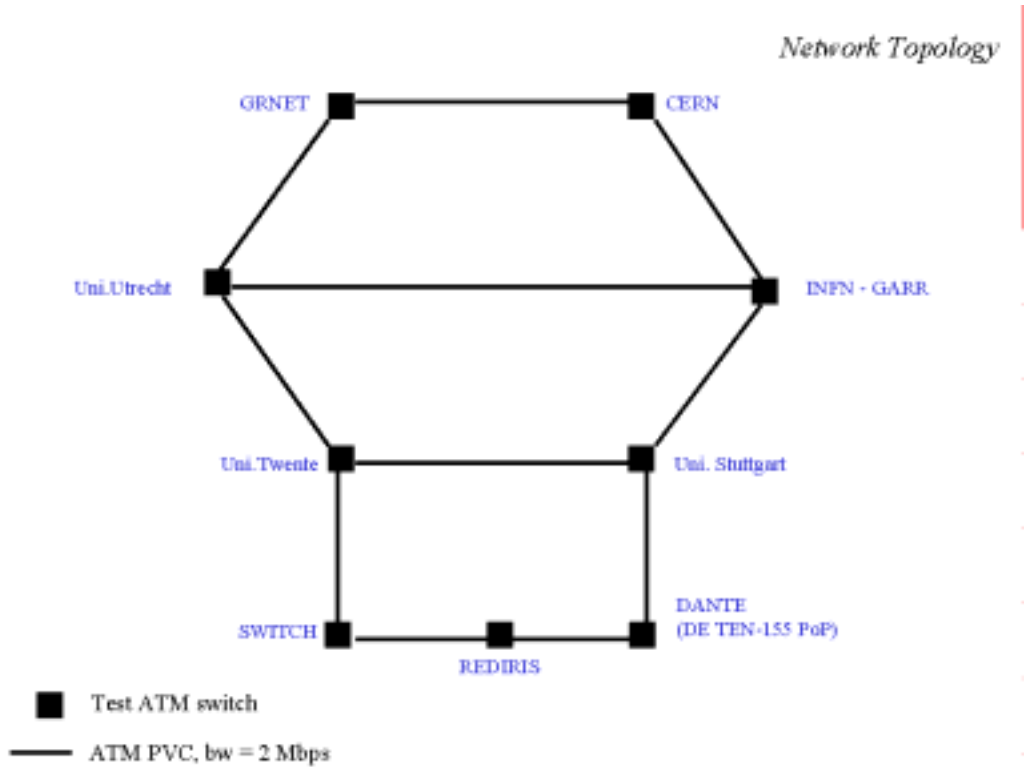


Fig. 4.1: Diffserv test network

[8] An amount of bandwidth like the one available in the test network does not give the possibility to run stress tests. For this reason, additional testing will be necessary in the local area network to try QoS features at higher speed.

4.6.1 Software

Cisco routers

- C7200: IOS 12.0(5.0.2)T1
- C7500: IOS 12.0(5.0.2)T1 on almost all the routers, IOS 12.0(5)S (DANTE)[9].

IBM routers

- IBM 2212: code version 3.3
- IBM 2216: code version 3.3
-

Netcom Systems SmartBits

- SmartBits 200: Firmware version 6.21
- 10/100 Ethernet: MPL-7710, beta build of 01 V1.06
- SmartApplications vers. 2.22

4.6.2 Addressing

Static routing has been deployed in order to avoid packet drop on control routing traffic when connections are tested under congestion.[10]

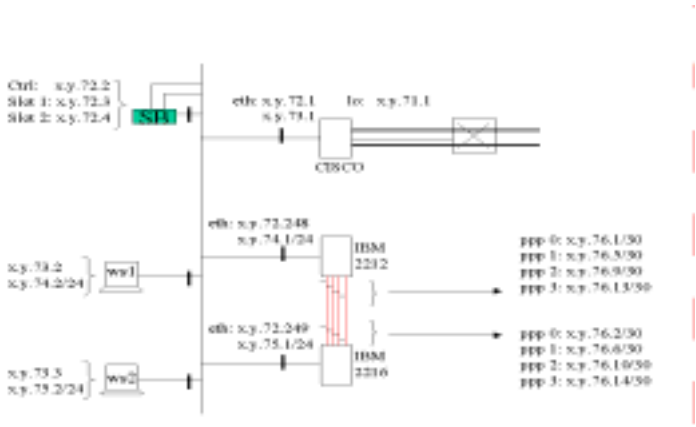


Fig. 4.2 Local addressing schemes

Addresses in the range [192.168.0.0, 192.168.255.0] have been deployed and a block of 10 class C network addresses was assigned to each site.
A homogeneous addressing scheme has been deployed in each site, according to figure 4.2. The global addressing scheme is presented in figure 4.3.

[9] This version did not support per ATM VC Class Based Weighted Fair Queuing (CB-WFQ), a necessary feature to enforce fair bandwidth utilisation on ATM PVCs. This feature is now supported by IOS 12.0.5-XE.
[10] Congestion is often needed to test traffic isolation among different classes of service. In a second phase, some QoS techniques themselves will be applied to control traffic. This type of configuration itself is an interesting example of traffic differentiation deployment.

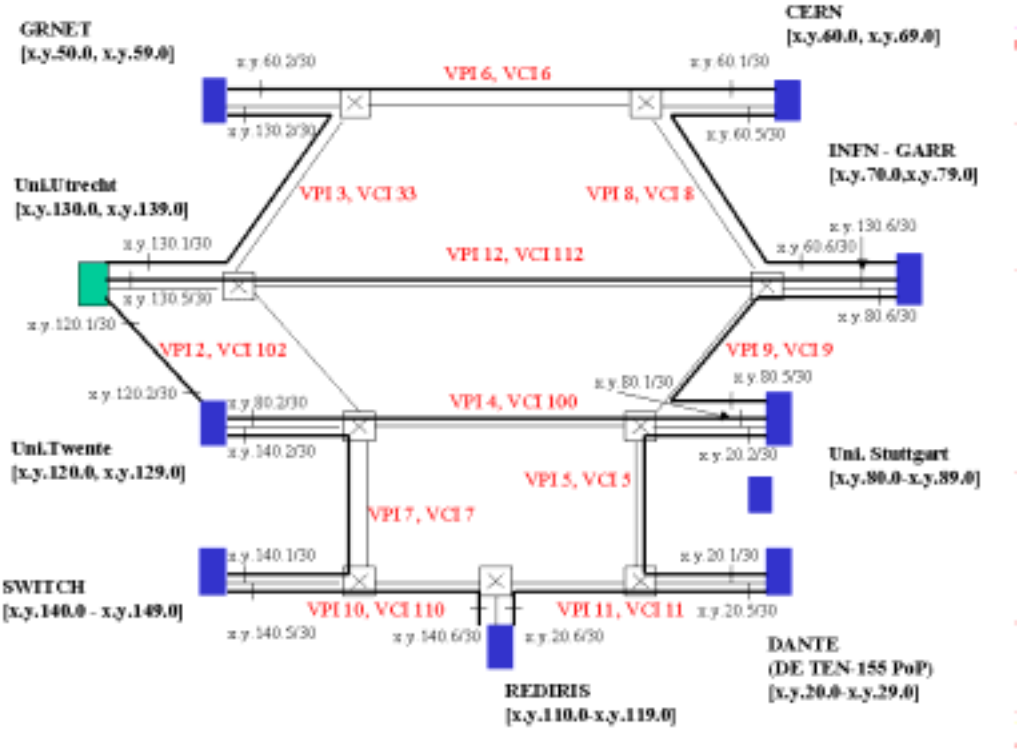


Fig. 4.3: Global addressing scheme

4.7 Planned Timetable and Work Items

PART 1: Jun 21st, Aug 31st 1999:

- network set-up
- test of basic CISCO QoS features:
- Committed Access Rate (CAR) – functionality, tuning of parameters
- Class Based Weighted Fair Queuing (CB-WFQ) – traffic isolation capability
- start of tests on IBM routers:
- SCFQ (premium, assured and best-effort traffic)
- TCP premium traffic and policing
- MONARC testing
- configuration of GPS based traffic generators for performance measurement
- definition of some services of interest for production networks

PART 2: Sep 1st, Dec 31st 1999

- Random Early Discard (RED) testing on CISCO equipment
- Completion of IBM testing
- Interoperability testing
- Configuration and test of services (study of QoS features deployment)
- Definition of a QoS performance measurement programme, measurement and performance analysis of end-to-end services
- Introduction of new diffserv capable platforms. Tentative list: Linux, NORTEL, Telebit and Torrent
- Test of mixed intserv and diffserv architectures (Phase 3 of the test programme)

- Bandwidth brokerage testing
- Policy deployment in diffserv networks
- interoperability between diffserv and MPLS
- test of a prototype of AAA server (Authentication Authorisation and Accounting)

The possibility of an additional test extension will be evaluated according to the results and problems encountered during part 2.

4.8 Results of the Experiment

4.8.1 Baseline testing

Some baseline testing was carried out in order to monitor the network performance with best effort traffic. RTT, TCP and UDP throughput figures were collected.

RTT

RTT is important for proper dimensioning of TCP socket buffer sizes, since the TCP window size is a function of the socket buffer. In addition, in case of large RTT, the stop-and-wait syndrome has to be avoided.

Given the high RTT values as reported below by Table 4.1, in order to optimise the performance of TCP applications large TCP socket buffer sizes need to be configured.

According to Table 4.1, some connections are not totally loss free: direct links between Twente and Utrecht and between Twente and SWITCH, and some other multiple hop connections are affected by packet loss.

From/to	CERN	DANTE	GRNET	INFN	RedIRIS	SWITCH	Uni Stutt.	Uni. of Twente
DANTE	116/116/116 0%	/						
GRNET	108/113/146 0%	221/221/222 2%	/					
INFN	51/51/52 0%	66/68/72 0%	156/156/157 3%	/				
RedIRIS	159/159/160 0%	46/46/46 0%	240/250/337 1%	110/110/110 0%	/			
SWITCH	227/227/229 0%	112/113/118 0%	153/157/209 0%	179/179/180 0%	67/67/68 0%	/		
Uni Stutt	100/101/107 0%	20/21/26 0%	219/431/653 1%	51/52/58 0%	65/65/67 1%	132/132/133 0%	/	
Uni. of Twente	188/188/188 0%	103/112/167 0%	110/110/112 0%	84/91/183 0%	104/104/108 0%	44/44/45 2%	NA	/
Uni. of Utrecht	170/170/176 0%	129/129/131 0%	91/110/212 0%	65/78/149 1%	124/142/170 0%	63/63/65 0%	115/127/268 2%	19/29/183 1%

Table 4.1: RTT between pairs of end-systems with packet sizes of 1420 bytes. NA = Non Available (link down during the test)

Baseline TCP throughput

Throughput of single and multiple TCP best-effort streams were measured in order to estimate the maximum performance achievable on each PVC. The maximum is approximately 1.6 Mbps, which corresponds to 2 Mbps when including ATM, IP and TCP overheads. The maximum TCP

rate as reported by counters in the router is 145 packets/sec, which corresponds to 1.74 Mbps (by including TCP and IP overhead). Traffic was generated using *netperf*.
Since netperf generates packets of 1500 bytes on average, the actual bandwidth utilisation can be estimated in the following way:

1500 bytes = 32 ATM cells
32 * 53 * 8 * 145 = 1.967 Mbps

The resulting capacity, 1.967 Mbps, corresponds to almost the whole line capacity.

TCP throughput depends on the number of parallel TCP connections and on the TCP window size. Table 4.2 shows the direct relationship with aggregate throughput and number of parallel TCP streams.

	tcp_cwnd_max (bytes)	**Avg RTT (msec)**	**Throughput 1 conn (Mbps)**	**Throughput 3 conn (Mbps)**	**Throughput 6 conn (Mbps)**
INFN -> CERN	262144	51	1.59	1.65	NA
INFN -> Uni Stutt	65535	52	1.62	1.65	NA
INFN -> DANTE	262144	68	1.55	1.65	NA
INFN -> Uni. Utrecht	262144	78	1.54	1.63	NA
INFN -> Uni. Twente	65535	91	1.20	1.64	NA
INFN -> RedIRIS	NA	110	1.40	1.55	NA
INFN -> GRNET	65535	156	1.30	1.54	1.62
INFN -> Switch	262144	179	1.15	1.50	1.60

Table 4.2: Relationship between RTT and best-effort TCP aggregate throughput for different numbers of TCP connections

Baseline UDP throughput
Full line-speed throughput can be achieved using 1 or more UDP streams. Full-line rate can be achieved when generating a stream of 202 or 203 packets/sec with a datagram size of 1000 bytes.

Bi-directional throughput
UDP
In our test-bed each PVC is allocated with 2 Mbps of capacity in each direction. The PVC capacity can be fully deployed in both directions at the same time, as the following results show.

When a two-way UDP stream between Uni. of Utrecht and INFN is deployed - by transmitting around 202 datagrams per second, where each datagram is 1000 bytes long - almost the full link capacity is consumed. In each direction the UDP stream rate estimated by the receiver is 1.6 Mbps, which is equivalent to the whole link capacity if the UDP, IP and ATM overhead is taken into account.

TCP
When two TCP streams are run in parallel, the performance achieved in one direction depends on the type of network interface. For senders with FastEthernet or Ethernet interfaces the TCP throughput can be

between 800 and 1000 kbps, whist for end-systems with native ATM interfaces the throughput is 1600 kbps.

4.9 Interim results

4.9.1 Committed Access Rate (CAR)

Description
Committed Access Rate is a CISCO feature that combines several functions:

multi-field (MF) packet classification: classes of traffic can be defined through extended access lists.

packet marking or re-marking (precedence marking): with marking, even traffic generated by non diffserv capable applications can be labelled with a given precedence. Re-marking can be fundamental when the router is located at the boundary between two diffserv domains and the current precedence value of a packet needs to be replaced by a different one. For this reason, re-marking enables interoperability between different diffserv domains.

policing: the upper threshold of the rate is defined and bound to a traffic class. Like marking, policing is an edge or boundary router function. It can be deployed to enforce a service level agreement, for example to limit a given class of traffic to a specified rate. Policing is important to enforce fair resource allocation.

The implementation of a policer requires the deployment of a traffic meter. On CISCO equipment, metering and policing are implemented through a token bucket.

Given an interface, CAR can be deployed on both input and output traffic on both physical and logical interfaces.

The command syntax is the following:

rate-limit [input | output] access-group <access-list> <rate> <normal_burst> <excess_burst> conform-action <action> exceed-action <action>

Packet marking
The marking function of CAR was tested, which resulted in correct packet identification. Conformant traffic can be associated to a given precedence, whilst traffic exceeding the contract can be marked accordingly, for example with a lower precedence value, or it can be dropped. Several other types of exceed actions can be chosen.

Marking was tested by applying CAR at the ingress interface of an edge router of the test network.
Another marking feature called *Policy Based Routing* could be deployed, but CAR is recommended since it gets better performance. When CAR is deployed just for its marking capability, then conform-action and exceed-action must be set to the same value.

