Project Number:	**IST-1999-11253-TEQUILA**
Project Title:	**Traffic Engineering for Quality of Service in the Internet, at Large Scale**

IP Fast Translator (IFT) – TEQUILA white paper

Author:	**FTR&D**
Contribution File Name:	**IFT_whitepaperv5.doc**
Version:	**final**
Company:	**FTR&D**
Date:	**Thursday, October 29th 2001**
Distribution:	**public**

The TEQUILA Consortium consists of:

Alcatel	Coordinator	Belgium
Algosystems S.A.	Principal Contractor	Greece
FT R&D	Principal Contractor	France
IMEC	Principal Contractor	Belgium
NTUA	Principal Contractor	Greece
RACAL	Principal Contractor	United Kingdom
UCL	Principal Contractor	United Kingdom
TERENA	Assistant Contractor	The Netherlands
UniS	Principal Contractor	United Kingdom

Table of Contents

IFT-based routers: the best of both worlds?

IFT[1]-based routers: the best of both worlds?

1 INTRODUCTION

There is a constant pressure on the capabilities of IP routers from the networking market for:

- Performance increases, to meet the demands of a continuously increasing number of users and their traffic volumes. The line rate of today's IP networks is reaching 10Gbit/s in the core and 1Gbit/s at the access. The number of entries in a full Internet routing table was 40K in late '98, it has reached 100K in early '01.
- New features and services. Traffic engineering and DiffServ are generally considered as useful tools for the deployment of value-added IP service offerings, where guarantees on the level of quality is required by customers of such services. Security is also a prime concern, as well as Multicast, Voice over IP, and Virtual Private Networks services.

Although commercially available routers address day-to-day networking requirements, in terms of switching performances, building proof of concept networks and services prior to their operation in real networks requires more than what is currently available on commercial IP routers or PC-based ones.

During the initial design of the architecture, which was presented at the first project review (September 2000), the project has evaluated the necessity of an experimental (IP edge) routing platform. The experimental routing platform is based upon the use of the IP Fast Translator (IFT) which has been developed by France Telecom.

The necessity for the IFT resides from the requirement to have a high-speed DiffServ edge router, i.e. a platform able to process at very high speed the full DiffServ functionality (with thousands of per-flow traffic conditioners, policers, markers, etc) to asses that Tequila can be integrated on a system with known performance characteristics. The commercial available (Cisco) routers do provide the processing speed, but lack the functionality required to evaluate the first project objectives. The Linux routers, enhanced by the project, provide full-blown DiffServ (edge) functionality, but not at the required processing speed or performance predictability. The IFT combines a hardware-based fast forwarding engine with (part of, i.e. the essential) DiffServ edge functionality.

Thus, developing a new router concept that addresses flexibility, performances and scalability issues fits in the 4th Tequila Project objective, which is "***To validate the theoretical models, architectures, algorithms and protocols developed by the project through experimentation using both simulation tools and physical testbed prototypes"*** as initially stated in the "Description of Work" of the Tequila project [TEQUILA] or recently reformulated as ***"Validate the above through both simulation and testbed experimentation using the appropriate (routing) platforms and enhancing these platforms as required for the validation of the above objectives..."*** [DGO]. A more detailed discussion on the ideas and innovations presented below, can be found in a paper entitled "High Router Flexibility and Performance by Combining Dedicated Lookup Hardware (IFT), off the Shelf Switches and Linux" submitted to the IFIP Networking 2002 conference [DRLL-VVD].

[1] IP Fast Translator

2 WHICH WORLDS: THE STATE OF THE ART

2.1 Linux-based routers

Over the past decade, Linux has attracted a considerable amount of interest from the research community, and also from industry. The numerous advantages of this Unix-based operating system, include:

- Its availability for low-cost PC-based platforms;
- Its development capabilities under open and distributed conditions: the full source code is freely available to users;
- A kernel that meets the current requirements of modern operating systems;
- An extensive set of networking features, including common network adapters (Ethernet, Serial Link…), ATM (Asynchronous Transfer Mode), traffic classification and conditioning, forwarding, queuing and monitoring mechanisms, software packages for routing (GateD, Zebra…) or configuration and management (SNMP (Simple Network Management Protocol), CORBA (Common Object Request Broker Architecture)…)

An extensive description of Linux features and a comprehensive bibliography can be found in [D2.1].

Within the context of the TEQUILA project, extensions have been added for implementing traffic control and engineering functionalities. Although of the majority of Intserv and Diffserv building blocks are available on Linux, the most important and yet missing feature is a complete Diffserv over MPLS (Multi-Protocol Label Switching) implementation. The detailed specifications of the TEQUILA data plane functionalities can be found in [D2.2r2].

