ISTITUTO NAZIONALE DI FISICA NUCLEARE

CNAF Bologna

INFN/TC-99/24
28 Ottobre 1999

ARCHITECTURE AND PERFORMANCE OF A TAG SWITCHING WIDE AREA NETWORK

T. Ferrari[1], J.M Uzè[2], C. Vistoli[1]

[1)]INFN–CNAF, V.le Berti Pichat 6/2, I–40127 Bologna, Italy
[2)]French Academic and Research Network, RENATER, France

Keywords: MultiProtocol Label Switching (MPLS), Tag Switching, ATM, WAN, IP routing, scalability, BGP, OSPF, performance.

*Published by **SIS–Pubblicazioni***
Laboratori Nazionali di Frascati

Architecture and Performance of a Tag Switching Wide Area Network

Tiziana ferrari <tiziana.ferrari@cnaf.infn.it>[1],
Jean-Marc Uzé <uze@renater.fr>[2],
Cristina Vistoli <cristina.vistoli@cnaf.infn.it>[1]
[1]Istituto Nazionale di Fisica Nucleare, INFN-CNAF, Italy
[2]French Academic and Research Network (RENATER, France)

Abstract

Tag switching and MPLS (MultiProtocol Label Switching) combine IP routing flexibility with the efficiency of cell switching techniques to address the need of scalable infrastructures, of a wider range of services and of the support of advanced applications. We present the model and the configuration of several tag switching networks in the local and wide area. In particular, we focus on the design of a scalable IP architecture and its application in a wide area testbed based on tag switching, a MPLS implementation by CISCO. The protocol and the network performance have been analysed in terms of functionality, software stability, round trip time, route recovery time and throughput. Results show that tag switching is a promising and viable technique for the implementation of scalable and integrated networks.

Keywords: MultiProtocol Label Switching (MPLS), Tag Switching, ATM, WAN, IP routing, scalability, BGP, OSPF, performance.

1 Introduction

Network technologies need to be enhanced to support new applications and to cope with a larger number of users. However, increasing the availability of network resources is not enough to achieve such a goal, since new applications also require scalable network architectures, increased packet forwarding capabilities and a wider range of services. Tag switching, a specific implementation of MultiProtocol Label Switching (MPLS), is a new technology which addresses these needs by combining the flexibility of the IP protocol routing scheme with the efficiency of cell switching.

We present the design and implementation of a wide area ATM backbone based on tag switching and connecting several exterior networks. A set of performance tests were carried out to estimate functionality and performances of tag switching to analyze its applicability for the implementation of an ATM-based backbone.

The label-switching test program [1] was developed by the task force *tf-ten* [2] of TERENA [3]. The ATM European backbone *JAMES* [4] (connecting European National Research Networks and co-founded by the European Community) was used as core infrastructure for the tests.

Section 2 introduces the main tag switching features, while in section 3 we provide several examples of tag switching network set-up in the local and wide area, in particular we analyze the protocol functionality (par. 3.1) and the tunnelling mechanism to run tag switching on top of an ATM public infrastructure offering CBR PVCs (par. 3.2). The design of a scalable tag switching network and of its IP architecture is illustrated in par. 3.3. Then, section 5 describes the performance tests carried out in a wide area European testbed to measure round trip times (par. 5.1), route recovery times (par. 5.2) and the comparison of TCP and UDP throughput achieved with and without tag switching (par. 5.3 and 5.4). Section 6 provides an example of configuration and use of traffic engineering. In section 7 we draw some conclusions.

2 Protocol overview

Tag switching and MPLS [5, 6], integrate layer 2 switching with layer 3 routing. They are designed for high speed networks to merge ATM performances IP routing flexibility in a single infrastructure. The idea is to set up ATM VCs to avoid per hop routing. Protocol implementations can be based on flow detection (e.g. *IP switching*) or IP routing (e.g. *Tag Switching*, a proprietary implementation by CISCO [1]).

Tag switching [7, 8] assigns a label (*tag*) to each packet depending on its final destination. In a conven-

[1]Tag switching is the specific protocol implementation chosen to carry out the test program because of the availability of tag switching beta software versions for both routers and ATM switches.

tional IP network packets are processed by each router on the path. On the other hand, with tag switching the ingress router assigns a label to each destination network and packets are then switched towards the egress router thanks to he tag.

A tag switching network consists of a core of tag switches with tag edge routers at the periphery. Tag edge routers and tag switches use standard routing protocols (BGP, OSPF...) to build routing tables. Then, a tag – represented by the VCI in ATM networks – is assigned to each route in the tag network. Tag Distribution Protocol (TDP) achieves two important goals: firstly it distributes the tag information among switches and routers in the tag cloud and secondly it sets up the ATM VC connection mesh in the backbone. To support TDP a control PVC is automatically set-up between adjacent tag devices during the initialization phase.
For a comprehensive description of MultiProtocol Label Switching (MPLS) refer to [5].

3 Design of a label-based switching network

In this section we present several testbeds which provide examples of label-based switching networks with different degrees of complexity, which show the functionality of the protocol both in the local and in the wide area.

3.1 Functionality in the local area

The configuration of a simple local area network based on label switching is illustrated in figure 1. Two Cisco 7505 routers (tag edge routers) are connected to a LightStream 1010. The IP routing protocol is OSPF on all systems.

On Cisco 7505 "tag-top" three IP networks (14.0.0.0, 15.0.0.0, 16.0.0.0) are associated with corresponding ethernet interfaces, while Cisco 7505 "tag-bottom" is configured with one IP network (7.0.0.0) subnetted on Ethernet interfaces (7.1.0.0, 7.2.0.0, 7.3.0.0 and 7.4.0.0). OSPF was selected as tag cloud routing protocol. An ATM interface running tag switching can also support traditional static PVPs for classic IP over ATM traffic, like in router "tag-bottom".