Policing
Policing was tested with both UDP and TCP traffic.

In both cases, the policer works correctly and the results were as expected.

When choosing *drop* as action to be applied to exceeding traffic, the resulting UDP data rate - as reported

by the receiving end-system - is exactly equivalent to the threshold rate defined by CAR. Similarly, the TCP throughput is equivalent to the CAR rate, which means that metering and dropping are effective and that also TCP can adjust to the threshold configured in the router.
At any instant the router gives the possibility to check the current amount of conformant and exceeding traffic experienced by the router. The command line is the following:

show interface <int> rate-limit

Example of output:

```
qos#sh in faste0/0 rate
FastEthernet0/0 test LAN
  Input
    matches: access-group 112
      params:  1296000 bps, 82000 limit, 164000 extended limit
      conformed 549 packets, 830066 bytes; action: set-prec-transmit 5
      exceeded 0 packets, 0 bytes; action: drop
      last packet: 48ms ago, current burst: 3028 bytes
      last cleared 00:00:10 ago, conformed 656000 bps, exceeded 0 bps
```

Exceed Action
Two types of exceed action were analysed: *drop* and *set-precedence* according to the following scenario.

UDP traffic was deployed.

As illustrated in Figure 4.4, traffic is generated at INFN and terminated at DANTE or SWITCH Figure

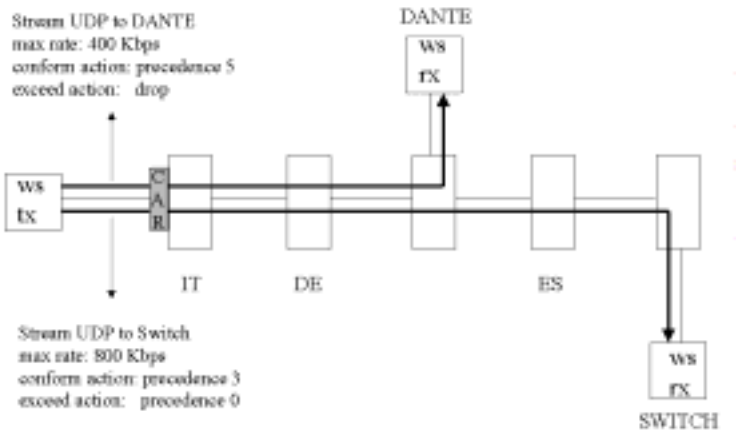


Fig. 4.4: Test of different exceed actions adopted by CAR

For traffic to DANTE the exceed action is *drop*, whilst it is *set precedence* for traffic to SWITCH. Excessive packets in the stream to SWITCH are transmitted with precedence 0.

Table 4.3 compares the throughput figures measured by the two receivers.

Throughput at the rx site (Mbps)	
SWITCH Exceed action = set precedence to 0	**DANTE** Exceed action = drop
1.200	0.386

Table 4.3: Effect of different CAR exceed actions on UDP traffic (application throughput not including overhead)

The table shows that throughput of rate limited traffic subject to *drop* (traffic to DANTE) equals the threshold specified by CAR. On the other hand, in the other case (for traffic to SWITCH) the UDP stream can grab more resources: its throughput exceeds 800 kbps and the stream uses the capacity unused by the stream to DANTE.

Normal and Excess Burst Sizes
The policer's behaviour is defined by two important configuration parameters, namely: normal burst *nb* and excess burst *eb*.

The policer's drop algorithm is described by the following formulas. Given a packet $pack_k$ of size s_k, its drop probability $p(pack_k)$, the current number of tokens available in the bucket $buck_k$ and the compounded debt $comp\_debt_k$ of the stream to which packet $pack_k$ belongs, then:

$$p(pack_k) = 0 \quad iff \quad (s_k \leq buck_k) \; or \; (buck_k = 0 \; and \; comp\_debt_k \leq eb)$$

$$p(pack_k) = 1 \quad iff \quad comp\_debt_k > eb$$

A detailed definition of compounded debt and a description of the algorithm are available in an appendix included as section 4.12 of this chapter.

If normal and exceed burst size are equal, then the policer's algorithm is comparable to a traditional token bucket with single bucket size, which is characterised by tail dropping in case of token unavailability.

The deployment of the excess burst size is important to minimise the effect of packet drops on TCP throughput. In fact, thanks to the drop probability, which increases gradually, significant packets drops are avoided and only isolated drops occur. This gives a TCP stream the possibility to gradually adapt to packet loss when the aggregate rate gets closer to the CAR rate threshold.

The effect of the normal and excess burst size on TCP performance[11] was quantified by running several tests with different buffer size configurations. The results show that for small normal bucket sizes the overall throughput is lower because of the limited burst tolerance of the policer. The recommended optimal normal burst size is a function of the maximum allowed data rate and the value can be set

[11] Normal and excess burst sizes are less relevant when CAR is applied to a UDP traffic class, since unlike TCP, UDP generates inelastic traffic which does not adapt in case of packet drop.

according to the following criteria:

$$nb = 0.5 * max\_rate$$
$$eb = 2 * nb$$

However, even with small normal bucket sizes, performance can be improved by increasing the excess burst size well above the normal burst size. In this way even if the fixed rate threshold is enforced, a real gain in performance is achieved.

The test shows that the tuning of normal and exceed burst sizes can be relevant when the rate-limited traffic class corresponds to a single micro-flow or is represented by the aggregation of few TCP streams. For any general-purpose service built on top of CAR the configuration of normal and exceed burst sizes indicated above is recommended. However, for application dependant services a different setting may be necessary, for example when CAR is applied to delay or delay variation sensitive applications. The impact of normal and exceed burst size on one-way delay is subject to future testing.

Parameter tuning is particularly important with few TCP streams. Table 4.5 shows that the impact of normal and exceed burst size is negligible when there is a small number of concurrent TCP flows. In this case the full line rate defined by CAR is achieved, but the nominal throughput of a single TCP connection is still low.

Table 4.4 reports on the throughput of a single TCP connection, whilst Table 4.5 is for 5 TCP connections. TCP streams are generated with netperf from the University of Twente test workstation 192.168.123.2 to the RUS test workstation 192.168.83.2.

The committed access rate configured in the following test is 1.296 Mbps, the exceed drop rate is *drop*. The following is an extract from the router configuration.

```
!
interface ATM1/0.7 point-to-point
 description connection to Ferrari (192.168.123.2)
 ip address 192.168.123.1 255.255.255.252
 no ip directed-broadcast
 rate-limit input access-group 130 1296000 48000 96000 conform-action  \
        set-prec-transmit 5 exceed-action     drop
 no ip route-cache
 no ip mroute-cache
 atm pvc 300 0 300 aal5snap
```

Throughput of 1 TCP connection (Mbps) (target throughput: 1.25 Mbps[12])					
Normal (bytes)	**Exceed (bytes)**				
	32000	48000	64000	96000	128000
32000	0.98	1.23	1.23	1.25	1.25
48000		1.09	1.21	1.25	125
64000			1.18	1.24	1.25
96000				1.24	1.25
128000					1.25

Table 4.4: Throughput of 1 TCP connection for increasing values of the normal and exceed burst size

Aggregate throughput of 5 concurrent TCP connection (Mbps) (target aggregate throughput: 1.25 Mbps)					
Normal (bytes)	**Exceed (bytes)**				
	32000	48000	64000	96000	128000
32000	1.26	1.26	1.25	1.26	1.25
48000		1.25	1.26	1.25	126
64000			1.25	1.27	1.25
96000				1.26	1.26
128000					1.25

Table 4.5: Throughput of 5 TCP connections for increasing values of the normal and exceed burst size

4.9.2 Class-Based Weighted Fair Queuing (CB-WFQ)

Weighted Fair Queuing is a well-known scheduling algorithm that can be deployed on a congested line to enforce fairness in resource allocation among different streams according to a configurable policy. With WFQ pre-emption of low bandwidth streams is avoided.

Each stream is provided with a dedicated queue, whose size is configurable, and queues are serviced at a rate proportional to the queue weight, which is a function of the bandwidth assigned to it.

With CB-WFQ classes can be defined through match criteria (access-lists), and a specified amount of bandwidth can be allocated to a given class.

CB-WFQ can be enabled per interface, per sub-interface and per ATM connection if its traffic class is VBR or ABR.

In addition, only 75% of the link capacity can be allocated through CB-WFQ, whilst the remaining part is distributed among the classes proportionally to their bandwidth share.

[12] Given a rate limit of 1296 Kbps and the exceed action equal to *drop*, the maximum throughput at the application layer is of 1.25 Mbps, which gives 1.3 Mbps by adding the overhead of the protocol stack.

For each class, CB-WFQ specifies the minimum amount of bandwidth that has to be allocated. This means that WFQ does not prevent any stream from getting more capacity when traffic is not congesting the line.

The following paragraphs, which present our CB-WFQ test results, refer to test scenarios only involving routers C7200, since at the time of the tests only a C7200 CB-WFQ capable version was available: IOS 12.0(5.0.2)T1.

A CB-WFQ capable version for C7500 is now available at the time of the writing: IOS 12.0(5) XE, which will be installed and tested on the C7500s in our network.

CB-WFQ configuration

The following is an example of CB-WFQ policy configuration as recommended by CISCO engineering. Bandwidth values are selected considering that the line rate is 2 Mbps:

```
policy-map wfq
 class wfq
    bandwidth 1300
    random-detect[13]

 class class-default
    bandwidth 200[14]

!
interface ATM1/0.8 point-to-point
 description to CERN (diffserv)
 bandwidth 2000
 ip address 192.168.60.6 255.255.255.252
 no ip directed-broadcast
 pvc 8/8
  service-policy output wfq
  vbr-nrt 2000 2000 1
  encapsulation aal5mux ip
 !
!
```

Buffer management with CB-WFQ

CB-WFQ takes effect only in case of congestion, as a consequence, its packet counters which can be monitored through the command:

```
show policy interface <interface>
```

are only updated when CB-WFQ is active. When attached to an ATM connection, CB-WFQ is active if and only if the ATM VC is congested.

“For any VC managed by the PA-A3, a per-VC tx queue is dedicated by the PA-A3's driver. These per-VC tx queues have a maximum depth. When this is reached, the VC is said to be "congested". At this

[13] The configuration of RED can help TCP but is not a strict requirement in the CB-WFQ configuration.

[14] Bandwidth values are chosen according to this rule: 1300+200=75% (2000). This effectively means that 1300/1500=87% of the bandwidth is assigned to the first WFQ class and 13% to the rest.

point, the PA-A3's driver will refuse to read any new packets delivered by the VIP-SRAM, hence causing the packets to be delayed in the VIP SRAM's CBWFQ system dedicated to this VC. Thus, NO drop should occur by design between the VIP per-VC CB-WFQ and the PA per-VC TxQ" [15] (C. Filsfils, CISCO).

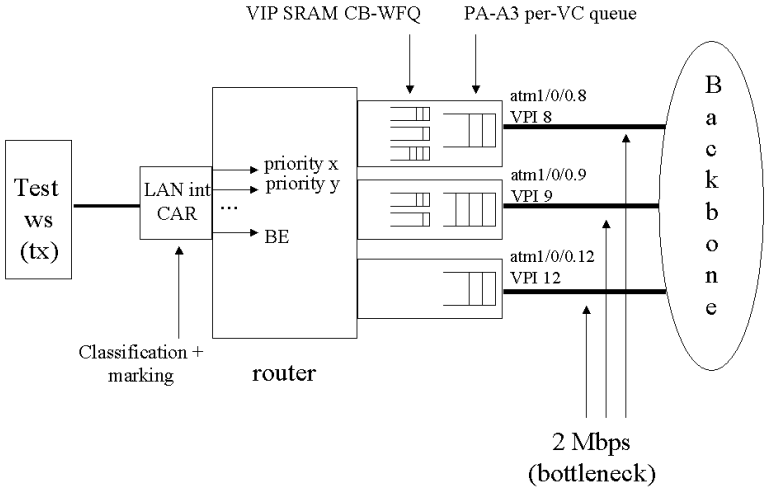
Testing of bandwidth management
The capability of a single TCP/UDP stream to get more bandwidth than the value stated by the CB-WFQ configuration in case of lack of congestion was verified. The result confirmed what is stated by the feature specification: without congestion, both TCP and UDP streams can achieve aggregated throughput figures well above the minimum stated by the class configuration.

Testing of traffic isolation
The capability of CB-WFQ to isolate traffic classes was tested to verify, for example, that high priority streams are protected from invasive best-effort traffic. Classification has been tested through several traffic pattern combinations:

1. UDP high priority traffic and UDP best-effort traffic
2. TCP high priority traffic and UDP best-effort traffic
3. TCP high priority traffic and TCP best-effort traffic

TCP high priority traffic was tested with both single micro-flows and TCP traffic aggregations.

The three test scenarios above were tested by deploying the marking via CAR on the edge router and by configuring CB-WFQ on the egress interface to the backbone, where classification was based on the precedence values set by CAR. CAR exceeding traffic was dropped in order to prevent it from getting



more bandwidth than specified by CB-WFQ. An example of this set-up is represented in figure 4.5.

Fig. 4.5: Example of CB-WFQ testing set-up

[15] This is monitored by the OutPktDrops variable, which is displayed through the command: show atm vc

The same model was applied to all the sites involved in this type of test.

CAR and CB-WFQ configurations are as follows:

```
class-map wfq-tcp
  match access-group 177
!
policy-map wfq
 class wfq
   bandwidth 1300
   random-detect
  class class-default
   bandwidth 200
!
interface FastEthernet0/0
  [..]
 rate-limit input access-group 140 1296000 26000 32000  \
          conform-action set-prec-transmit 5 exceed-action drop
  [..]
!
interface ATM1/0.9 point-to-point
 description to Uni. of Stuttgart (diffserv)
 [..]
 pvc 9/9
  service-policy output wfq
  vbr-nrt 2000 2000 1
  encapsulation aal5mux ip
 !
access-list 140 permit udp host 192.168.73.2  192.168.23.0   0.0.0.255
access-list 140 deny   ip any any
access-list 177 permit udp any any precedence critical
access-list 177 deny   ip any any
```

UDP high priority and UDP best effort traffic

The deployment of UDP traffic is quite useful to verify the basic functionality of features given its constant behaviour that is packet loss independent. In addition, the aggressiveness of UDP traffic when deployed as background traffic helps test the robustness of features for traffic isolation.

According to our tests, CAR and CB-WFQ seem to be an effective mechanism for traffic isolation.

Test A: exceed action = drop

In fact, two UDP streams were generated, each injecting traffic at 2 Mbps to the same output interface of the router, so that the ATM connection (at 2 Mbps) is the bottleneck.

The scenario consisted of two UDP classes, a high priority one (rate limited to a given amount of kbps, for example 1300 kbps) and a best-effort one (not rate limited). The high priority class is prevented from getting more than 1300 kbps by CAR, but it can still get the whole capacity assigned to it. This means that the best-effort background stream does not interfere with high priority traffic and perfect traffic isolation is achieved.