Although Linux-based routers provide rich functionality which is easily extensible, there is a significant drawback when it comes to using them in operational environments or even in representative testbed experiments: their switching performances. The performance may be:

- difficult to predict since both the data and control planes mechanisms may run on the same CPU.
- bounded by the CPU performances themselves, and such performances may be poor compared to commercial routers.

2.2 Commercial routers and network processors

While commercial routers provide more than acceptable switching performances their main drawback within an experimental context is that researchers are limited to using standard versions of their controlling software as released by their manufacturers. It is impossible to get access to the source code of this software or to extend its functionality with experimental features without close commercial ties to the vendors. Thus, whenever a draft standard is not yet implemented or some functionality is missing it is necessary either:

- To rely on the roadmap of a given manufacturer for the introduction of new features;
- To add extra adaptation boxes, where it is feasible, for working around the present limitations.

Router architectures that are based upon a high performance CPU and (high speed) interface cards linked together by a shared bus, are no longer sufficient to keep pace with the constant increase of Internet traffic. Packet queuing, classification and monitoring tasks contribute to the hardware bottlenecks and most CPU resource consuming tasks. For this reason a new class of router components, dedicated to high speed network layer processing has emerged over the last year: the network processor.

Unfortunately, network processors are clearly designed in an opposite way as the Linux paradigm. Indeed, they are either:

- based on proprietary designs which are exclusive to a particular router manufacturer, or,
- designed for general purposes, but the cost of both acquisition and training remains extremely high.

Furthermore, in most cases, detailed specifications are not made available even following the signature of a NDA (Non Disclosure Agreement) and development platforms - both hardware and software - are not available at a reasonable cost.

3 IFT-BASED EXPERIMENTAL ROUTER

3.1 IFT: a high speed H/W classification and monitoring engine

Several years ago, FTR&D (formerly known as CNET) initiated a research program on high speed networking techniques - initially ATM - to address the issues discussed in section 2 of this document.

One way for performance improvement is system optimisation. Looking at a conventional router, one can see that less than 5% of the system software runs in the data path but is responsible for more than 95% of execution time. A reference document for the qualification of an IP router's requirements is given in [RFC-1812]. Only a small part of the related functions has to be "wired" to reach the performance level needed today, this level being around 1.5 10^6 packets/second per Gigabit/s bit rate at the interface level.

A relatively small set of wired basic functions must be able to process most of the traffic. Classification[2] is a critical functionality that should be as flexible as a pure software-based implementation to handle forwarding decisions. Other functions at this level include filtering such as Access Control Lists (ACL), an increasing set of encapsulations headers, and support for forthcoming protocols (IP v6).

A simplified architecture of a forwarding engine is depicted in figure 1. The functions to implement in hardware, assuming an underlying ATM link, are the AAL5 (ATM Adaptation Layer, type 5) related processes, L2 (Layer 2) de-capsulation, extensive header field analysis at various layers (address look-up and access list processing), corrupted datagram detection, counters, and header editing at output interfaces. Header editing means the update of IP header (TTL (Time To Live) and Checksum) and also pushing in front of the last analysed header (IP in most cases) all the logical output interface related data, that is to say MPLS labels, encapsulating headers and ATM VC (Virtual Circuit).

The software, which runs on a Linux workstation (although it was initially developed for Sun Solaris), is responsible for:

- Implementing the control path, in the same way as on a regular PC-based Linux router;
- Processing the “slow path” packets (those that have not been directly forwarded through the high speed hardwired path) through regular kernel processes;
- Mapping the Traffic Conditioner configuration commands and routing updates onto header pattern entries;
- Maintaining the pattern entries, header field analysis sequencing and counters allocation and retrieval through a dedicated driver.

[2] the process by which a data packet is examined and policy decision are made which affect down-stream processing [ROTHFUS]