For each network announced by an egress router a TVC is created to connect it to any other ingress router, i.e. there is no IP route aggregation. This set-up is not suitable for a wide area network with many prefixes. To achieve scalability in a wide area environment we need a hierarchical structure like the model presented in paragraph 3.3.

3.2 Tunneling

Figure 2 provides an example of simple wide area label-based switching network. We interconnected two sites (the Idris laboratory, Orsay, and the University of Jussieu, Paris) through an ATM PVP (CBR service, 10 Mbps of bandwidth).

On both user sides one CISCO 7505 router and one LS1010 ATM switch run tag switching. On the Orsay side a DEC GIGAswitch/ATM running IP switching (Ipsilon router) was connected to the CISCO router through the ethernet interface eth 3/2. On the other hand, on the Jussieu side ATM connectivity was achieved through a Fore ASX 200-BX switch performing VP switching on VPI 5. GDC switches provided ATM connectivity on the public network side. OSPF was the IP routing protocol running in the tag cloud.

This testbed gave the opportunity to try connectivity between remote sites through an ATM public operator offering CBR VP service. Tunneling was the pre-requisite for the implementation of the wide area set-up described in paragraph 5.

In order to run tag switching on a public ATM infrastructure like this, tag switching VCs (including the control VC) need to be tunneled into the PVP offered by the Public Network Operator. In this way tag switching is completely transparent to the public network ATM equipment. Interoperability with CBR service is achieved by shaping cells at 10 Mbps (the VP bandwidth) on the physical ATM interface connecting the LS1010 switch to the public network.

A dedicated VP must be used for tag switching tunneling since other signaling protocols (e.g. UNI and PNNI) cannot be tunneled on the same VP.

3.3 Scalability in the wide area

An experimental overlay network connecting several countries: Austria (ACONET), France (RENATER), Germany (DFN), Italy (GARR), Spain (REDIRIS) and Switzerland (SWITCH), was designed to carry out a tag switching European test. The supporting network infrastructure was the European ATM network *JAMES* [4].

The overlay network consists of ATM CBR permanent virtual circuits with different capacities depending on the link: 4515 cells/sec on the France-Spain and France-Germany link, and 4750 cells/sec in the rest of the overlay network. The infrastructure is a tag switching cloud (corresponding to the area inside the dashed line in picture 3) and a set of peripheral non tag switching local networks, one per country. The loop between Austria, Italy and Switzerland was configured to verify the correctness of route computation between the three sites.

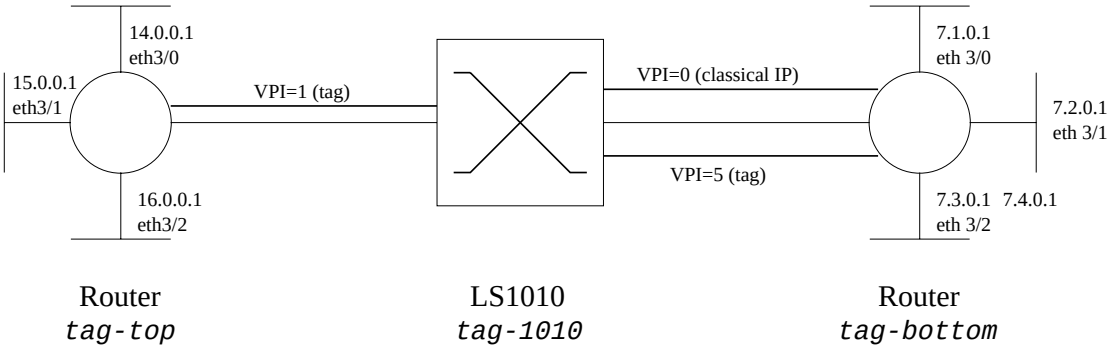

Figure 1: Network infrastructure for tests in the local area.

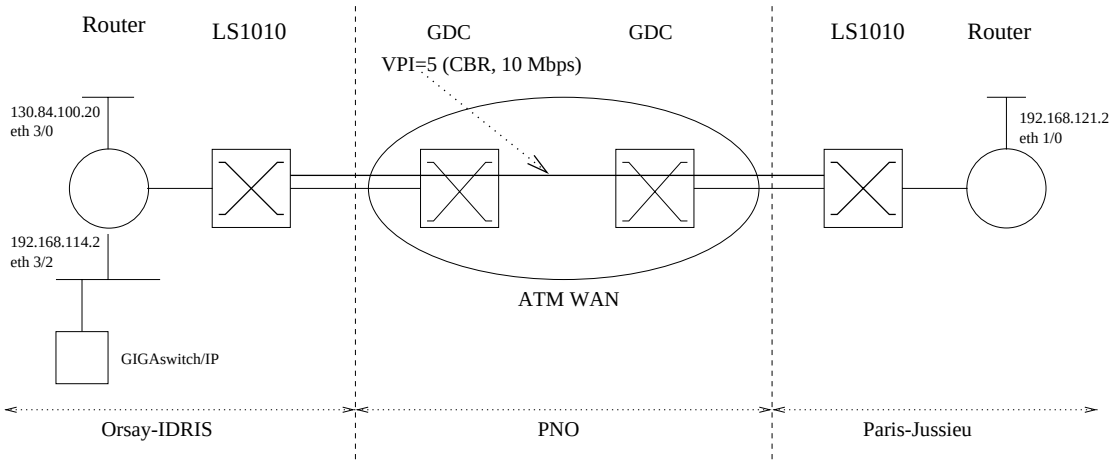

Figure 2: Network infrastructure for tunneling tests in the wide area.