Test B: exceed action = transmit

If the CAR exceed action is *transmit* instead of *drop*, also the traffic to DANTE can get more than 400 kbps, despite having a lower priority than the traffic to SWITCH, and the excess capacity is shared between the two classes. The actual amount of excess capacity allocated to the high priority class depends on the aggressiveness of the best-effort stream, on the metering of CAR (normal burst and exceed burst sizes) and on the amount of bandwidth guaranteed to the default class.

Test C: exceed action = set precedence

In this test the CAR exceed action is *set-precedence 3* instead of *drop*, and CB-WFQ guarantees a minimum rate of 400 kbps to precedence 3 traffic according to the following configuration:

```
policy-map switch
  class switch-pr5  /* access-list 190 permit ip any 192.168.143.0 0.0.0.255 precedence critical
                       access-list 190 deny   ip any any */
   bandwidth 800
  class switch-pr3  /* access-list 191 permit ip any 192.168.143.0 0.0.0.255 precedence flash
                       access-list 191 deny   ip any any */

   bandwidth 400
```

In this test the expected aggregate UDP rate of traffic to SWITCH should be at least 800 + 400 = 1200 kbps. This is confirmed by the test, according to which the throughput - as measured by the UDP receiver - is 1.39 Mbps (overhead included). This figure is slightly bigger than 1.2 Mbps since both precedence 5 and precedence 3 traffic compete against best effort traffic for the allocation of the remaining part of bandwidth (800 kbps).

Test D: test of CB-WFQ with multiple classes

UDP was also deployed to try CB-WFQ in presence of multiple classes[16]. Traffic from INFN was divided into 4 different classes and assigned a different share of resources according to the following policy:

```
policy-map ch-sp-dante-de
  class switch-I              /* TCP traffic from 192.168.73.2 to SWITCH */
   bandwidth 800
  class rediris-li            /* TCP traffic from 192.168.73.2 to RedIRIS */
   bandwidth 400
  class dante-lli             /* TCP traffic from 192.168.73.2 to DANTE */
   bandwidth 200
  class us-llli               /* TCP traffic from 192.168.73.2 to Uni. of Stuttgart */
   bandwidth 100
```

When running 1 UDP stream for each class, each receiver gets at least as much as stated by the policy configuration. Receivers get even slightly more as a consequence of the distribution of the exceeding 500 kbps capacity amongst the 4 classes. When a bundle of best-effort TCP streams (10) is added, part of the excess bandwidth is allocated to them, so that the measured aggregate TCP throughput is approximately 400 kbps, each UDP streams gets its capacity share and the total data rate is equivalent to the PVC line rate.

[16] In this example CB-WFQ classification is based on the destination address. CB-WFQ is not necessarily based on precedence values, even if according to the diffserv architecture the most important way to differentiate packets in core routers is through the DSCP (DiffServ Code Point).

TCP high priority and UDP best effort traffic

We ran the same test as test A above with the only difference that the high priority traffic class matched TCP traffic instead of UDP traffic. Unlike UDP, TCP adapts to losses (caused by policing or CB-WFQ dropping) and a high percentage of packet loss can prevent TCP from getting its target rate.

We run two types of tests:

- Test A: with 1 TCP stream matching the high priority class
- Test B: with *n* TCP streams matching the high priority class ($n > 1$)
-

Test A

With just one TCP high priority stream, results all over the diffserv network are not consistent, since different throughput figures on the site hosting the transmitter were obtained. The test was repeated by deploying different transmitters in the following sites: CERN, GRNET, INFN, University of Twente and University of Utrecht, and for each site the test was repeated by alternatively selecting different neighbours. Results are illustrated in the following table:

Throughput of a single TCP high priority connection: CB-WFQ rate = 1300 Kbps		
Test site (tx)	**Neighbour (rx)**	**Throughput (Kbps)[17]**
CERN	GRNET	1250
	INFN	1210
GRNET	CERN	700
	Uni. of Utrecht	710
INFN	CERN	100
	Uni. of Stuttgart	100
	Uni. of Utrecht	100
Uni of Twente	SWITCH	NA
	Uni. of Stuttgart	880
	Uni. of Utrecht	880
Uni of Utrecht	GRNET	1110
	INFN	1160
	Uni. of Twente	1180

Table 4.6: Results of CAR and WFQ tests with 1 TCP high priority stream and UDP background traffic

As the table above shows, some sites (CERN and Uni. of Utrecht) obtain figures close to the full target rate assured by WFQ (1.25 Mbps are equivalent to 1.3 Mbps when including the overhead). On the other hand, GRNET and University of Twente get approximately ¾ of it. INFN is the worst case: throughput is only 100 Kbps, which is the same value achieved by TCP when it is run in parallel with UDP but without any guarantee from WFQ (best-effort).

Debugging of the problem of low performance

GRNET and Uni of Twente obtain similar results and in both cases the transmitter is connected via ATM. Similarly, both CERN and Uni of Utrecht have end-systems with an Ethernet connection (FastEthernet and GigaEthernet respectively).

[17] Application throughput, not including the overhead.

The lowest performance is measured at INFN, but the understanding of this inconsistency is rather difficult. Low performance occurs only in the direction INFN → CERN, whilst in the opposite direction perfect stream isolation is achieved. In addition, the laboratory at INFN is such that the layout is equivalent to the one deployed at Uni of Utrecht in terms of components: GigaEthernet interfaces, C7200 router and a Cabletron Smart Switch Router.

Low performance is not dependent on the type of operating system of the transmitter since both a Linux RedHat 2.2.5-15 workstation and a Solaris 2.7 workstation were deployed as transmitter with the same results. It can also be excluded that the problem is due to the type of network interface on the end-system, since several interface cards were tested: GigaEthernet, FastEthernet and Ethernet gave same results.
In addition, it can be excluded that low performance is due to a local problem in the Smart Switch Router according to the results of the following test.

Figure 4.6 illustrates the network configuration.

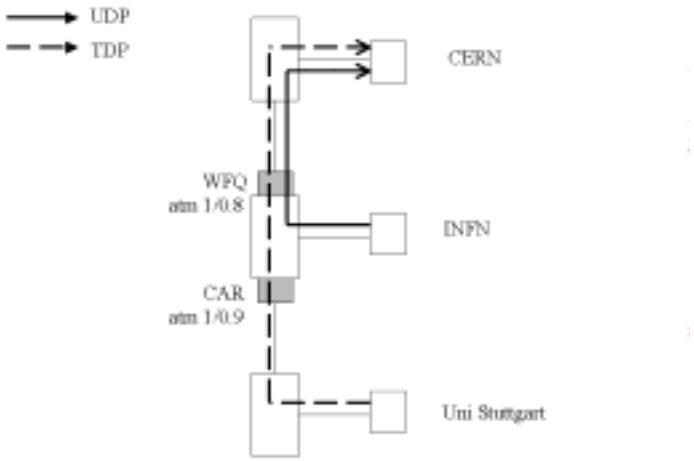


Fig. 4.6: specific test scenario for the debugging of performance inconsistency with 1 high priority TCP stream and background UDP traffic.

According to the network configuration above, the C7200 at INFN is a transit router for the Uni of Stuttgart, which means that TCP traffic and UDP traffic converge into the same interface ATM 1/0.8. Even in this scenario TCP suffered from the UDP load and the same throughput figure (100 kbps) was obtained.

Interpretation of this inconsistency is difficult and is subject to further analysis. Also, the problem is under evaluation by CISCO engineering.

Test B
Test A was repeated by increasing the number of TCP high priority streams to verify if traffic aggregations can get better performance when packet drop is distributed among several streams. Two sites experiencing poor traffic isolation were selected: GRNET and INFN, and tests were performed to one neighbour, since according to test A results are independent of the destination. The length of each TCP stream was set to 200 sec and all the TCP streams were run in parallel.

Throughput of multiple high priority TCP connections: CB-WFQ rate = 1300 kbps, UDP background traffic at 2 Mbps			
Source site	**Destination site**	**Number of TCP streams**	**Aggregate application TCP throughput (Kbps)**
INFN	CERN	1	100
		3	130
		10	221
		20	680[18]
GRNET	CERN	1	700
		3	1000
		10	1180
		20	1270

Table 4.7: Results of CAR and WFQ tests with 1 or more high-priority TCP streams and constant UDP background traffic

As expected, the table shows that by increasing the number of TCP streams, i.e. with traffic aggregations, performance increases. This is probably due to the fact that when one stream is affected by packet drop, there is some chance that other TCP streams are expanding their congestion window, so that the overall sensitivity to packet drop is reduced.

TCP high priority and TCP best effort traffic

The test configuration deployed in this case is equivalent to the one applied to the "TCP HIGH PRIORITY AND UDP BEST-EFFORT TRAFFIC" test above. The only difference is that in this case background traffic is TCP, instead of UDP.

This test was run by selecting a source at INFN (since this is the worst case scenario) and by generating traffic to CERN[19]. The number of both high-priority and best-effort TCP streams was varied, as the following table shows.

[18] The same figure is achieved when the 20 connections are distributed among different receivers.

[19] Since according to the previous tests, in several cases TCP prioritisation is effective even with a considerable load of UDP background traffic, we run this test in the case in the worst scenario, i.e. where traffic isolation is more critical.

Aggregate throughput of high priority TCP connections: CB-WFQ rate = 1300 Kbps, TCP background traffic			
Number of high-priority **TCP streams**	**Number of best-effort TCP streams**	**high-priority aggregate throughput (Mbps)**	**best-effort aggregate throughput (Mbps)**
1	0	1.6	/
1	1	0.8	0.8
10	10	1.17	0.57
10	1	1.46	0.12

Table 4.8: results of CAR and WFQ tests with variable number of high-priority and best-effort TCP streams

The table confirms that even with TCP background traffic, prioritisation of TCP traffic through CAR and WFQ starts to be effective for a relatively large number of TCP streams. With just one high-priority TCP stream, even only one TCP best-effort stream is enough to prevent traffic isolation.

However, the case of 10 high-priority and 10 best-effort streams shows that for an identical load on both classes, CB-WFQ succeeds at providing high-priority traffic with a real preferential treatment.

4.9.3 Premium, assured and best-effort testing with Self Clocked Fair Queuing (SCFQ)

SCFQ is a scheduling algorithm deployed on IBM routers which is a variant of WFQ.
In IBM routers policies are a combination of three components:

Policy = (profile, validity_period, diffserv action)

A *diffserv action* defines the type of marking or re-marking which has to be applied, the queue type to which the packet has to be assigned and the amount of bandwidth that has to be assigned to the queue. Bandwidth limits are enforced through proper allocation of buffers: the number of buffer assigned to each queue is proportional to the amount of bandwidth assigned to a queue.

Memory is divided into two main blocks: a premium area and a best-effort/assured area. The first area is dedicated to premium packets, whilst the second is deployed for the allocation of both assured and best-effort packets. Both areas reserve a given configurable percentage of their size to a shared buffer pool, which can be deployed by all the streams of the same type (premium, assured and best-effort).

While premium traffic is policed and cannot exceed the amount of bandwidth assigned to it (whatever the load is), assured and best-effort traffic are guaranteed a minimum amount of capacity, which can be exceeded in case of lack of congestion. Assured and best-effort buffers can be released after a timeout if they are not deployed and can be allocated to other (active) streams demanding for more buffer space.

Diffserv was enabled on the PPP interface of the IBM 2212. Then a combination of premium, assured and best-effort steams was tested in a local area network illustrated in figure 4.7.

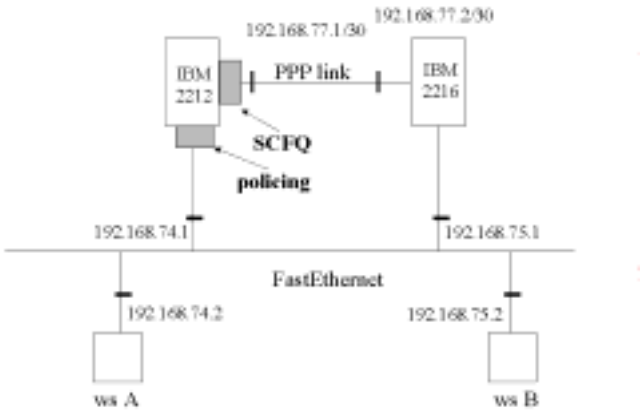


Fig. 4.7: example of network set-up for IBM testing in the local area

Test of premium, assured and best-effort class with UDP traffic
Three different UDP streams, each at a rate of 2048 Kbps, were run. Each stream was associated to a different class, and premium and assured queuing were configured as follows (default configuration):

- premium: 163.8 kbps (8% of PPP int. Bandwidth)
- assured: 819.2 kbps (40% of PPP int. Bandwidth)

The resulting overall router configuration was the following (as reported by the IBM 2212):

				Premium				Assured			
Net If Num	Status	NumQ	Bwdth (%)	Wght (%)	OutBuf (bytes)	MaxQos (%)	Bwdth (%)	Wght (%)	OutBuf (bytes)	MaxQos (%)	
0	PPP	Enabled	2	20	90	5500	95	80	10	27500	80

Seven different tests were run by injecting several combinations of traffic:

1. Best Effort only
2. Assured only
3. Premium only
4. Best Effort + Assured
5. Best effort + Premium
6. Assured + Premium
7. Best Effort + Assured + Premium

Results are reported in Table 4.9:

Premium TCP traffic throughput, target rate: 163 kbps					
Test number	**Streams**	**BE throughput (kbps)**	**Assured Throughput (kbps)**	**Premium throughput (kbps)**	**Total throughput (kbps)** [20]
1	BE	1967.7	/	/	1967.7
2	A	/	1968.0	/	1968.0
3	P	/	/	159.8	159.8
4	BE + A	649.8	1367.0	/	2016.8
5	BE + P	1852.5	/	159.8	2012.3
6	A + P	/	1852.0	159.8	2011.8
7	BE + A + P	617.8	1236.9	159.8	2014.6

Table 4.9: Per class performance for combinations of premium, assured and best-effort traffic

The table shows that premium traffic is correctly isolated, as expected. The performance achieved by the UDP premium stream is equivalent to the one defined by the router configuration and is completely independent of the traffic mixture under test. The threshold set by the router is never exceeded.
As stated by the product documentation, assured and best-effort traffic can get more bandwidth than stated by the router configuration when spare capacity is available (test 2 and 3), unlike the premium class.

The bandwidth guarantee of the assured class is always enforced, since for each test, where present, assured traffic gets more than 819.2 Kbps.

In addition, when best-effort and assured traffic are mixed, best-effort traffic is not starved and excess capacity (i.e. the capacity not guaranteed to the assured class) is shared.

Test of premium class with TCP traffic
The impact of the policing algorithm for premium traffic on TCP performance was tested. The policer is implemented with a token bucket, and as a consequence, the strictness of the policer depends on the token bucket depth.

The default value is 2200 bytes, and is particularly small in order to guarantee low delay jitters and to minimise delays due to queuing.

For each test a single TCP connection was run and the token bucket size was modified. Results are illustrated in Table 4.10.

[20] The aggregate throughput achieved in this tests is higher than the one achieved in the tests relating to CISCO equipment, since in this case no ATM technology is involved in the test with a consequent reduction in protocol overhead and a benefit in overall throughput.