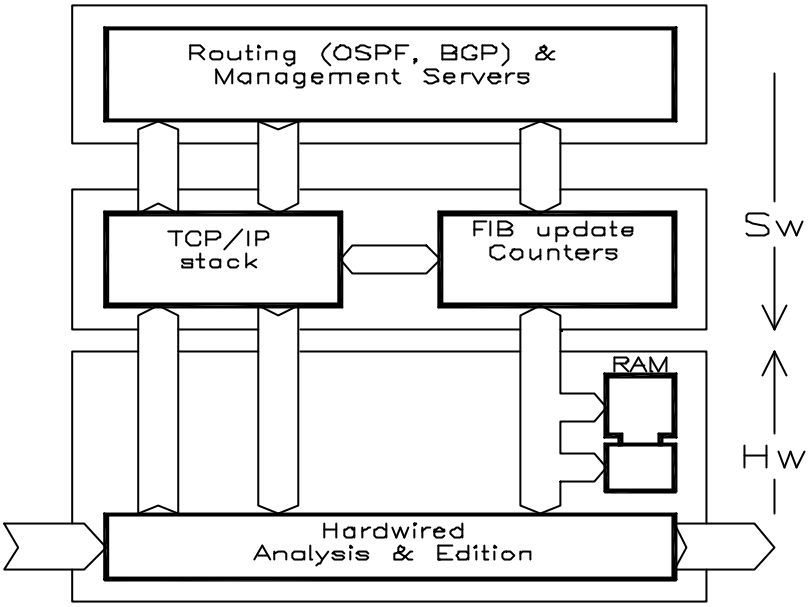

Figure 1. Architecture of the forwarding engine

The switching performance reaches the million of packets per second range, and the storage capacity allows an extensive set of access lists as well as the full Internet table entries to be processed in the same device.

In addition, the IFT functionalities have been tuned to the TEQUILA system requirements both for hardware and software purposes. The following operations have been added in the IFT hardware:

- Push, Pop and Swap operations on 2 levels of MPLS labels;
- MPLS EXP (EXPerimental) field tagging;
- TTL update (LSR (Label Switch Router)) and TTL exchanges between IP and MPLS layers (LER (Label Edge Router));
- Insertion of an LLC-SNAP (Logical Link Control – Sub-Network Access Point) header instead of 2 levels of MPLS headers at LER egress.

3.2 Building scaleable and high performance systems

The limitations of routers based upon shared bus systems may be overcome by hardware-based distributed forwarding mechanisms implemented directly at the level of interface boards with inter-board communications within the router being provided through a switching fabric. A variation in this architecture, well suited for systems that are not purely devoted to processing IP traffic but also have to inter-work with legacy devices, is the implementation of forwarding functions within server modules.

The combination of high speed forwarding engines with advanced header analysis capabilities, together with ATM switch fabric with state-of-the-art ATM transfer capabilities and the scalability inherent to switching networks allows high-end systems to be built. Note that most of these considerations also apply to a design, which is based upon an Ethernet switch. In addition to OC-12 ATM interfaces, OC-12 Packet Over SONET and Gigabit Ethernet interfaces are currently under development.

The communication within the IFT-based experimental router is performed through an ATM switch fabric (in the present design). Packets, once processed by the IFT, are directed to appropriate external or internal interfaces, which are handled by the Linux host and the control plane processes. Two router architecture models are currently under study: the server model and the distributed model (Figure2).

In the server model, all the traffic generated by accesses that needs to be processed at level 3 is directed to dedicated IFT modules that are capable of handling data at the IP layer. In this scheme, the amount of data to be processed at layer 3 must be consistent with the capacity of the overall IFT modules.

If the required capacity at layer 3 is large, a distributed model is much more suitable. In this case, the IP forwarding capacity is distributed on every interface. In both models, IP dynamic routing protocols (OSPF (Open Shortest Path First), BGP (Border Gateway Protocol)) run on a single Linux workstation. The routing process is in charge of computing IP routes and downloading the corresponding Forwarding Information Base "FIB" to each IFT module.