The tag cloud was made of LightStream 1010 switches and CISCO routers of the 7500 and 7200 series, all running tag switching beta software. The ATM switches represent the core of the tag backbone, because they provide very high performance switching, while routers are deployed in the periphery to connect to external networks. In each country we set up a tag core switch and a tag edge router. The routers are connected to their adjacent switch with a STM1 link, while switches are connected through the JAMES infrastructure. The tag switching protocol is entirely tunneled in the JAMES infrastructure. In this way tag switching is completely transparent to the ATM equipment on the public network operators side.

Tag switching needs an IP routing protocol in order to exchange all routes through the backbone and to set up the corresponding Tag VCs (TVCs). On each tag apparatus an IP loopback address was configured to be used by the TAG Distribution Protocol (TDP). Furthermore, a single area OSPF routing process was configured in each tag switching apparatus. In order to establish a full mesh of TAG VCs in the core network, TDP uses a dedicated control PVC which is automatically configured at initialization between adjacent devices.

Routers (C7500 and C4500) and workstations, were connected to the tag cloud to add external networks. Hosts were used for performance tests (section 5) by generating TCP and UDP memory-to-memory traffic. The three workstations (a Sun Ultra in France and Switzerland and a Sparc Station 10 in Italy) were connected to the backbone respectively through a Fast Ethernet, an Ethernet and an ATM network interface.

4 IP architecture design

Several IP architectures can be devised to interconnect external networks through a tag backbone. For example OSPF can be configured in each router and switch as unique IP routing protocol. In this way a TVC is created from each tag edge router to each external route announced to the backbone because of

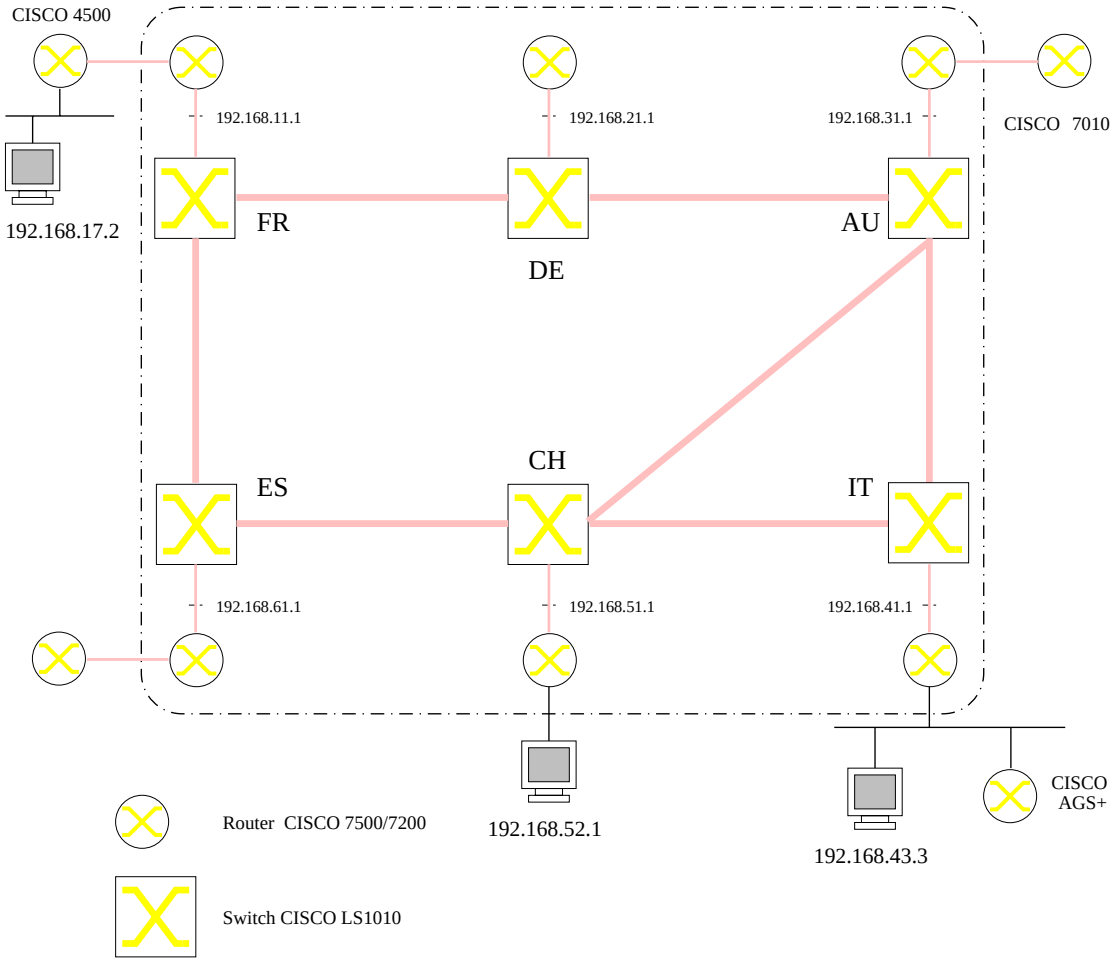

Figure 3: Tag switching scalability in an extended wide area network.

external route redistribution into the tag cloud (see par. 3.1).

Scalability is a fundamental requirement for a wide area network. To achieve this requirement and to carry out a comprehensive tag switching test we configured the following set-up: BGP in the edge routers and interior routing protocol OSPF inside the tag cloud as shown in figure 4.