Bucket size (bytes)	Test length (sec)	TCP Throughput (kbps)
2200	Connection stalled	~ 0
4400	60	0.97
6600	60	35.2
	120	74.7
	180	89.8
	240	88.5
	300	95.6
	360	98.2
	420	99.3
	480	100.6
8800	300	118.9
11000	300	124.4
13200	300	124.8
15400	300	126.0
17600	300	125.3
64000	300	125.0

Table 4.10: Effect of token bucket depth on premium traffic

Important considerations can be deduced from the table above.

First of all, the bucket size is of great importance in order to allow a given TCP stream to achieve its target rate (163 kbps). In this test the bucket is fundamental since the input traffic profile is bursty. TCP input traffic is not subject to shaping or to any form of conditioning since it directly comes from the source.

For small bucket sizes (e.g. 2200 bytes, the default value), the TCP stream does not make any progress and almost null throughput is achieved, since packets are continuously retransmitted. For slightly larger sizes, the improvement in performance is significant, but for bucket sizes satisfying the following formula:

$$Bucket\_size > 7 * MTU$$

performance is constant. Even with very large bucket sizes TCP throughput does not increase.
In addition, also the test length can effect throughput.

The important conclusion is that in order to prevent excessive packet drop by the policer, premium traffic has to be shaped first. If this is not possible the bucket sizes must be appropriately tuned according to the average packet size and rate of the incoming premium stream.

4.10 Difficulties Encountered

The network configuration was rather simple, but the debugging of the ATM connection between INFN and University of Stuttgart, suffering from excessive packet drop, took a long time to be solved. Also, the diffserv network was not complete for several weeks from the start of the testing due to some connectivity problems and to the lack of hardware in some sites. The network was partially down for a limited amount of time because of a request for extension, which did not take effect in some sites.

In some cases, testing was slowed down by the lack of suitable software versions for all the router platforms in the network, by poor documentation and by the time needed to interact with the support engineers in case of problems. An additional amount of work needed for the debugging of the performance problems we encountered was also needed.

The real complexity of diffserv stems from the need of a full understanding of QoS features and of their deployment. Also, the large number of parameters which can to be tuned, makes testing and debugging rather complex. Apart from the testing of QoS mechanisms, another element of complexity is the definition of relevant services and the study of their implementation through a set the elementary QoS features.

4.11 Implications for Future Services

QoS testing is very important for the deployment of QoS services in production networks. Several types of services can be defined according to the needs. Afew service examples are provided.

4.11.1 Virtual leased line

QoS features can be combined so that bandwidth capacity at network interfaces is distributed among several configurable traffic classes and hard traffic isolation is achieved. This service can be deployed to support the managed bandwidth service to some sites, for example when ATM is not available.

4.11.2 Capacity allocation on congested links

Traffic can be divided into several classes so that in case of congestion on network interfaces, TCP packets belonging to low priority classes start experiencing packet loss before any other high-priority TCP class. This type of service can be useful to differentiate and regulate the access to expensive network resources.

4.11.3 Capacity allocation on lightly loaded links

On router interfaces which occasionally experience congestion, a bandwidth allocation scheme can be deployed such that different traffic classes are provided with a minimum amount of guaranteed bandwidth which anyhow does not prevent them from grabbing more capacity - if available - without damaging other traffic classes.

This service can also be deployed to protect important traffic like control packets in case of congestion.

4.11.4 Best than best-effort service

Specific applications which are packet loss tolerant but delay or delay jitters sensitive may benefit from a specific type of service by grouping packets from applications with similar requirements together and by queuing them in dedicated and properly tuned queues. In this way, the application requirements are enforced, but packets can still experience packet loss in case of congestion.

4.11.5 Rate limiting

Some traffic classes (for example UDP traffic), which can be configured through fine granularity multi-field classification, may be rate limited in order to protect other traffic categories from excessive “dangerous” traffic.

4.12 Appendix: Metering algorithm deployed by CAR[21]

A pure token bucket defines the following compliance 'zone':

```
---------------------------------------------------->
                    burst time
<====== compliance =====> <=== non-compliance =====>
        'White'                    'black'
```

TCP hates this kind of black and white world and thus needs some kind of graduation of grey from white to dark to find a stability point. This is what the excess-burst offers.

4.12.1 Detailed explanation

As each packet has the CAR limit applied, tokens are removed from the bucket in accordance with the byte size of the packet. And tokens are replenished at regular intervals, in accordance with the configured committed rate. The maximum number of tokens that can ever be in the bucket is determined by the normal burst size. (So far this is just standard token bucket). Now, if a packet arrives and available tokens are less than byte size of the packet, then the extended burst comes into play.

If there is no extended burst capability, which can be achieved by setting the extended burst value to equal the normal burst value, then operation is as in a standard token bucket (i.e., the packet will be dropped if tokens are unavailable).

However, if extended burst capability is configured (i.e., extended burst > normal burst), then the stream is allowed to borrow more tokens (if under standard token bucket there would be none available). The motivation is to not enter into a tail-drop scenario, but rather gradually drop packets in a more RED-like fashion. This works as follows.

If a packet arrives and needs to borrow n tokens, then a comparison is made between two values:

1) the extended burst parameter value
2) 'compounded debt', where 'compounded debt' is computed as sum of a_i, where i indicates the i^{th} packet that tries to borrow tokens since the last time a packet was dropped, and a_i indicates the 'actual debt' value of the stream after packet i is sent (if it is sent).

Note that 'actual debt' is simply a count of how many tokens the stream has currently borrowed.

If 'compounded debt' is greater than the extended burst value, then the packet is dropped. And note that after a packet is dropped, compounded debt is effectively set to 0 and the next packet that needs to borrow will have a new value computed for 'compounded debt', which will be equal to 'actual debt'. So, if 'actual debt' is greater than extended limit, then all packets will be dropped until 'actual debt' is reduced via accumulation of tokens.

Also note that if a packet is dropped, then of course tokens are not removed from the bucket (i.e., dropped packets do not count against any rate or burst limits).

[21] Quotation from Clarence Filsfils's e-mail of the 10th of August 1999.

Though it is true the entire compounded debt is forgiven when a packet is dropped, the actual debt is not forgiven. And the next packet to arrive to insufficient tokens is immediately assigned a new compounded debt value equal to the current actual debt. In this way actual debt can continue to grow until which time it is so large that no compounding is even needed to cause a packet to be dropped. So, in effect at this time the compounded debt is not really forgiven. This would lead to excessive drops on streams that continually exceed normal burst (and thereby discourage that behaviour).

5 RSVP to ATM Signalling Mapping (Author: Tiziana Ferrari – INFN)

5.1 Introduction

The goal of this is experiment is to verify if the integration of the IP and ATM QoS mechanisms is viable and applicable for the support of applications devised by the academic and research community.

The interest in this type of integration is many-fold:

- It fits the TEN-155 network infrastructure and type of traffic, since the backbone is ATM based and applications are IP based;
- For some NRNs the access to the backbone is IP based, but with the integration of ATM and RSVP applications may still make use of the ATM QoS features in the backbone, whilst the access to the backbone is through RSVP;
- According to the TEN-34 test programme results, SVCs can be deployed in a wide area ATM network and several vendors now support RSVP. In addition the protocol standard development process has now reached its maturity;
- It gives the possibility to make use of the ATM QoS features (even if limited to the backbone of the network) to a wider range of user applications.

Since the ATM QoS features need the support of the ATM signalling protocol in the backbone, the pre-requisite of this test is the existence of a stable ATM overlay network with UNI access points at the edge, or if possible, supporting PNNI in the backbone. This implies the need of dedicated VPs for the tunnelling of signalling protocol until the SVC service will be offered in the backbone.

5.2 ATM QoS features vs RSVP

ATM QoS and RSVP are complementary in many respects.

In the first place, ATM relies on dedicated circuits and offers a very rich variety of QoS parameters. The disadvantage of this approach is that in order to benefit from it, ATM signalling support is required and end-systems must be directly connected to an ATM infrastructure. In addition, only few applications can use ATM natively.

On the other hand, RSVP does not require a specific type of layer 2 technology (it runs on top of IP), so it's more flexible. Three different QoS services are available with RSVP: best-effort, controlled-load and guaranteed.

The effective deployment of RSVP in TEN-155 requires a mapping scheme to ATM since the IP-based QoS guarantees depend on the traffic congestion at the ATM layer (the production network is based on UBR connections, which means no guarantees at the ATM layer).

The problem of the RSVP-ATM mapping is challenging for various reasons (see RFC 2382):

- RSVP is receiver oriented (because the receiver issues the reservation request), whilst ATM signalling is sender oriented (the connection set-up request is issued by the sender and forwarded hop by hop to the receiver).
- RSVP reservations are based on "soft states", i.e. reservations need not to be explicitly torn down (they timeout) and their profile can be dynamically updated through re-negotiation. On the other hand, ATM connections have an "hard state".
- In RSVP connection set-up and data transfer are not separate, since data, PATH and RESV messages are issued in parallel. On the other hand, in ATM the data transfer phase starts only after the connection set-up acknowledgement is received and consequently ATM connections are rather delay sensitive.
- There is no straightforward mapping between the Integrated Services and the ATM classes of service. In particular, RSVP QoS services are mostly defined in terms of delay, whilst the ATM classes are defined in terms of bit rate.

Most of these issues have been recently addressed by the IETF working group issll with a set of RFCs. The integration of RSVP and ATM is a relevant subject which appears in several test programmes devised for Academic and Research backbones (see for example the vBNS network test programme).

5.3 Description

The testing activity has not yet started, as it relies on the availability of ATM SVCs within TEN-155. This is expected to be available by mid October 1999.

The detailed test plan is available at: http://www.cnaf.infn.it/~ferrari/tfng/rsvp-atm.html and is summarised as follows:

5.3.1 RSVP-ATM mapping only in the router

In this scenario the router gives the IP-based application the opportunity to take advantage of the capabilities of the core ATM network.

In the transport router the Rspec parameters are mapped into the ATM UNI QoS parameters. In addition, after the virtual circuit set-up phase, the VPI/VCI numbers used to send the flow packets through the TEN-155 ATM cloud are added to the flow state information in the transport router. During the ATM connection establishment, IP packets may cross the ATM cloud either through permanently configured best-effort PVCs or through best-effort SVCs. After the connection set-up completion the IP packets will be switched to the SVC. The same connection is also used by the RSVP PATH messages. If the SVC is a two-way connection (the reverse has been set up), the PATH messages will use this SVC in both directions, otherwise the RSVP messages will keep using the best-effort PVC.

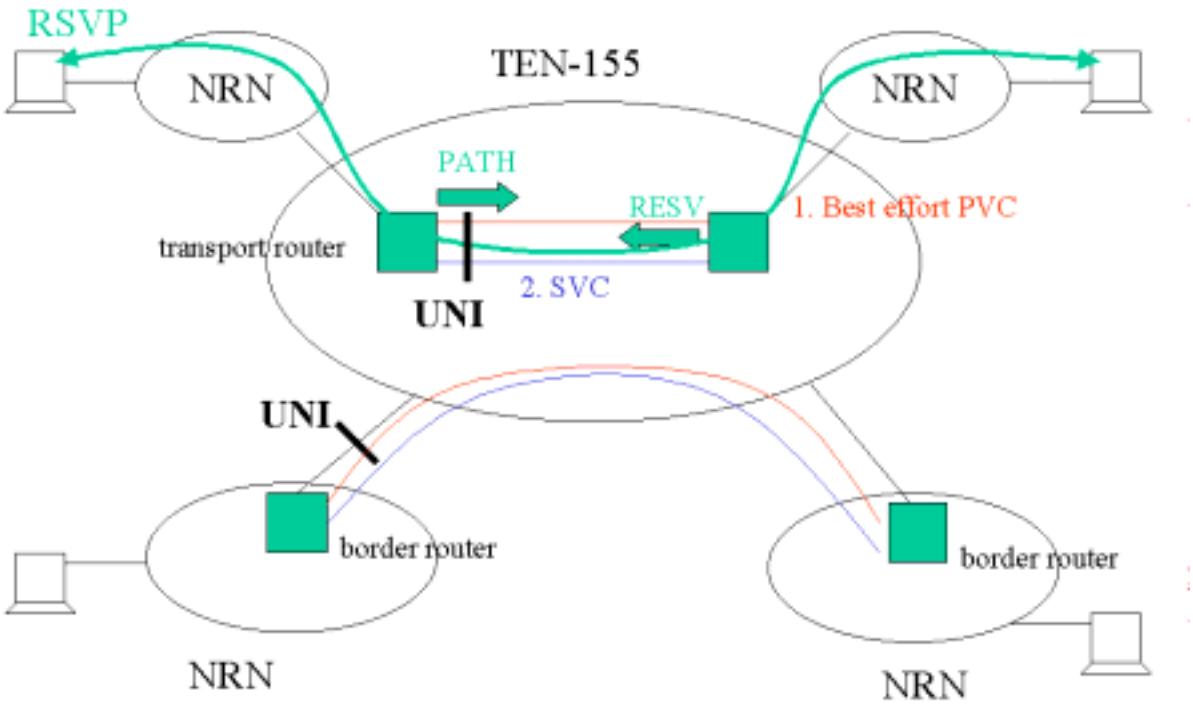


Fig. 5.1 RSVP-ATM mapping only in the router

Admission control relies on the result of both the initial RSVP Call Admission Control and the CAC on the boundary between the IP and ATM cloud.

The interoperability between RSVP re-negotiation and ATM connection profile re-negotiation is another interesting issue to be investigated.

5.3.2 RSVP-ATM mapping both in the router and in the end-system

In this second scenario the RSVP-to-ATM mapping functionality is implemented by the routers which interconnect the LIS (Logical IP Subnets) into which the test network is divided, together with the ATM drivers of the end-systems (which have a native ATM connection to the router of the LIS they belong to).

The testbed can be divided into several LIS, for example one for each NRN and one for the TEN-155 backbone, each interconnected through a router which implements the RSVP-to ATM mapping. The resulting ATM connectivity is made of several hops.