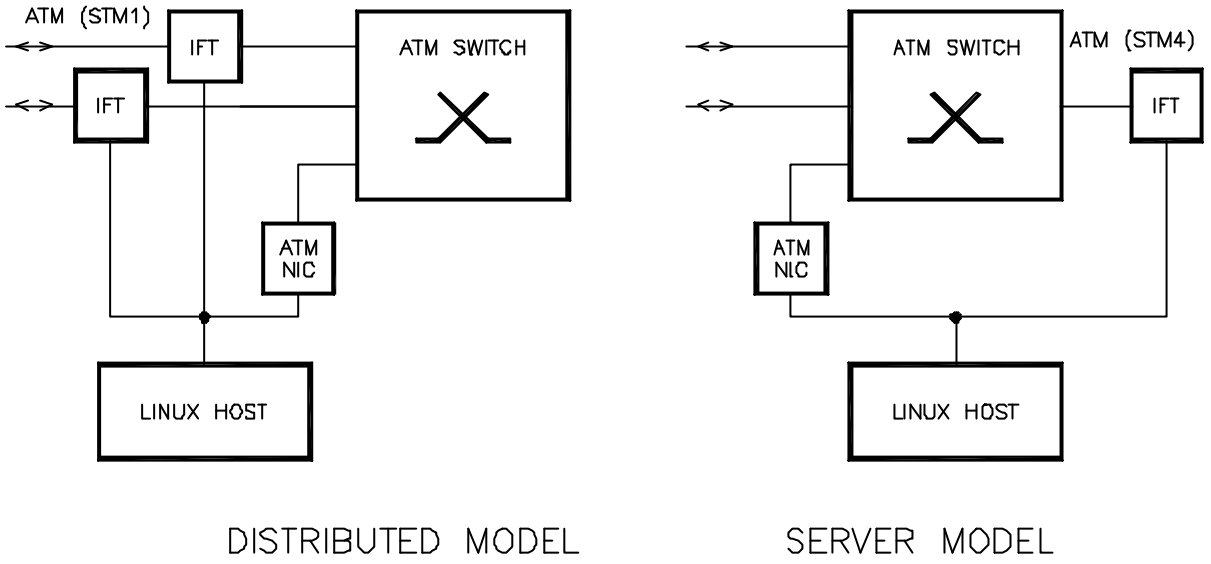

Figure 2. Server and distributed models Tequila Generic Adaptation Layer and the IFT

Within the context of the TEQUILA architecture, the system described above implements the following “Data Plane” functionalities:

- Traffic Conditioning,
- MPLS switching,
- PHB enforcement.

When operating as part of a network delivering QoS-based services, these functions need to be configured and monitored by management algorithms which reside outside of the IFT-based routers themselves. In the TEQUILA project, a set of management systems covering Service Management, Traffic Engineering and Monitoring functionalities have been defined [T4GJE]. A management interface is needed to support interactions with these external algorithms.

A GAL (Generic Adaptation Layer) component has been introduced to de-couple the TEQUILA algorithms from the specific underlying implementation of the data plane functionality of each kind of Network Element (NE) that is handled by the TEQUILA Project (Cisco, Linux-based and IFT-based).

The GAL component captures the common functionality of the different router types with respect to monitoring, traffic classification and conditioning, MPLS routing, scheduling and buffer management, and export it to the Tequila system through a generic interface which provides configuration and reporting capabilities.

The GAL component is further decomposed into a generic Interface Layer and a NE specific Interface Driver. The Interface Driver for the IFT-based router is implemented in its Linux host, according to the approach described in [DRLL-VVD].

4 CONCLUSION

The IFT-based Experimental Router as developed by the TEQUILA project provides several benefits at once when compared to the options of commercial routers, one the one hand, and experimental software-based routers deployed on Linux-based PCs, on the other hand:

- A performance level which is comparable to the switching performances of commercial routers;
- Scalability through the use of off-the-shelf switching fabric (currently ATM, to be ported to Fast and Gigabit Ethernet switches in the near future)
- The extensive developments that have been engaged on Linux-based routers.
- A clear separation between forwarding and control planes.

In addition to the role of the Experimental Router in the TEQUILA project, the IFT developments are also being considered for use in a commercial context for the implementation of a Multimedia Switch Router through an industrial partnership between FTR&D and an equipment vendor [ABDL]. The Multimedia Switch Router is a node that is capable of processing different types of data traffic such as voice, video or non-real time data over the same physical infrastructure. Security applications are also considered in collaboration within an academic partnership [PAUL].

5 REFERENCES

[ABDL] Michel Accarion, Christophe Boscher, Christian Duret, Joël Lattmann "Extensive Packet Header Lookup at Gb/s Speed for an Application to IP/ATM multimedia switch router" In World Telecommunication Congress – International Switching Symposium, Birmingham May 2000

[D1.2r2] P. Trimintzios *et al*, "Protocol and Algorithm Specification", Deliverable D1.2, IST TEQUILA Project, 2001.

[D2.1] D. Griffin *et al*, "Selection of Simulators, Network Elements and Development Environment and Specification of Enhancements", Deliverable D2.1r2, IST TEQUILA Project, 2000.

[DGO] Danny Goderis *et al*, "TEQUILA reformulated Objectives", November 2001

[DRLL-VVD]. C. Duret, F. Rischette, J. Lattmann, V. Laspreses, P. Van Heuven, S. Van den Berghe, P. Demesteer "High Router Flexibility and Performance by Combining Dedicated Lookup Hardware (IFT), off the Shelf Switches and Linux" submitted to IFIP Networking 2002

[PAUL] Olivier Paul, Maryline Laurent, Sylvain Gombault, "A Full Bandwidth ATM Firewall" in Proc. of the 6th European Symposium on Research in Computer Security, Toulouse, France, October2000

[RFC1812] F. Baker "Requirements for IP Version 4 Routers"

[ROTHFUS] "Programming & Reprogramming: Keeping the speed without Losing your Mind" in Network Processor Summit – Networld+Interop 2000

[T4GJE] P. Trimintzios *et al*, A Management and Control Architecture for Providing IP Differentiated Services in MPLS-based Networks in IEEE Communications Magazine May 2001

[TEQUILA] "Traffic Engineering for Quality of Service in the Internet, at Large Scale" Annex 1 - "Description of Work" Proposal number: IST-1999-11253 October 19, 1999