In the configuration of figure 4 tag routers (192.168.11.1, 192.168.21.1, 192.168.31.1, 192.168.41.1, 192.168.51.1 and 192.168.61.1) belong to the same autonomous system (AS) and run exterior BGP sessions with the local exterior routers, whose networks are associated to a different AS number.

A complete mesh of internal BGP sessions is set-up between tag edge routers. These IBGP sessions use tag switching virtual circuits (TVCs) established by tag. These IBGP sessions permit to exchange external routes between tag edge routers. In this way a *hierarchical* IP routing configuration is achieved.

This architecture is more scalable, since IP datagrams with destinations reachable through the same tag edge router, use a unique TVC. For instance traffic between France (AS 110) and Spain (AS 160) is forwarded through a single TVC set up between the French tag router (192.168.11.1) and the Spanish (192.168.61.1).

5 Performance

A set of performance tests was done to verify the correct functionality of routing and switching of tag switching equipment in the testbed of figure 4 and to compare results achieved in the same network testbed with and without tag switching. Without tag switching, tests were performed by configuring static IP routes between the routers directly connected to the switches, in this case static ATM circuits were used

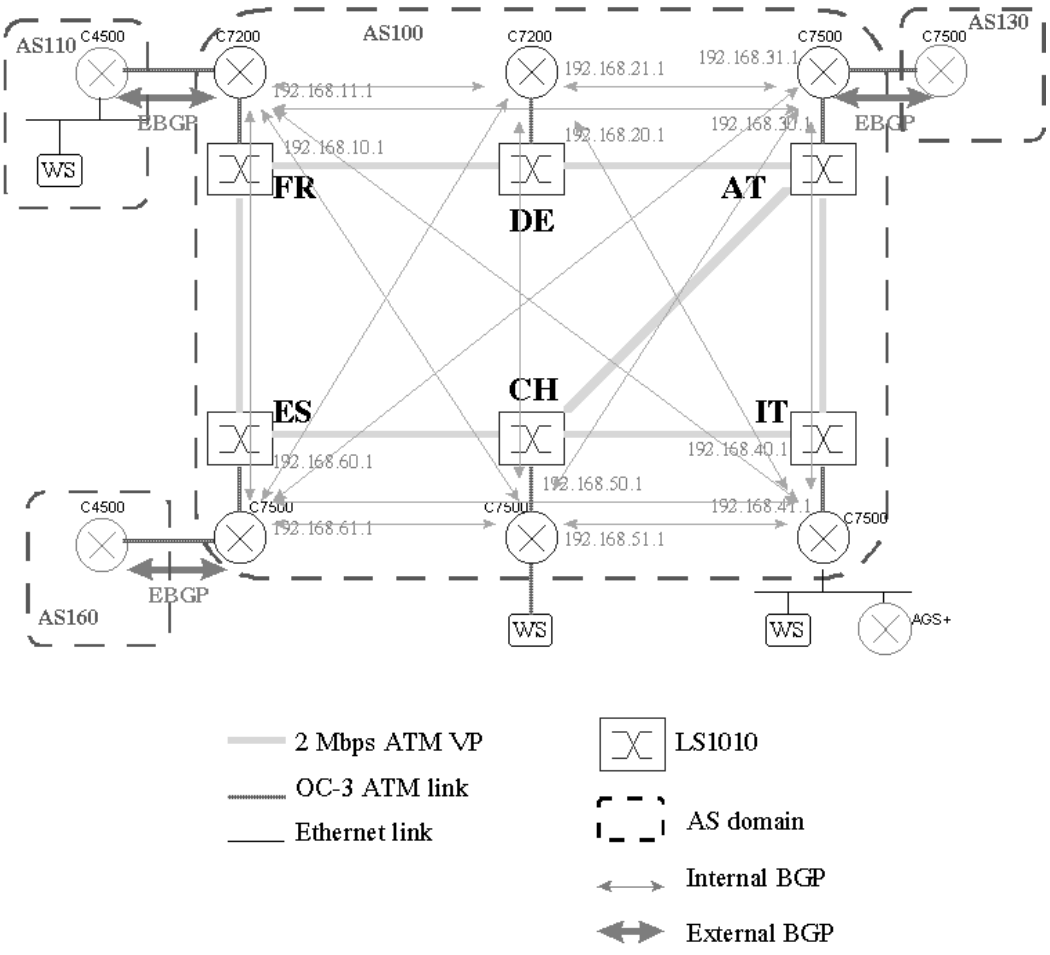

Figure 4: IP infrastructure for scalability.

as point-to-point permanent links with classical IP on ATM. *Netperf 2.1* [10] is the application used for measurement of TCP throughput, while *Mgen 3.1* [11] was deployed to generate UDP streams at a given user supplied rate.

Single and multiple TCP or UDP connections with and without tag switching were set up between France, Italy and Switzerland. Results obtained with and without tag switching have been compared. The test parameters are:

1. round trip time,
2. route recovery time,
3. throughput of TCP connections,
4. packet loss for UDP streams,
5. average CPU utilization in the routers.

Traffic was generated from three end-systems located in Italy (192.168.43.3), Switzerland (192.168.52.1) and France (192.168.17.2) (respectively a Sparc Station 10 and two Sun Ultra, all mounting Solaris 2.5.1).

Measuring the real benefit of tag switching in terms of packet forwarding efficiency and amount of protocol overhead is not easy: It requires a loaded high speed backbone with a complex IP topology like in a production environment. On the contrary in our testbed the IP protocol overhead is not a critical factor because of the network simplicity, and the link capacity is not enough to load the routers.