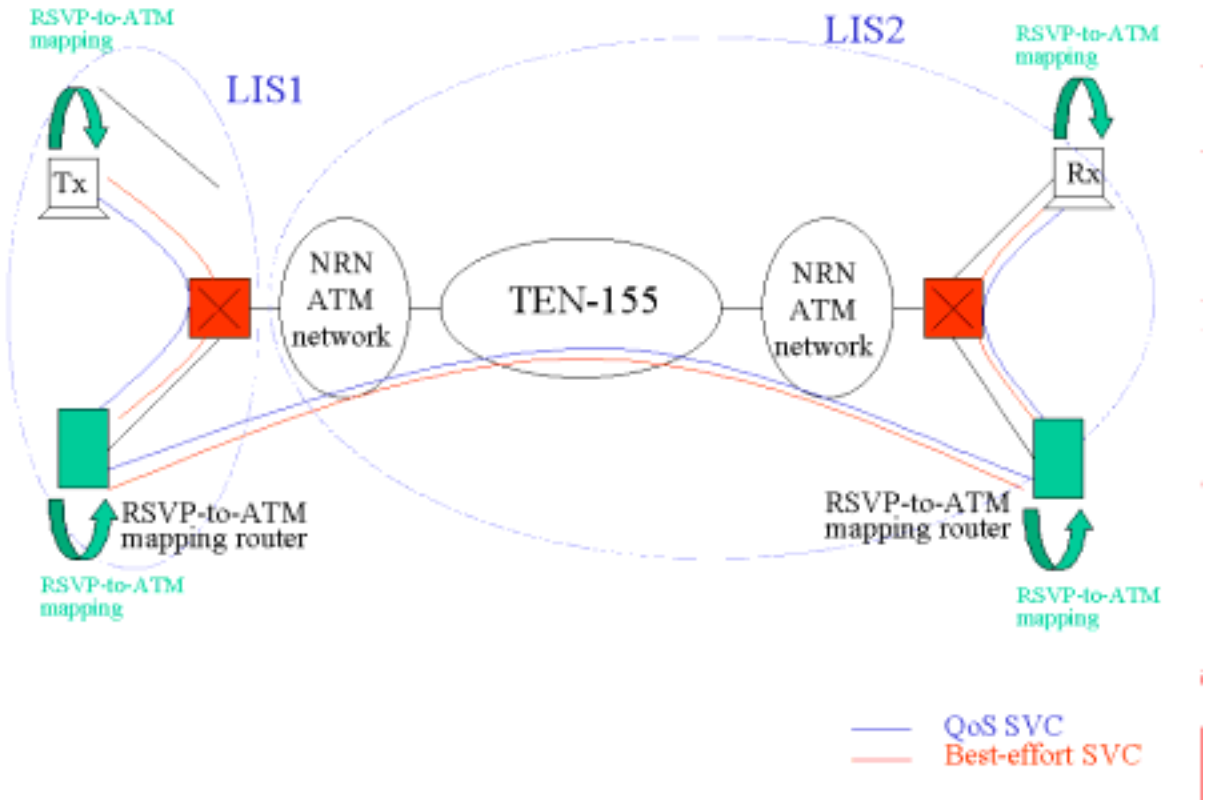


Fig. 5.2 RSVP-ATM mapping in end-systems and routers

Initially a best-effort ATM SVC is used by the source to transfer both data (for example an audio and a video stream) and RSVP messages to the first router. Data and RSVP messages are transparently carried through a best-effort SVC that traverses several LIS and ends up in the destination host. A best-effort ATM connection is also used by the destination to send back RESV messages. If the CAC succeeds, n dedicated SVCs are set-up, where n is the number of flows generated by the source.
The QoS parameters of each SVC depend on the IP flow characteristics specified by the source through the RSVP parameters.

In order to put the equipment to test, generate more sources/receivers can be added to background traffic.

5.4 Participants

The following organisations have expressed interest in participating:

EPFL (CH), SWITCH (CH), GRNET (GR), GARR/INFN (IT), RedIRIS (SP), Dante (UK), CSELT (IT)

5.5 Future work

All the testing activity will start once ATM SVCs are available on TEN-155.

6 ATM Signalling (Authors: Jan Novak, Agnes Pouélé – DANTE)

6.1 Technical objectives

- Implement an ATM SVC service on TEN-155 on 1 June 1999, early deployment possible by QTP during summer 99
- Investigate SVCs capabilities' developments on ASCEND side and developments related to the SVCs deployment areas (RSVP, I-PNNI, p2mp) - longer term

6.2 Description

KPN is contractually obliged to deliver ATM SVCs as a service on TEN-155. The subject of this activity is to discuss with KPN the ATM switch possibilities, to propose a set-up, to test it in KPN/Lucent laboratory, and to implement it on TEN-155 as pilot service for the purposes of QTP and MBS. QTP participants will perform LAN emulation set-up including TEN-155 workstations and workstations in NRN premises to verify the capability of the TEN-155 backbone to provide SVC service.

6.3 Participants

ACONET, DANTE, GRNET, RENATER, SWITCH

6.4 Required Resources

Human resources – staff at DANTE and participating NRNs

Time schedule:
10 - 30 September 1999

Hardware in NRNs:
ATM connected workstations
ATM switch in NRNs: ACONET, RENATER, SWITCH (LS1010), GRNET (FORE/ASX400)

6.5 External Participation

ATM NOC for configurations and debugging in the backbone

6.6 Time Table

1) 13. - 17. 9. 1999 LANE set-up in the backbone
2) 20 - 30.9 1999 LANE set-up including NRNs
3) 1- 8. 10.1999 result report

6.7 Description - Phase 1

6.7.1 Phase 1 - Backbone set-up

Objectives:

1. Configure backbone part for LANE
2. Test SVC routing in the backbone part

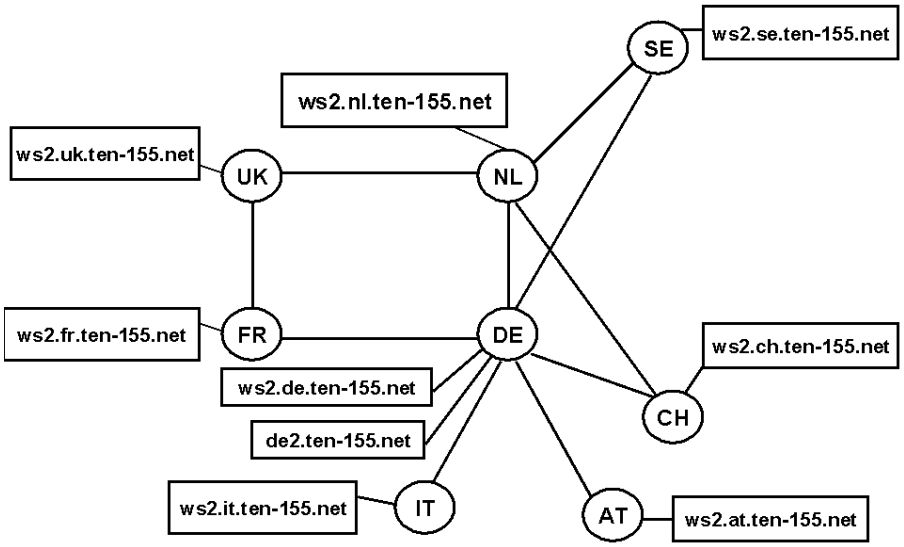


Fig. 6.1 - Backbone hardware set-up

Port configuration: "native" signalling service on physical port
Signalling channels on all ports: VPI/VCI= 0/5

ILMI channels on all ports - 0/16 for all ports, dynamic exchange of LANE servers address, prefixes and ESI

System	Port	ATM address - prefix	ESI	IP address (/24)
ws2.at	WIEN 11/2 2000134	45.043011120000000F.000000 00	08:00:20:97:31:50	212.1.220.1
ws2.ch	GNVE 10/3 2000039	45.041011030000000F.000000 00	08:00:20:99:66:f0	212.1.220.2
ws2.de	FFM 11/2 2000027	45.049011120000000F.000000 00	08:00:20:98:ed:70	212.1.220.3
de2.de	FFM 11/3 1000001	45.049011130000000F.000000 00	00:90:b1:de:10:80	212.1.220.31
ws2.fr	PRS 13/2 2000054	45.033011320000000F.000000 00	08:00:20:98:42:00	212.1.220.4
ws2.it	MLA 10/3 2000049	45.039011030000000F.000000 00	08:00:20:98:42:e0	212.1.220.5
ws2.nl	AMS 11/3	45.031011130000000F.000000 00	08:00:20:98:41:30	212.1.220.6

ws2.se	STHM 11/2 2000066	45.046011120000000F.000000 00	08:00:20:98:42:70	212.1.220.7
ws2.uk	LDN 11/2 2000044	45.044011120000000F.000000 00	08:00:20:97:34:50	212.1.220.8

2. Test results

For simplicity it was decided to configure CLIP and perform the routing and data transfer test on CLIP set-up. This avoids the debugging of point-to-multipoint SVCs and gives the possibility to check the signalling functionality itself.

Data transfer and SVC routing test results

Hardware limitations on SUN ATM cards set to 5 Mbit/s:

From - to	SVC routing path	500 x ping RTT in ms min/avg/max with packet sizes (bytes) of: 100 500 1024 2048 4096	ttcp - udp, 300s
at-de	WIEN - FFM	13.2/13.4/67.8 14.6/14.7/15.1 16.3/16.4/16.7 19.4/19.5/20.0 25.8/26.0/26.4	at-de 5.08 Mbit/s de-at 5.04 Mbit/s
at-fr	WIEN - FFM - PRS	21.5/21.6/22.0 22.9/23.0/23.3 24.6/24.7/25.0 27.8/27.9/28.2 34.2/34.3/34.7	at-fr 5.09 Mbit/s fr-at 5.08 Mbit/s
at-it	WIEN - FFM - MLA	21.8/22.0/22.3 23.3/23.4/23.8 24.9/25.1/25.5 28.1/28.2/28.6 34.5/34.7/35.1	at-it 5.06 Mbit/s it-at 5.08 Mbit/s
at-nl	AMS - FFM - WIEN	18.9/19.0/19.5 19.7/19.8/20.1 20.6/20.8/22.3 22.3/22.4/23.0 25.7/25.9/26.5	at-nl 5.07 Mbit/s nl-at 5.07 Mbit/s
at-se	WIEN-FFM-STH	47.9/48.0/48.5 49.3/49.5/49.9 51.0/51.2/53.8 54.2/54.3/54.9 60.6/60.7/61.9	at-se 5.05 Mbit/s se-at 5.08 Mbit/s

at-uk	WIEN-FFM-PRS - LDN	39.5/39.6/40.0 40.9/41.0/41.4 42.6/42.7/43.1 45.8/45.9/47.1 52.2/52.3/52.8	at-uk 5.07 Mbit/s uk-at 5.08 Mbit/s
de-fr	PRS-FFM	9.19/9.41/56.6 (std = 2.11) 10.6/10.7/11.0 12.3/12.4/12.7 15.4/15.5/15.9 21.9/22.0/22.3	de-fr 5.08 Mbit/s fr-de 5.08 Mbit/s
de-it	FFM - MLA	9.39/9.51/10.1 10.8/10.9/11.2 12.4/12.5/12.8 15.6/15.7/16.1 22.0/22.1/22.5	de-it 5.07 Mbit/s it-de 5.08 Mbit/s
de-nl	FFM - AMS	6.43/6.51/6.82 7.21/7.29/7.58 8.10/8.22/8.67 9.75/9.92/10.4 13.1/13.4/13.9	de-nl 5.06 Mbit/s nl-de 5.06 Mbit/s
de-se	FFM-STH	35.9/36.2/72.3 37.5/37.9/83.7 39.5/39.8/52.4 43.4/43.9/154 51.2/51.4/66.6	de-se 5.05 Mbit/s se-de 5.08 Mbit/s
de-uk	FFM-PRS-LDN	22.0/22.3/53.8 23.7/33.7/164 25.7/26.1/61.1 29.5/29.7/35.1 37.4/37.6/62.5	de-uk 5.03 Mbit/s uk-de 5.07 Mbit/s
fr-it	PRS-FFM-MLA	17.7/17.8/18.1 19.1/19.2/19.7 20.8/20.9/21.2 23.9/24.1/24.4 30.4/30.5/30.9	fr-it 5.08 Mbit/s it-fr 5.04 Mbit/s
fr-nl	AMS - FFM - PRS	20.9/21.0/21.4 21.7/21.8/22.3 22.6/22.8/23.2 24.3/24.5/25.4 27.7/28.0/28.7	fr-nl 5.07 Mbit/s nl-fr 5.08 Mbit/s
fr-se	PRS-FFM-STH	43.8/44.1/125 45.3/45.4/45.7 47.1/47.2/47.5 50.2/50.2/50.7 56.6/56.7/57.2	fr-se 5.06 Mbit/s se-fr 5.08 Mbit/s
fr-uk	PRS-LDN	18.8/18.9/19.6 20.2/20.3/20.6 21.9/22.0/22.4	fr-uk 5.08 Mbit/s uk-fr 5.08 Mbit/s

		25.1/25.2/25.6 31.5/31.6/32.0	
it-nl	AMS - FFM - MLN	23.9/24.0/24.4 24.7/24.8/25.1 25.6/25.8/26.1 27.3/27.5/28.4 30.7/30.9/31.4	it-nl 5.06 Mbit/s nl-it 5.07 Mbit/s
it-se	MLA-FFM-STH	44.1/44.2/44.5 45.5/45.6/46.1 47.2/47.3/47.6 50.4/50.5/50.8 56.8/56.9/64.0	se-it 5.08 Mbit/s it-se 5.07 Mbit/s
it-uk	MLA-FFM-AMS-LDN	21.4/21.5/21.9 22.8/22.9/23.2 24.5/24.7/40.5 27.6/27.7/28.1 34/34/34	it-uk 5.07 Mbit/s uk-it 5.08 Mbit/s
nl-se	AMS - STHM	6.43/6.51/6.82 7.21/7.29/7.58 8.10/8.22/8.67 9.75/9.92/10.4 13.1/13.4/13.9	nl-se 5.07 Mbit/s se-nl 5.06 Mbit/s
nl-uk	AMS - LDN	6.79/6.89/7.35 7.59/7.70/8.4 8.53/8.67/9.37 10.2/10.3/11.0 13.5/13.8/14.4	nl-uk 5.09 Mbit/s uk-nl 5.05 Mbit/s
se-uk	STH-AMS-LDN	37.7/37.8/38.9 39.1/39.3/39.9 40.9/41.0/41.6 44.0/44.1/45.3 50.4/50.6/51.6	se-uk 5.07 Mbit/s uk-se 5.03 Mbit/s

Call set-up measurements

The following table summarises the results of call set-up and address resolution measurements. SUN ATM CLIP software was used for this purpose. This software has the following properties:

- When there is no traffic between CLIP clients, SVCs are torn down after 300 sec
- The IP to ATM address mapping is kept in the cache for 600 sec
- The SVC to the ARP server was always up during all measurements

This allowed measuring the following parameters using ping to all other clients:

1. Call set-up plus address resolution time - ARPS (in the table) - difference between RTT of first and second packet when there is no SVC and ARP cache is empty
2. Call set-up time - S (in the table) - difference between RTT of first and second packet when there is no SVC but ARP cache is still populated
3. Simple RTT between next subsequent packets to the clients - RTT (in the table)

The columns to the right of the blank one have the following meaning:
1. S – RTT: difference of time in columns S and RTT
2. Number of ATM switches involved in the call set-up - 2 for direct neighbours (AT-DE) etc.
3. S-RTT divided by the number of switches: estimation of call set-up time per switch
4. ARP - estimated time for address resolution itself on the cisco router, calculated as difference between ARPS - S - (RTT to ARP server e.g. RTT to DE)

Each row in the table is the result of 5 or 6 measurements, all measurements have been very stable.

from AT to:	ARPS [ms]	S [ms]	RTT [ms]		S - RTT [ms]	Number of ATM switches	per switch [ms]	ARP [ms]
DE	68	53	13		40	2	20.0	2
FR	105	85	21		64	3	21.3	7
IT	154	112	39		73	3	24.3	29
SE	183	132	48		84	3	28.0	38
UK	143	118	34		84	4	21.0	12

from DE to:	ARPS [ms]	S [ms]	RTT [ms]		S - RTT [ms]	Number of ATM switches	per switch [ms]	ARP [ms]
AT	68	54	13		41	2	20.5	8
FR	56	45	9		36	2	18.0	5
IT	120	83	27		56	2	28.0	31
SE	135	98	35		63	3	21.0	31
UK	94	80	21		59	3	19.7	8

from FR to:	ARPS [ms]	S [ms]	RTT [ms]		S - RTT [ms]	Number of ATM switches	per switch [ms]	ARP [ms]
DE	55	44	9		35	2	17.5	2
UK	103	66	19		47	2	23.5	28
AT	103	80	21		64	3	21.3	21
IT	120	90	26		81	3	27.0	37
SE	170	124	43		59	3	19.7	14

from IT to:	ARPS [ms]	S [ms]	RTT [ms]		S - RTT [ms]	Number of ATM switches	per switch [ms]	ARP [ms]
DE	109	80	27		53	2	26.5	2
AT	158	112	39		73	3	24.3	19
FR	144	107	35		72	3	24.0	10
SE	230	160	61		99	3	33.0	43
UK	176	124	39		85	4	21.3	25

from NL to:	ARPS [ms]	S [ms]	RTT [ms]		S - RTT [ms]	Number of ATM switches	per switch [ms]	ARP [ms]
DE	56	40	6		34	2	17.0	4
SE	135	92	31		61	2	30.5	37
UK	70	41	7		34	2	17.0	23
AT	98	75	19		56	3	18.7	17
FR	85	66	15		51	3	17.0	13
IT	113	87	24		63	3	21.0	20

from SE to:	ARPS [ms]	S [ms]	RTT [ms]		S - RTT [ms]	Number of ATM switches	per switch [ms]	ARP [ms]
DE	136	98	36		62	2	31.0	3
AT	185	133	48		85	3	28.3	17
FR	171	125	44		81	3	27.0	11
IT	200	143	53		90	3	30.0	22
UK	175	109	38		71	3	23.7	31

from UK to:	ARPS [ms]	S [ms]	RTT [ms]		S - RTT [ms]	Number of ATM switches	per switch [ms]	ARP [ms]
FR	86	54	13		41	2	20.5	11
DE	106	79	21		58	3	19.3	37
SE	167	109	37		72	3	24.0	6
AT	150	115	34		81	3	27.0	14
IT	193	141	47		94	4	23.5	31

6.7.2 Phase 2 – interoperability with NRNs

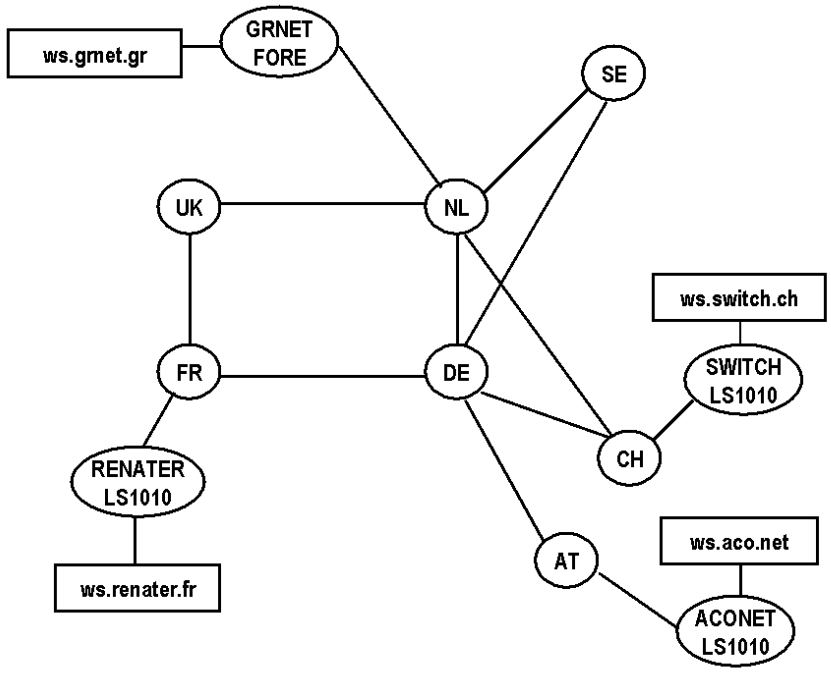


Fig. 6.2 - Hardware set-up

Port configuration: CBR VP 9 (2 Mbit/s) configured towards NRNs with signalling service enabled
Signalling channels on all ports : VPI/VCI=9/5
ILMI channels on all ports - 9/16 for all ports, dynamic exchange of LANE servers address, prefixes and ESI

System	Port	ATM address - prefix	ESI	IP address (/24)
ws.aco.net	WIEN 11/4 2000135			212.1.220.11
ws.grnet.gr	ATNI 14/8 2000075			212.1.220.61
ws.renater.fr	PRS 9/4 2000056			212.1.220.41
ws.switch.ch	GNVE 9/4 2000041			212.1.220.21

6.8 Conclusions

The ATM signalling was successfully configured in AT, DE, FR, IT, SE, UK. It was not possible to set-up SVCs towards NL (although outgoing SVC worked properly). The service was totally unusable in CH - the SVCs have been set-up, but round trip times of IP packets varied from 50ms to 1.5 seconds from one ping to another. Despite a thorough debugging process, it was not possible to determine the cause of

this discrepancy. It was therefore decided, to concentrate on the testing of the SVC functionality on the nodes where the round trip times were consistent. The investigation of the problems in CH and NL is still open.

The IP level RTT (round trip time) between all clients was measured - the values have been very stable, so they were used for the estimation of call set-up times on the switches - as described in the results. A data transfer between all clients with limited bandwidth to 5 Mbit/s produced also very stable results. In some rare cases, for debugging purposes, data transfers were run at full line rate available on the production trunks with satisfactory results – no difference with the production PVC service was noticed. Potentially the use of CLIP for the current production service could be investigated. This requires investigation of operational/management issues, load balancing/sharing issues and redundancy issues (the ARP server being the single point of failure for whole network operation).

The measurement of call set-up times provided very stable results and it seems that the interpretation of the data provided in the results table is correct. The result of call set-up measurements is (23.2 + 4.4) ms per one ATM switch, where the first value in the bracket is the arithmetic average of all measurements from tables in section H, the second one is the standard deviation. Given the broad range of Round Trip Times from min 6 ms to max 61 ms the standard deviation confirms the assumptions/interpretations made. On the other hand, the estimation of ARP request time did not provide similar results, probably due to too many processes (and too much uncertainty in subtracting many numbers with large relative errors) involved in clients, ATM network and ARP server and the results have not yet been analysed any further.

Interoperability tests with NRNs have not been performed to date.

7 Policy Control (Author: Leon Gommans – University of Utrecht)

7.1 Description

Policy Control mechanisms are required to allow allocation of resources to processes or persons according to a defined set of rules based on some form of contract between organisations that offer a service or individuals/organisations that require a service.

7.2 Approach

The workgroup will focus on solutions that implement standards set by organisations such as the IETF and DMTF. As most related work is currently in progress, the working group is active by observing the developments and actively participating in IETF activities. The IETF Authentication Authorisation and Accounting (AAA) work promises to be an overall activity overlooking work of various other IETF working groups such as Mobile IP, RAP, Roam ops and others where AAA has been recognised as a requirement.

The AAA working group is chartered to establish a common AAA protocol allowing policy based decisions to be made across independent organisations.

The Policy Control WG seeks co-operation with TF-TANT WGs that need a policy-based decision on access control and resource allocation such as the Diffserv WG.

The Policy Control WG does not actively investigate any Authentication or Accounting related mechanisms, however investigations include possible ways to interact with these mechanisms.

Contributions to the IETF AAA working-group are being made via the University of Utrecht, however any member of TF-TANT is free to participate according to the rules of the IETF.

7.3 Resources required

The resources that are currently required are manpower to collect requirements of the specific environments the TF-TANT members operate in and to disseminate information regarding the status of the standardisation work and its implementations. This is necessary to seek possible collaborations with TF-TANT experiment members.

If implementations of standards based AAA solutions become available, a more elaborate resource requirement description will be provided. Single laboratory experiments are currently being set up in the University of Utrecht and are covered by the current resources available.

7.4 Description of the Experiment

Although not many details can be given, possible experiments include:

- Distance Learning applications where students need authorisation from one or more independent organisations.
- Access to quality network resources based on a student's budget.
- Automated reprint delivery as an extension to current abstract services from libraries.

7.5 Technical set-up

Considering 6.4, experiments might involve the deployment of

- AAA server(s)
- Network Access equipment that allow policy enforcement and detection of users.
- Application server(s) that request authorization from one or more AAA servers.
- The Diffserv infrastructure as offered by the Diffserv WG.

7.6 Planned timetable

Planned availability of a AAA server: Q1/2000
Planned availability of an AAA based application: Q3/2000

7.7 People/organizations involved

Currently the work actively involves the following people and organisations:

University of Utrecht, Computational Physics group: Cees de Laat,
Hans Blom, Linda Penneweert and Leon Gommans.
Surfnet, AAAA working group lead by Ton Verschuren.
Merit Networks, Ann Arbor MI: John Vollbrecht, Dave Spence
Interpay Nederland BV: Betty de Bruijn.
Cabletron Systems Inc.: John Roese, Franke Lau
Ellacoya Networks Inc.: Kurt Dobbins.
Lucent Technologies: Brian Lloyd.

7.8 Work done; progress so far

Contributions were made to the IETF AAA working groups as:
draft-ietf-aaa-authorization-reqs-00.txt

7.9 Results of the Experiment

No experiments have been conducted yet.

7.10 Future Activities

No future activities are yet planned; however the intention is to expand the experiments by including more and more complex service authorisation structures.

8 IP over ATM (Author: Roberto Sabatino – DANTE)

8.1 Introduction

TEN-155 offers an IP service based on ATM as the underlying infrastructure. The mapping between IP and ATM is done according to RFC 1483. On TEN-155 a mix of two ATM traffic classes (ATC) are used to ensure efficient backbone usage and fair sharing of capacity between competing flows in cases of congestion. These ATCs are DBR and SBR3, which correspond to CBR and VBR-nrt (with SCR=10, PCR=line rate) respectively. These traffic classes have shown to be suitable for meeting the requirements of TEN-155 and have been thoroughly tested prior to their use on TEN-155.

Other ATCs are available, such as ABR, SBR3 (with SCR >> 0) and SBR2, and it is suggested to investigate further the use of these ATCs to verify if they are suitable for use on TEN-155. Theoretically using SBR3 with SCR>>0 could be advantageous in situations where a minimum guarantee is required as well as the ability to exceed the minimum guarantee. ABR in theory meets these same requirements, except that in this case the ATM layer provides feedback to the end systems for congestion control and therefore avoids cell drops at the ATM level. SBR2 has a similar behaviour to SBR3, except that the end systems rather than the ATM switches tag the cells exceeding SCR.

8.2 Objectives

The objectives of the experiment are to understand exactly how the different ATCs behave in isolated situations and in co-existence and competition with other ATCs. This will enable network managers to configure the most appropriate set of ATCs for their specific needs. On TEN-155 there is a specific interest in verifying if a different mix of ATCs is more suitable than the one currently used.

8.3 Test description

The idea is to have ATM switches from various vendors tested in a laboratory environment. End systems such as workstations, routers, ATM traffic analyser should be connected to these switches and the various ATCs should be tested. Once tests are successful in a laboratory environment, a test on TEN-155 may be performed with live traffic. From the Dante perspective, it is interesting to perform these tests on Ascend CBX 500 switches as these are deployed in TEN-155.
For laboratory tests a simple set-up such as the following is required

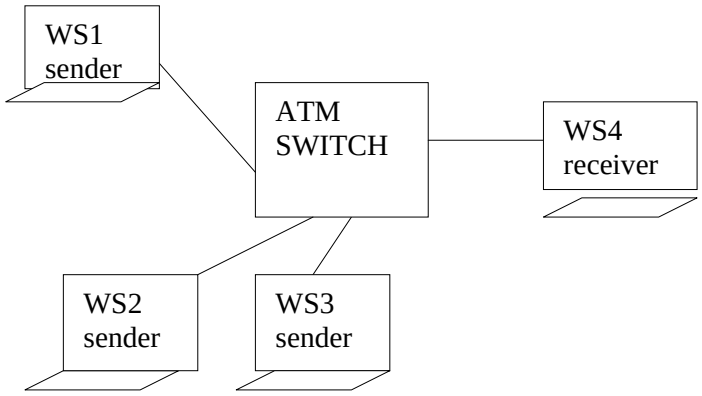


Fig. 8.1 Test set-up for SBR3 (SCR >0) testing

This set up allows to create competing streams from WS1, WS2 and WS3 towards WS4, therefore the co-existence of different ATCs in situations of congestion can be tested. The ttcp program is used to generate artificial traffic between the workstations. The connections between workstations and ATM switch are STM-1.
The laboratory tests enable to verify the behaviour of the various ATCs in conditions of extreme load, but with artificial traffic. Field tests on TEN-155 with live traffic are required to verify the behaviour of the ATCs with real traffic patterns. With field tests though it is neither possible nor desirable to artificially create situations of congestion.

8.4 Planned timetable

March 99: testing of SBR3, with SCR >>0 in laboratory environment.

From April 99 onwards: testing of SBR2 and ABR, subject to the availability of suitable ATM switches and end systems. Note that ABR is available on TEN-155 since August 1999.

8.5 Organisations involved

DANTE staff and KPN staff are carrying out this activity. The laboratory testing is done at the Lucent laboratories in Hilversum (NL), on Ascend CBX 500 switches.

8.6 Progress and results

Initial tests with SBR3 with SCR>>0 have been performed in Hilversum, NL. These tests are described in detail in http://www.dante.net/staff/roberto/docs/1999/qtp/RS-99-10.html. ATM PVCs are set up between the sending workstations and the receiving workstation. The 3 PVCs have different SCR, but all have PCR=line rate. The sum of the SCR must not exceed the capacity of the link towards the receiving workstation.

Four sets of tests were performed:

1. all senders sending at a rate less than the corresponding SCR of PVC to receiver;
2. all senders sending at a rate > SCR, but without creating congestion towards receiver;
3. some senders sending at >SCR, others at <SCR, creating congestion towards receiver;
4. all senders sending at a rate > SCR, creating congestion towards the receiver

Tests 1 and 3 are important to verify that the ATC basically works, and that in situations of under utilisation of the PVCs, traffic on the PVCs is not affected by traffic on other PVCs which may be overloaded.

Test 2 is required to verify that in cases of extra capacity being available, this may be used by traffic flows sending at a rate greater than the SCR of the corresponding PVC.