This is why the primary goals of our tests are: the design of a scalable tag switching infrastructure, the analysis of tag switching functionality and of its applicability and the measurement of traffic performance.

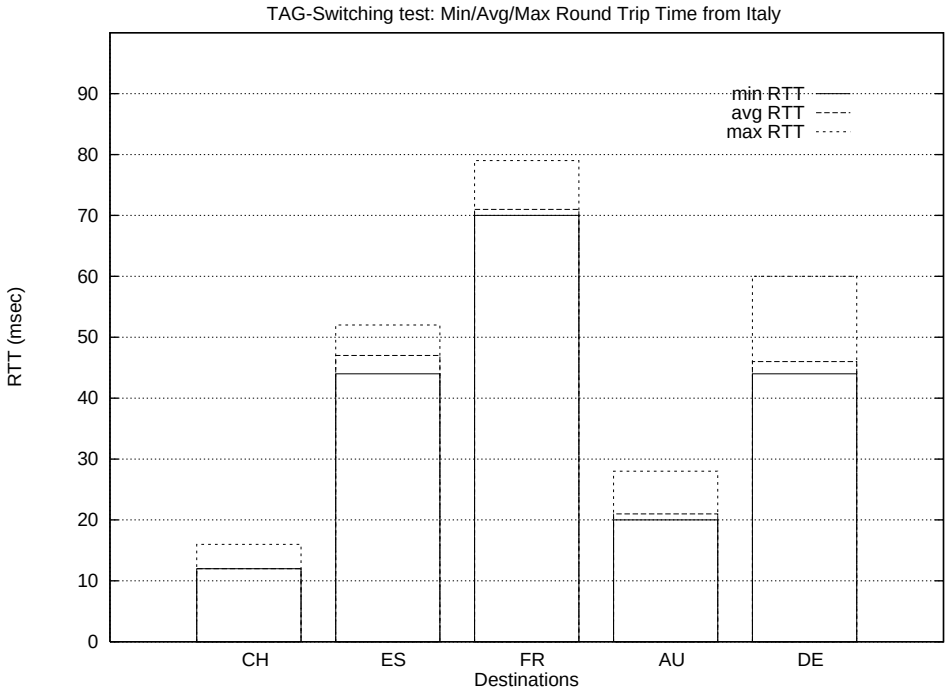

Figure 5: Round trip times between edge routers of the tag cloud.

5.1 Round trip time

Picture 5 illustrates the round trip times from edge router 192.168.41.1 to any other edge router. Differences in RTT are mainly due to different propagation delays between the couples of routers.

RTTs are the same both with and without tag switching. This shows that the packet forwarding speed achieved with tag is as good as the one obtained when edge routers are physically adjacent, since permanent ATM CBR connections are equivalent to point-to-point physical links.

Tag edge router packet forwarding is not faster than in legacy routers because upon receiving input packets, tags need to be assigned them. RTTs were the same because in our testbed we had no routers inside the tag cloud. The real benefits of tag switching are the possibility of integrating routers with switches, the IP architecture simplification and the availability of new services like class of service support and traffic engineering (section 6).

5.2 Route recovery

In order to quantify the route computation efficacy in an unstable tag switching environment, we measured the *route recovery time*, i.e. the time necessary to the router to compute the path towards a given destination when link failure occurs.

During the tests link failure was artificially generated in some switches by shutting down the ATM subinterface corresponding to a given destination. Link failure was generated several times in different parts of the backbone.

In the core ATM switched OSPF recovery time varied approximately in the range [12..38] sec. We were able to observe the recovery time after link reactivation from the external network point of view (they were connected through BGP). The aggregate time that external routers need to recalculate the best routes and the backbone to establish the right TVCs, was in the range [10..40] sec. Measuring the recovery time after link failure was not possible in our testbed and we can't deduce it from the previous one. This is argument of further study.

5.3 TCP Performance

Tests were done both generating half duplex connections (i.e. connections with a single source and destination) and also with full duplex streams (i.e. streams in which end-systems act as sender and receiver at the same time). The following paragraphs illustrate the results obtained in the two environments.

TCP half duplex connections Performances have been gathered for concurrent half duplex TCP flows. Picture 6 and 7 show the throughput obtained by a single TCP stream on the path from Italy to Switzer-

land and from France to Italy respectively. Several values have been collected for different socket buffer sizes (buffer sizes were set consistently on both the sending and receiving side). 4750 cells/sec (the capacity of the ATM link Italy-Switzerland) correspond to 1.78 Mbps of application data throughput (considering an IP MTU of 1500 bytes). On the other hand, 4515 cells/sec (the capacity of the link between Italy and France) give 1.68 Mbps.

Results show that in both tests the ATM link capacity is available. The socket buffer sizes are not relevant in the first case (see figure 7) since the minimum RTT is 12 msec, while in picture 6 throughput decreases for small socket sizes and bandwidth utilization turns to be rather inefficient, because of the *stop and wait* behaviour, which is due to the combination of small buffers and large RTT (about 70 msec).

The throughput achieved by connections to France is less than to Switzerland because of the smaller PVC capacity on the path Italy-France. Results refer to tests with message size equal to 16,000 bytes. Tests with different message sizes showed that such a parameter is not relevant in our testbed.

Bandwidth utilization in an ATM PVC infrastructure with static IP routes is *less* efficient than with tag switching. Picture 6 and 7 compare the two results: *With* tag switching streams to Switzerland reach 1.75 Mbps against 1.69 Mbps. We had the same for streams from the workstation in Italy to the one in France, in this case *with* tag switching throughput is 1.63 Mbps against 1.57 Mbps.

The throughput gain is due to a different encapsulation scheme used in the tag switching test and classical IP test: TVCs use AAL5 VC based multiplexing encapsulation, while ATM PVCs deploy *AAL5 LLC-SNAP encapsulation* [12]. With LLC-SNAP 8 bytes (LLC header plus SNAP header) are added to the IP PDU when it is encapsulated into the AAL5 CPCS PDU payload. On the other hand, with *VC based multiplexing* no overhead is added at all with a consequent performance gain which depends on the IP PDU size distribution, i.e. on the number of padding bytes added in the AAL5 CPCS PDU. [2]

This set of tests was repeated with even more concurrent connections: The number of flows did not influence the aggregate throughput achieved either with or without tag switching.

[2] PVCs can be configured to use several encapsulation schemes, VC based multiplexing included, so this tag throughput gain depends on the PVC configuration.

TCP full duplex connections Full duplex traffic consists of two concurrent TCP connections in opposite directions. The performance test was repeated with and without tag switching and the results are compared in figure 8, which also plots the throughput obtained in both directions for a half duplex streams to provide a term of comparison.

With full duplex streams the performances achieved in each direction is less than in the half duplex case. Performance loss depends on the socket size. The maximum is achieved with socket buffer dimensions around 128 kbytes: aggregate throughput loses approximately 250 Kbps. Full duplex connections achieve less throughput both with and without tag.

In this test performances in the two opposite directions were not the same, i.e. the direction from Italy to Switzerland was less penalized then the opposite one. This problem requires further investigation. In addition, with full duplex streams between France and Italy we had high peaks in CPU utilization (a 5 sec average equal to 15 %) in router 192.168.41.1 (C7200). This was probably due to the beta software versions running on the routers, which are still under development, and require further investigation.

5.4 UDP performance

With Mgen one or more half duplex UDP streams can be activated by specifying a given application datagram rate. We used Mgen to produce UDP traffic at increasing data rates between end-systems in France and Italy.

Performances are rather good from Italy to France, since as expected, packet loss starts when the datagram rate overcomes the link capacity. On the other hand, results in the opposite direction from France to Italy are not as satisfactory as these, since datagram loss appears with data rates equal or bigger than 0.8 Mbps and the packet loss rate is directly proportional to the application data rate.

We think that this is not due to the protocol itself, but to the beta versions running on the routers, especially on C7200. This hypothesis is confirmed by the peaks in CPU utilization (up to 19 %) in router 192.168.11.1, while CPU utilization on C7500 was always less than 3 %. When repeating the same test between Italy and Switzerland, we had no packet loss in *both* directions for flow rates not exceeding the ATM connection capacity (in this case on the path we just had routers C7500). This confirms the correctness of tag switching functionality in presence of UDP traffic and the need of stable software versions.

6 Traffic engineering

Traffic engineering, one of the main tag switch-

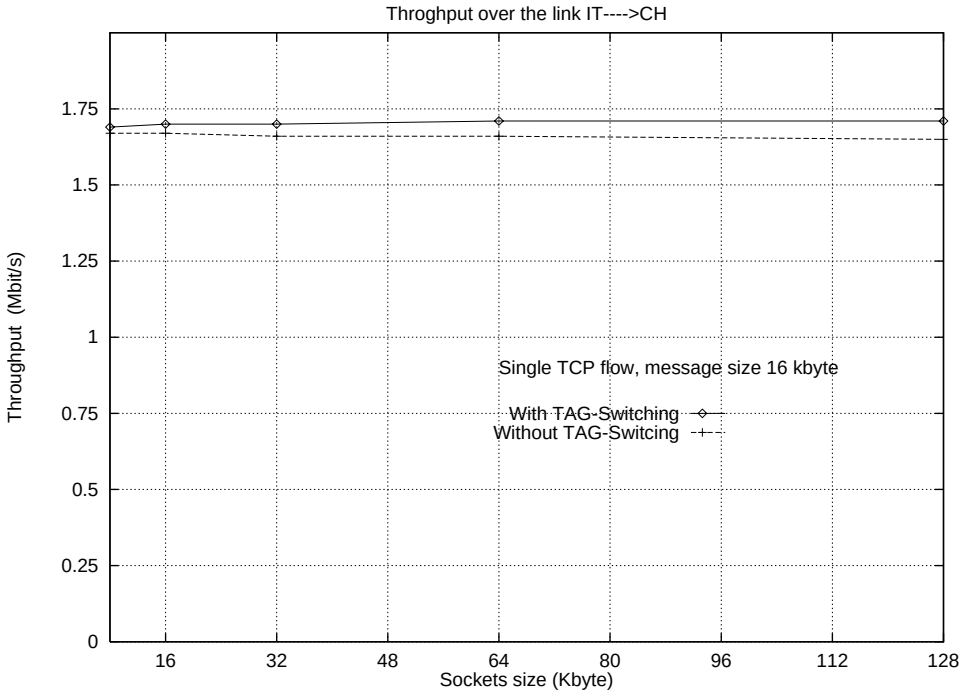

Figure 6: Single half duplex TCP stream from Italy to Switzerland.

ing features, allows one or more streams, specified through filters, to be forwarded according to a predefined path. It gives the opportunity to tailor and balance traffic in the network so that standard routing information can be overridden and well defined streams can be routed differently. The preferential path can be defined as a unidirectional tag switching tunnel to be configured in the ingress router. The rest of the tunnel is automatically and dynamically configured by a signalling protocol based on RSVP.