Test 4 is required to demonstrate that in cases of severe congestion, each flow receives at least the SCR of the corresponding PVC to the receiver. As the sum of the SCR is less that line rate, the remaining capacity should be distributed amongst the competing streams.

The tests have outlined that in a situation with no congestion the ATC SBR3 with SCR>>0 behaves well, in that SCR is guaranteed and extra capacity available may be used. The tests have however outlined that the ATC does not work in situations of congestion: the SCRs are in effect not guaranteed.

The details of the results are subject to a non-disclosure agreement with Ascend. While the unexpected behaviour is being investigated, to date this problem has not been solved.

8.7 Future work

To perform tests with ABR and SBR2 it is necessary to have access to end systems with these capabilities. Tests with ABR have been performed with ATM traffic analysers, but this is viewed as a limited test as these are not systems that users use for real traffic. Future work includes identifying systems with the necessary capability to generate ABR and SBR2 ATCs.

8.8 Implications for future services

If the problems detected with ATC SBR3 with SCR>>0 are resolved, it may be possible to use this ATC on circuits which often experience congestion, in order to guarantee a minimum capacity and the ability to use more in situations of spare capacity being available.

9 Flow-Based Monitoring and Analysis (Author: Simon Leinen – SWITCH)

9.1 Abstract

This work package investigates the class of ``flow-based'' accounting methods for IP traffic, and their applicability to the design and operation of large backbone networks such as TEN-155 or a National Research Network (NRN).

9.2 Introduction

Traffic measurement has traditionally been an important component of network design, capacity planning, and operational monitoring. In the past, such measurements have generally been restricted to relatively simple metrics per interface or trunk, such as total input and output traffic rates. Thus, many interesting aspects of network traffic were relatively hard to measure. Examples of those are: the ``application mix'' of the traffic over a particular trunk line; distribution-valued parameters such as packet-size distributions; transport-oriented measures such as TCP connection throughput; classification of traffic on trunk lines according to customers etc.

In recent years, flow-based accounting mechanisms have been proposed that can be used to provide these types of information. One example is the IETF (Internet Engineering Task Force) RTFM (Real-Time Flow Measurement) architecture proposed by Nevil Brownlee of the University of Auckland. Another example is Cisco's NetFlow accounting.

As these systems reach maturity, and an abundance of software packages building upon them started to become available, this technology has become very interesting to backbone IP network operators. The goal of this work package is to evaluate available systems against the particular needs of backbone network operators.

9.3 Test Strategy

We have adopted the following strategy to evaluate available protocols and software products against the specific needs of backbone IP network providers:

1. Define a few sample applications for detailed accounting which are of actual interest to some members of the TF-TANT community.
2. Implement those applications with some of the available tools.
3. Evaluate the tools with respect to the suitability for the different tasks.
4. Provide feedback on possible improvements of the protocols and software products to the respective vendors/developers.

9.4 Status of the Experiment

As of early September 1999, preliminary work has been performed as described in the following paragraphs.

The TF-TANT community has provided valuable input on how traffic accounting is currently used by DANTE and some NRNs, as well as on potential new applications.

DANTE has opened a test account on the test workstation in the Geneva TEN-155 PoP for use by the test participants.

The NetFlow accounting stream generated by the Geneva TEN-155 router has been diverted to a ``flow replicator'' program. This program, which has been developed by GARR, is capable of copying a router's accounting stream and sending it to a set of receivers. This allows peaceful coexistence of the production statistics collector in the TEN-155 network and the programs evaluated within this experiment.

A selected set of software packages for the post-processing of NetFlow accounting data has been installed on the Geneva test workstation:

- CAIDA cflowd
- Cisco FlowCollector/FlowAnalyzer
- Fluxoscope (SWITCH's NetFlow accounting tool)

9.5 Applications to Investigate

Traffic Statistics at Exchange Points
One area where traditional traffic measurement is insufficient is where a network exchanges traffic with multiple peers over a shared medium such as a LAN-based Internet Exchange Point. Using per-interface counters, it is not possible to measure the traffic exchanged with individual peer networks.

Flow-based accounting can be used to overcome this by either:

- using next/previous-hop IP or MAC address information
- using BGP routing information (neighbour/origin AS or AS paths)

BGP information can either be provided by the router itself (NetFlow v5 and later) or by the collecting process which interacts with a routing registry or a BGP speaker.

Accounting for Volume-Based Charging
Network operators that use volume-based charging methods have found it problematic to charge an indiscriminate price for all traffic. Experience has shown that this leads to waste of resources, since people will save on local traffic that doesn't cost much to provide.
Therefore it is desirable to price traffic differentially according to destination/source. As an example, traffic within the NRN and with peers may be free (included in access fee), but traffic over expensive external connections such as TEN-155 or a US transit provider might be charged for.
Flow-based accounting can support such charging schemes, but the following conditions should be apply too:

- Separate traffic counts are generated for each (customer, external-network) pair.
- Optionally, ``itemise'' the counts for an organisation according to internal cost centres (as done in JANET).
- Near-real-time feedback should be provided to organisations (or cost centres, or even individual users).
- A trail of individual accounting records is kept to substantiate bills. This trail can become quite large, so won't be around for very long. Therefore it is important to provide
- near-real-time feedback so that users can check their data frequently.

In addition, such a differentiated charging model may be extended to support differentiated services such as those studied in the DiffServ experiment.

Abuse/Attack Detection
Flow-based accounting can be used to detect anomalous traffic such as

- denial-of-service attacks such as ``smurf''
- attempts to breach security: systematic scans, known bugs/backdoors
- high amounts of non-adaptive traffic
-

Care must be taken to ensure privacy and to avoid false positives (for example, WWW caches may look similar to port scanners at first sight).

Long-Term Traffic Analysis
As a prerequisite to intelligent connectivity and bandwidth provisioning, network operators want to be able to recognise large-scale trends about the traffic over their networks.

- application mix on different connections
- emerging applications
- interesting/important source/destination networks

Detection of Routing Anomalies
The aim in this case is to compare actual traffic flows with intended routing policy.
The motivation behind this is that providers try to send their outbound traffic over the optimal links (to TEN-155 rather than the US, or towards settlement-free peers rather than paid transit connections), but they cannot easily exert control over the paths through which inbound traffic comes back to their network.

9.6 Intermediate Results

Comparison Between Flow-Based Accounting Protocols
When comparing the IETF RTFM (Real-Time Flow Measurement) and Cisco NetFlow protocols, the following can be observed:

RTFM
RTFM is an IETF effort, which means that a priori it has a good potential to lead to a solution supported by multiple vendors. But so far, the only known implementations are Nevil Brownlee's NeTraMet system and another implementation from IBM.

The NeTraMet implementation runs on general-purpose computers under either MS-DOS or Unix. Data collection uses LAN interfaces such as Ethernet or FDDI in promiscuous mode. This makes it easy to deploy probes on shared LAN segments, but difficult to measure traffic over point-to-point links. As most of the interesting trunks in modern backbone networks run over such point-to-point links, this is a severe limitation. In particular we haven't found a way to test NeTraMet in a realistic setting within the TEN-155 network.

In the RTFM architecture, a measurement probe (traffic meter) is configured with a set of rules for classifying packets into flows. The Simple Network Management Protocol (SNMP) can be used both for installing those rule sets and for reading out measured data.

Cisco NetFlow
NetFlow is proprietary to a router vendor (Cisco), but well-documented and has seen wide deployment in the short period since its first release. There is a relatively large (and growing) amount of software that can process NetFlow accounting information.
In NetFlow, data collection is performed at the router. If NetFlow is enabled on an interface, the

packets received on that interface will be classified into flows according to a fixed set of parameters (source and destination IP address, protocol, source and destination TCP/UDP port, and TOS byte/DSCP). For each flow, octet and packet counters as well as some additional information is kept. Collected data for flows is then ``exported'' asynchronously to a specified management station. Whenever accounting data is exported for a flow, that flow is deleted from the flow cache. A flow can be deleted--and thus exported--for one of the following reasons:

1. The flow cache is full and a new flow must be created. In this case some kind of LRU scheme is probably used to determine which of the existing flows should be purged.
2. A packet has been received for the flow that signals the end of the flow. An example of this is a packet containing a TCP segment with the FIN bit set.
3. The lifetime of the flow in the cache has exceeded a limit. This limit used to be fixed at thirty minutes in early releases of the NetFlow router code. This is still the default in recent versions, but can be changed in a range of 1-60 minutes.
4. No new packets have been received for the flow for a certain amount of time. This inactive timeout can be configured to a range between 10 and 600 seconds in recent releases.

When flow accounting data is ready for export, the flow sender tries to ``batch'' multiple accounting records into a single UDP packet.

With the NetFlow v5 accounting format, every accounting record is 48 bytes long. Up to 30 flow accounting records are batched into a single UDP packet. A header of 24 bytes contains information about the entire set of flow accounting records in the packet.

The packet header includes a sequence number, so that missing packets can be detected. However, there is no possibility to have lost packets retransmitted.

Recent versions of Cisco IOS implement NetFlow accounting in ``distributed'' mode. In this mode, Versatile Interface Processors (VIPs) autonomously manage their own NetFlow caches, and export accounting data independently.

9.7 Planned Work

The following steps are planned for the remaining lifetime of the experiment:

The different products should be configured and evaluated for the applications at hand.

More software packages might be investigated. In particular, the current set doesn't include any package that is specifically oriented towards charging and billing. However, those systems have recently started to become available and might be quite suitable for some of the applications, especially the charging application described previously.

Deployment issues should be investigated and documented. The resulting document should explain the trade-offs in deciding where measurement probes should be deployed in the network, what must be measured, and so on.

The final report should include:

1. the description of flow-based approaches to traffic accounting and analysis (RTFM, Cisco NetFlow)
2. use of flow-based accounting for different applications (detailed description of deployment/scaling issues)

3. comparison of tools and suggestions for improvement.

9.8 Conclusions and Outlook

Flow-based accounting mechanisms such as RTFM or Cisco NetFlow have the potential to be very useful in short-term and long-term monitoring of backbone IP networks. Most of the applications that have been investigated so far are somewhat limited in the range of traffic categorisation they support, which results in restricted usefulness for some of the applications envisioned. Some level of programmability would seem to be useful in the accounting data collection components. While it is possible to implement programmable data collection systems on general-purpose workstations, this may no longer be an option if categorisation of accounting data is performed in the router, which may become necessary in some situations for performance reasons.

9.9 References

http://www.switch.ch/tf-tant/floma/

10 Multicast (IP and ATM) (Authors: Robert Stoy - RUS, Jan Novak - DANTE)

10.1 Problem Statement

For several years multicast on the internet has been supported by the Mbone, which is in effect an overlay network interconnecting several routers on the internet. The technology used is based on DVMRP, which is a multicast forwarding and routing protocol. DVMRP is a distance vector routing protocol and therefore does not have the ability to apply routing policies.

The current DVMRP based world-wide Mbone is reaching its limits in providing an IP multicast service in a production environment because of the following reasons:

DVMRP as a distance vector routing protocol does not scale with the growing numbers of multicast sessions and networks providing a multicast service;

Multicast forwarding with DVMRP is based on a *flood and prune* mechanism which in too many cases is not implemented correctly thus leading to a considerable amount of unwanted traffic on the internet and causing congestion on the Mbone itself;

The multicast infrastructure is based mainly on rate-limited tunnels, which are affected by the congestion due to the flood and prune mechanism. Tunnelling in itself is a cause of performance limitations;
The multicast topology is different from the unicast Internet, thus leading to sub-optimal utilisation of expensive international circuits.

As a result, multicast sessions are often suffering from high packet losses, and inefficient multicast routes leading overall to poor performance and quality.

The solution to these problems is the introduction of hierarchical multicast routing topologies by using new inter-domain and intra-domain routing protocols.

10.1.1 Inter-domain Multicast Routing

The IETF Mbone-WG and IDMR-WG have been working for several years on a new multicast forwarding and routing protocols. Both inter-domain and intra-domain protocols are being developed and standardised. For intra-domain forwarding and routing, the development of PIM-SM has reached stability that allows its deployment.

Inter-domain routing and forwarding, which are needed to provide a multicast service across multicast domain borders is, on the other hand, still being developed. The IETF has engineerd an interim solution to inter-domain multicast routing, which is the combination of the MSDP (Multicast Source Discovery protocol) protocol together with the MBGP protocol (Multiprotocol BGP. Implementations of these protocols are already available on Cisco routers.

On the longer term the IETF works on a “final” multicast distribution model, the BGMP/MASC architecture. First implementations of components of this architecture are expected to appear in fall 1999, and should be tested in the second phase of this experiment.

DANTE, with the co-operation of the NRNs, has carried out the migration from the DVMRP based European Mbone to a native multicast pilot service based on PIM-SM and MBGP.

10.1.2 Intra-domain Multicast Routing, point to multi-point ATM-SVC service

TEN-155 provides an IP over ATM service according to RFC 1483, and the same applies for the IP multicast service. With the forthcoming availability of ATM SVCs (autumn 1999), the mapping of PIM-SM to point-to-multipoint (p2mp) ATM SVCs is potentially an effective way of ditributing IP multicast within TEN-155. RFC 2337 *Intra-LIS IP multicast among routers over ATM using Sparse Mode PIM* describes this mapping in detail.

Some initial basic tests of this functionality have been successfully carried out within the TF-TEN project in 1998. Before this functionality can be introduced on TEN-155 p2mp ATM SVC must be tested as well as the QoS associated to these

10.1.3 User Site multicast performance and routing monitoring

Multicast network performance monitoring between user sites is currently done only from time to time and often only during troubleshooting procedures, after problems occurred and users send trouble tickets. Multicast network statistics are needed to analyse problems and can help to find and locate problems before they become urgent.

As a component of the test set-up in the inter-and intra-domain activities a multicast service routing and performance monitoring system at the user sites will be established, that could be later used for monitoring a production multicast backbone service. Besides the common tools, such as MRTG and HP Open-View, also Cisco's Multicast Router Monitoring (MRM) functionality will be tested when available.

10.2 Objectives of the Experiment

10.2.1 Inter-domain multicast routing

The migration from DVMRP to PIM-SM/MBGP/MSDP required the following steps to be performed :

1) Evaluation of Cisco's implementation of MBGP/MSDP on non-production equipment. Cisco was chosen as test equipment because these are the routers currently deployed on TEN-155;
2) Implementation of MBGP/MSDP on TEN-155 routers;
3) Connection of NRNs to MBGP/MSDP in the core and tests of data transfers between different domains.

As soon as test implementations of the MASC/BGMP implementations are available, these should be tested in non-production environment. Depending on the test results (and also on the status of deployment of these protocols in the general Internet) the migration from MBGP/MSDP to BGMP/MASC should be planned.

10.2.2 PIM-SM mapping to point to multi-point ATM-SVCs

This activity will focus on the correct functionality of TEN-155's point to multi-point ATM SVC service including the ability to provide requested QoS. Quantitative measurements like the SVC set-up times and the provided delays and bandwidth will be measured.

With respect to the interaction between PIM-SM and the ATM service, all phases of this interaction will be carefully tested and monitored for all steps of a multicast session, including creation, data distribution, and session deletion.