The following is an example of tunnel from router 192.168.41.1 to 192.168.31.1. Picture 9 shows the tag switching tunnel (red line) used to route traffic to the Austrian network 192.168.33.0 through Switzerland instead of the direct default link Italy-Austria used to reach the Austrian networks.

```
interface Tunnel2000
ip unnumbered Loopback0
transmit-buffers backing-store
tunnel mode tag-switching
tunnel tsp-hop 1 192.168.40.1
tunnel tsp-hop 2 192.168.50.1
tunnel tsp-hop 3 192.168.30.1
tunnel tsp-hop 4 192.168.31.1 lasthop
```

Traffic engineering works correctly. Preferential traffic to the selected network is routed by overriding the standard route entry:

```
show ip traffic-engineering
 Filter 1: egress 192.168.33.0/24
 Tunnel2000 route installed
 Installed traffic engineering routes:
 Codes: T - traffic engineered route
 T 192.168.33.0/24 (override of routing table entry)
 is directly connected, 00:59:30, Tunnel2000
```

7 Conclusions and future work

According to the test results, tag switching is a promising and applicable technique. Even if software implementations need to be improved, the protocol shows good functionality in terms of routing stability, interoperability with ATM, protocol tunneling on ATM PVCs, traffic engineering and maximum bandwidth utilization with both TCP and UDP.

The deployment of label-based switching techniques like tag switching combined with a carefully designed hierarchical IP architecture, is viable and achieves scalability in the wide area. These results are encouraging also for the future support of differentiated services in the Internet through MultiProtocol Label Switching.

A more detailed study of performance on a loaded tag switching network and the comparison with the results achieved in an equivalent legacy IP infrastructure

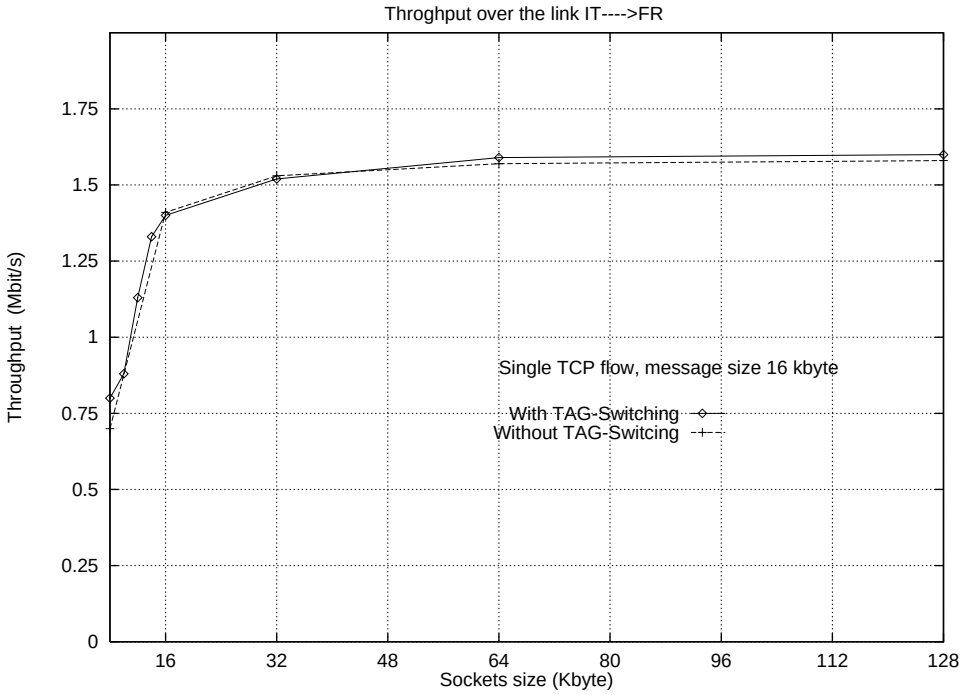

Figure 7: Single half duplex TCP stream from Italy to France.

needs further investigation. The test of the interesting advanced features like: VC merging, VPN, class of service support and PIM with multicast tag switching, need to been analysed for a more comprehensive understanding of the technology.

8 Acknowlegements

The test program was carried out thanks to the collaboration and support of Cisco Systems, which supplied both hardware and beta software versions for routers and switches, and of the task force *tf-ten*. The collaboration of all the test participants was fundamental to set up a European infrastructure: Simon Leinen (Switch, CH), Guenther Schmittner (Johannes Kepler University, AT), Robert Stoy (RUS, DE) and Celestino Tomas (Rediris, SP). A special acknowledgment goes to Alessandro Canzian for his valuable work during the performance tests.

References

[1] *Label-Based Switching Experiment*, `http://www.renater.fr/jmu/jameslbs.html`.

[2] *Task Force TEN Home page*, `http://www.dante.net/tf-ten/`.

[3] *The Trans-European Research and Education Networking Association*; `http://www/terena.nl`

[4] *Joint ATM Experiment on European Services* `http://www.labs.bt.com/profsoc/james/`.

[5] *MultiProtocol Label Switching* `http://www.ietf.org/html.charters/mpls-charterhtml`

[6] *Multi Layer Routing* `http://infonet.aist-nara.ac.jp/member/nori-d/mlr/`

[7] Cisco Systems, *Scaling the Internet With tag switching*, white paper, `http://www.cisco.com/warp/public/732/tag/pjtag_wp.htm`

[8] Cisco Systems, *Tag Switching*, `http://www.cisco.com/warp/public/732/tag/tag_resources.html`.

[9] B.Halabi, J.Lawrence, *Tag Switching in Service Provider ATM Networks*, white paper.

[10] *Netperf*, `http://www.cup.hp.com/netperf/DownloadNetperf.html`

[11] *Mgen*, `http://tonnant.itd.nrl.navy.mil/ipresearch/mgen.html`.

[12] Juha Heinanen, *Multiprotocol Encapsulation over ATM Adaptation Layer 5*, RFC 1483, July 1993.

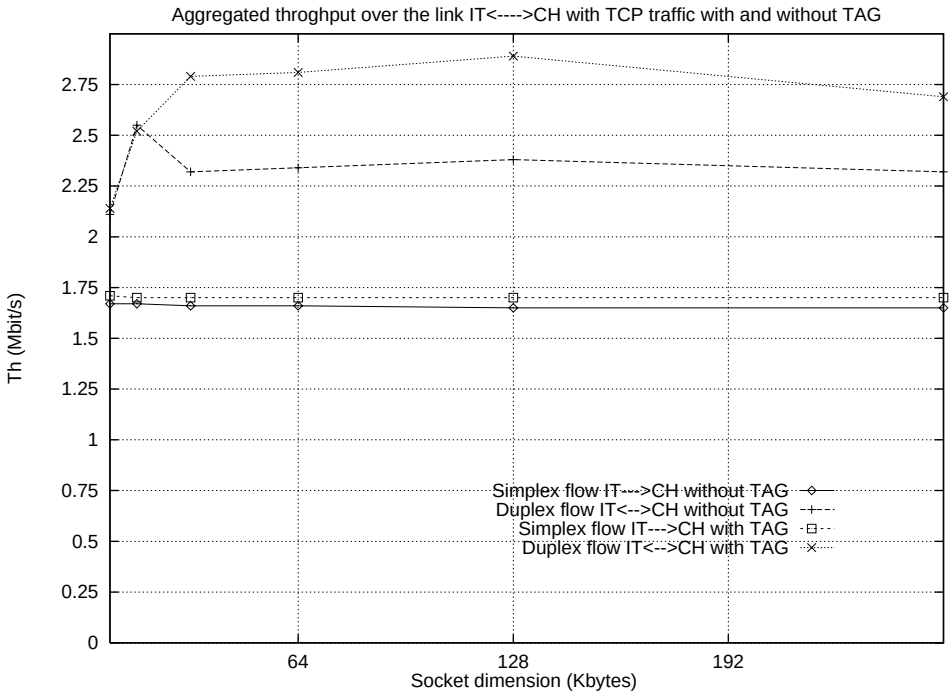

Figure 8: Full duplex connection between Italy and Switzerland.

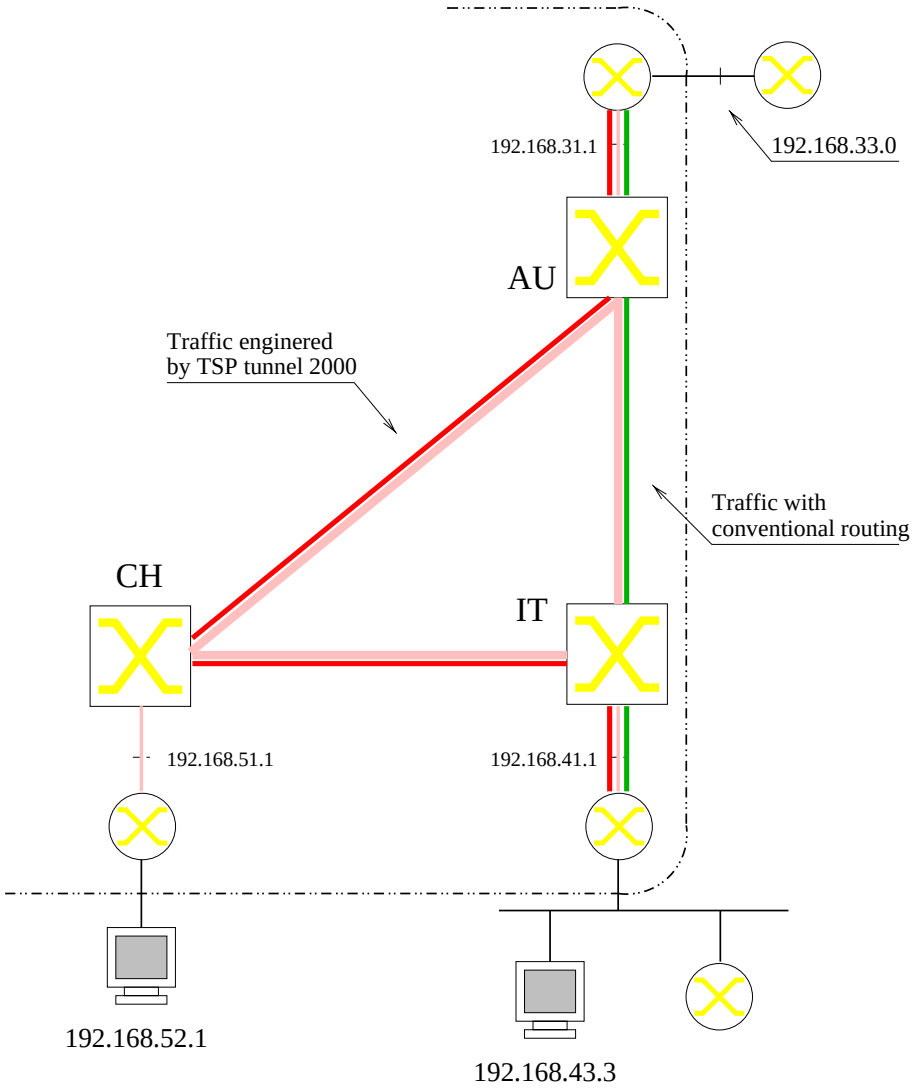

Figure 9: Tag switching tunnel used for traffic engineering.