10.2.3 User Site multicast performance and routing monitoring

The goal of the monitoring activity is to provide an environment that allows the monitoring on dedicated user sites of one permanent dedicated multicast test session, which is used to periodically generate a small load on the network for troubleshooting purposes. The system should have a WWW interface.

10.3 Outline Solution

The implementation of IP multicast on TEN-155 is described in detail in the following documents:

http://www.dante.net/mbone/mcast99/migration.html and
http://www.dante.net/mbone/mcast99/mphase2.html.

Basically, test equipment in the TEN-155 PoP in Frankfurt was installed and the TEN-155 DVMRP cloud was connected to it. DFN and CESNET were connected with the new protocol stack MBGP/MSDP. The stability of the test router with new software and data transfer between DVMRP and MBGP/MSDP domains were successfully tested. After this MBGP/MSDP was enabled on TEN-155 routers in Sweden, Netherlands, United Kingdom and France.

SurfNET, NORDUnet, GRnet, RCCN and RedIRIS have a native connection to TEN-155's multicast infrastcture, whilst Belnet, CESNET, DFN, Machba/ILAN, SWITCH and Renater use dedicated equipment. This dedicated equipment is connected either via a separate ATM PVC or an IP-in-IP tunnel to the TEN-155 equipment. The remaining NRNs still have to complete their migration to the new infrastructure.

TEN-155's multicast was successfully stress-tested during the IETF meeting in Oslo in July 1999 (when it delivered data to both DVMRP and MBGP clouds) during which two parallel sessions of 2Mbps each were transmitted via TEN-155.

In parallel to the migration of TEN-155 to MBGP/MSDP, the BGMP/MASC development at IETF was followed. A first implementation of BGMP is available in GATED, and has been retrieved for initial testing. A first MASC implementation is also available.

Discussions are ongoing with KPNQwest with the goal to define the configuration for the point to multi-point ATM-SVC test activity.

For the multicast monitoring the RTP/RTCP monitoring tools and the MULTIMON software are being evaluated.

10.4 Participants

All TEN-155 NRNs

10.5 Time Table

1. Basic testing - 30.4. - 10.5.1999
2. Backbone set-up - 12.5 - 3.8.1999
3. NRN connections - 12.5.1999 - ongoing

10.6 Current Status

The following figure illustrates the TEN-155 multicast topology:

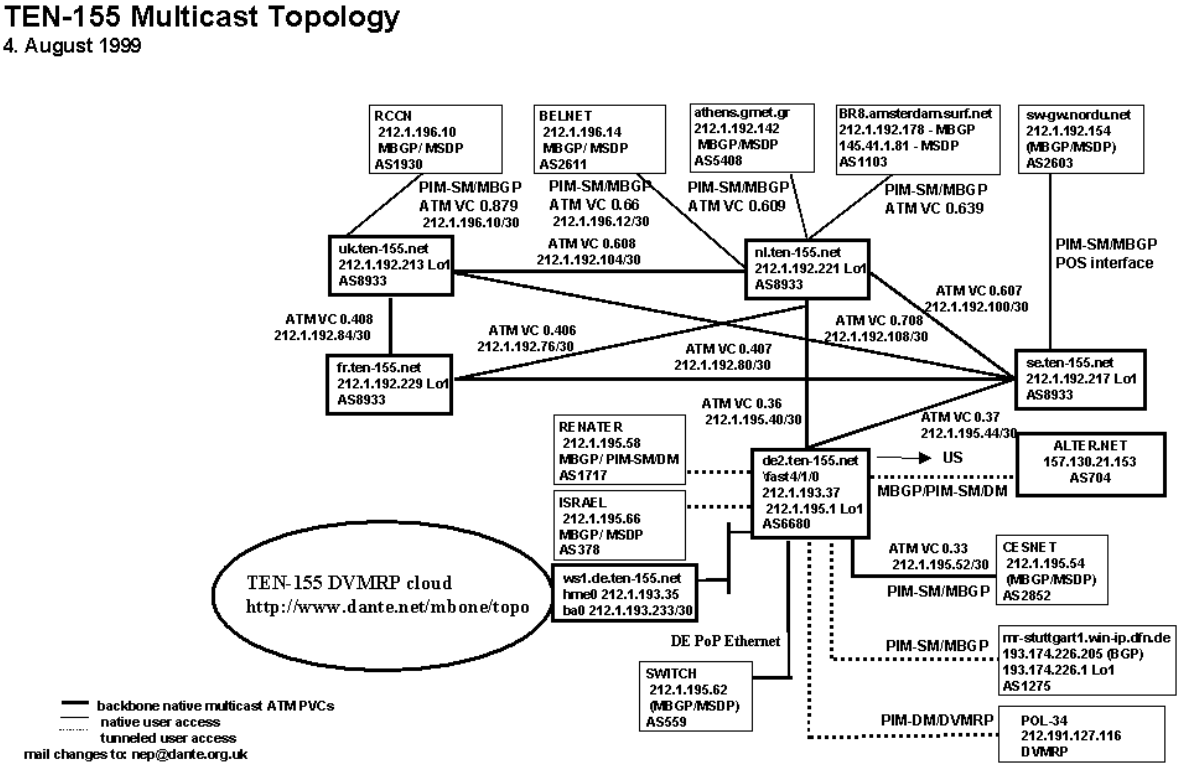


Fig. 10.1 TEN-155 Multicast Topology

10.7 Future Activities

The following activities will start in autumn 1999:

- Test of MASC/BGMP components, building up a test configuration, if possible in co-operation with ISI and the MECCANO project.
- Tests on point to multi-point SVC service of TEN-155 and its interaction with PIM-SM.
- Building up the multicast monitoring tools with a web interface.
- Testing other implementations on other (than Cisco) vendor equipment.

11 IPv6 (Author: Simon Nybroe – Telebit)

11.1 Objectives of the Experiment

The goal of this experiment is to provide the information needed to implement IPv6 connectivity as a service in the TEN-155 production network. This will provide input to DANTE with regard to a future IPv6 production service and on the same time work as a forum where NRNs (and possibly others) that have or want hands-on experience can share their findings.

11.2 Outline Solution

While IPv6 is a new protocol, the steps in providing a backbone service are very similar to those needed to follow in order to provide an IPv4 backbone service. This will hopefully shorten the learning curve and ease the adaptation of IPv6 into DANTE's existing management procedures.

The experiment will be performed in two locations: the Telebit test laboratory and in the field. The laboratory tests will focus on testing compliance, interoperability and performance of the IPv6 equipment. It will also be used to verify specific configurations before actually deploying them in the field and to enable closer examination of interesting issues that arise in the field tests.

11.3 Description of the Experiment

11.3.1 people/organisations involved

Below is the list of the people and organisations involved.

SURFnet (NL)	Ronald van der Pol
G6 (FR)	Bernard Tuy
GRnet (GR)	Dimitrios Kalogeras
ACOnet (AT)	Wilfried Woeber
Telebit Ericsson (DK)	Simon Nybroe/Alex van der Plas
INFN (IT)	Tiziana Ferrari
Uninett (NO)	Olav Kvittem
SWITCH (CH)	Simon Leinen
RedIRIS/CSIC (ES)	Celestino Tomas
CESNET (CZ)	Ladislav Lhotka
DFN (DE)	Juergen Rauschenbach
BME (HU)	Janos Mohacsi
UCL (UK)/SOTON (UK)	Peter Kirstein/Tim Chown
ILAN (IL)	Yaron Zabary
ICM (PL)	Wojtek Sylwestrzak

11.3.2 Field tests

Technical set-up

The goal of the field tests is to connect the various NRNs using ATM PVCs. The equipment consists of one Ericsson Telebit IPv6 capable router. This router is installed in the TEN-155 Amsterdam PoP where it will connect the participating NRNs in a star-like topology.

Planned timetable

The tests will last for the lifetime of the Quantum Test Programme.

Work done, progress so far

So far not much progress has been made, the Ericsson Telebit router has been installed in the PoP, but interoperability problems on the ATM have considerably delayed progress. Changing SDH configuration details on the Telebit router has solved these problems very recently.

11.3.3 Lab tests

Technical set-up

A basic test network has been set up at the Telebit laboratories. The setup allows simulating the proposed production network and it is possible to integrate the test routers into the test set-up as both backbone routers and as access routers for NRNs. This flexible test set-up allows both backbone issues, such as IBGP and internal routing, as well as NRN connectivity to be examined. The current set-up is depicted in the following figure:

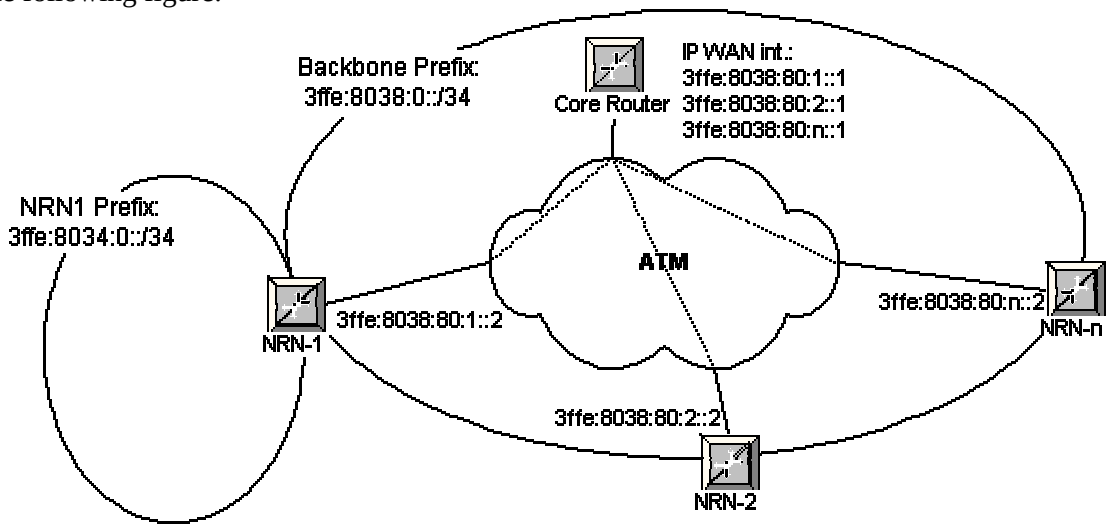


Fig. 11.1 IPv6 Laboratory Test Set-up

11.4 Current Status

Currently the test network consists only of Telebit routers. The NRNs are connected via IPv6 over ATM to the core router. Each NRN uses a single PVC with a capacity of 1204 cps. IP routing is enabled by BGP between core and NRN access routers. The core router performs all the IP routing between NRNs. The only vendor, other than Telebit, who is participating in the test network is Cisco. A Cisco 3600 multi protocol router, with Ethernet and ATM 25 module, will be connected as one of the NRN access routers to the core Router. Unfortunately there have been some problems with the Beta software of Cisco that supports IPv6 and the ATM 25 module. As soon as the fix is available the Cisco will be connected to the test network.

11.5 Scheduled Activities

Compatibility
As soon as possible, a Cisco router will be connected in order to perform compatibility tests between the Cisco router and Telebit routers.

DNS
DNS for the pTLA assigned to the IPv6 experiments will be implemented in the test network in October 1999.

Performance
Simple throughput tests will be performed for each of the routers using Telebit's internal traffic generator.

Compliance
Each router will be tested for compliance with the latest specifications using the http://www.tahi.org/ testing tool.
Using the Merit Multi-Threaded Routing Toolkit (http://www.merit.edu/) the routing protocols will be tested.

Interoperability
These tests will be performed in the test set-up that is as close to the set-up of the expected production network as possible, and will focus on the issues that need to be resolved before NRNs are able to peer with the backbone. Possible problems could be interoperability between routing protocols, IPv6 over ATM, IPv6 over other media (LAN) and IP6-in-IP4 tunnel implementations.

12 Glossary of Terms

AAA	Authentication, Authorisation and Accounting
ABR	Available Bit Rate
ARP	Address Resolution Protocol ??
AS	Autonomous System
ATC	ATM traffic class
ATM	Asynchronous Transfer Mode
BGP	Border Gateway Protocol
BGMP	Border Gateway Multicast Protocol
CAC	Call Admission Control
CAR	Committed Access Rate
CBR	Constant Bit Rate
CB-WFQ	Class Based Weighted Fair Queuing
CE	Customer Edge
CLIP	Connection Less Internet Protocol
DBR	Deterministic Bit Rate
DNS	Domain Name Service
DSCP	Differentiated Service Code Point
DMTF	Distributed Management Task Force
DVMRP	Distance Vector Multicast Routing Protocol
DWDM	Dense-Wave Division Multiplexing
FIN	Close connection (TCP message)
GPS	Global Positioning System
iBGP	internal BGP
IETF	Internet Engineering Task Force
ILMI	Interim Local Management Interface
IP	Internet Protocol
ISP	Internet Service Provider
LANE	Local Area Network Emulation
LIS	Logical IP sub-network
LDP	Label Distribution Protocol
LRU	Least Recently Used
LSP	Label Switched Path
MAC	Media Access Protocol
MASC	Multicast Address-set Claim
MBGP	Multiprotocol BGP
MBS	Managed Bandwidth Service (of the TEN-155 network)
MPLS	Multi Protocol Label Switching
MRTG	Multi Router Traffic Grapher
MSDP	Multicast Source Distribution Protocol
NHRP	Next Hop Resolution Protocol
NRN	National Research Network
OSPF	Open Shortest Path First
PATH	Path message for RSVP
PCR	Peak Cell Rate
PE	Provider Edge
PHB	Per Hop Behaviour

PIM	Protocol Independent Multicast
PIM-SM	PIM Sparse Mode
PNNI	Public Network to Network Interface
PoP	Point of Presence
PPP	Point to Point Protocol
pTLA	Pseudo Top Level Area
PVC	Permanent Virtual Circuit
QoS	Quality of Service
QTP	QUANTUM Test Programme
RAP	RSVP Admission Policy
RD	Route Distinguisher
RED	Random Early Discard
RESV	Reservation message for RSVP
RFC	Request For Comments
RRR	Routing for Resource Reservations
RSVP	Resource ReServation Protocol
RTCP	RTP Control Protocol
RTFM	Real Time Flow Measurement
RTP	Real Time Protocol
RTT	Round Trip Time
SBR	Statistical Bit Rate
SCFQ	Self Clocking Fair Queueing
SCR	Sustainable Cell Rate
SDH	Synchronous Digital Hierarchy
SLS	Service Level Specification
SNMP	Simple Network Management Protocol
SVC	Switched Virtual Circuit
TCP	Transmission Control Protocol
TOS	Type Of Service
TVC	Tagged Virtual Circuit
UDP	User Datagram Protocol
UNI	User-Network-Interface
VBR	Variable Bit Rate
VC	Virtual Channel
VP	Virtual Path
VCC	Virtual Channel Connection
VPC	Virtual Path Connection
VPN	Virtual Private Network
VRF	VPN Routing/Forwarding
WDM	Wave Division Multiplexing
WFQ	Weighted Fair Queuing
WG	Working Group
WRED	Weighted Random Early Discard