# ON THE DEPLOYMENT OF PIPELINE FORWARDING IN A NATION-WIDE INTERNET SERVICE PROVIDER NETWORK

*Mario Baldi, Riccardo Giacomelli, Guido Marchetto, Andrea Vesco*

Politecnico di Torino (Technical University of Turin)
Dipartimento di Automatica e Informatica (Department of Control and Computer Engineering)
+39 011 564 7067, {mario.baldi,riccardo.giacomelli,guido.marchetto,andrea.vesco}@polito.it

*Abstract*-**The paper studies the impact of the introduction of pipeline forwarding on the network of a large Internet Service Provider trying to provide a detailed analysis of the trade-offs between costs and benefits. The network of Telecom Italia is considered as a case study.**

## 1. INTRODUCTION

This work studies the implications of the introduction of pipeline forwarding (PF) of packets in the backbone network of a large Internet Service Provider (ISP), focusing on the main advantages and hurdles. The network of Telecom Italia (TI) is considered as a case study.

One of the most appealing features of PF is the capability of providing per-flow quality of service (QoS) without facing the scalability problem characterizing the IntServ model. Moreover the QoS in a PF network is independent of the bandwidth reserved to the single traffic source and the total network load. Each pipelined flow experiences (*i*) bounded deterministic end-to-end delay, (*ii*) bounded delay jitter, and (*iii*) no losses, because the PF eliminates communications link bottlenecks since it completely avoids congestion. Specifically, the end-to-end delay of pipelined packets is constant, independent of the total amount of pipelined traffic and can be calculated in advance, given the number of nodes traversed. These properties are ensured also when multicasting is implemented, as discussed in exiting literature.

Per-flow resource reservation is required to guarantee QoS to single flows. However, resource reservations can be aggregated in the network core so that per flow reservations are handled only at the edges [2]. In any case it is important to highlight that a PF capable router does not need to store per-flow information and to implement per-flow queuing. Therefore the approach scalability is granted.

Considering that QoS sensitive traffic might be the one for which customers are willing to pay higher rate, proper support for traffic with QoS requirements is extremely important as it is likely to provide a source of new revenue for service providers.

QoS provisioning is dependent on a successful scheduling operation at resource reservation time. (see Section 4.4 for a more detailed discussion about blocking). A non-null blocking probability can be experienced in a PF network before reaching full utilization. As a result, the total capacity allocated to pipelined traffic is not limited by the minimum level of QoS required, but by the maximum blocking probability the ISP will accept. As showed in previous work [3][4], in practical cases blocking probability is nearly negligible for network utilizations up to 80-90% of the total link capacity depending on the way PF is being deployed (see Section 2.1 for PF implementation flavors). The high link utilization allows an ISP to carry more traffic with QoS requirements on a PF network than over an asynchronous network based on DiffServ model. Consequently, given the growth trend of the traffic with QoS requirements, the deployment of PF can postpone the upgrade of the network infrastructure compared to the case in which other approaches based on network overdimensioning and resource overprovisioning are being deployed.

Finally, as it will be apparent from the PF working principles described in Section 2, PF provides segregation of traffic. Consequently, a PF network is inherently robust against security attacks, specifically denial of service (DoS) attacks. In fact, when a traffic class based solution is implemented, such as in the DiffServ model, the total amount of malicious traffic injected in a class can degrade the QoS provided to the entire traffic class. On a PF network only flows in which malicious traffic is generated can suffer a QoS degradation.

In Section 2 PF is introduced. Section 3 describes the TI backbone that is here considered as a case study for the deployment of PF, which is discussed in Section 4. Conclusions are drawn in Section 5.

## 2. PIPELINE FORWARDING

### 2.1. Operating Principles and Technologies

PF is a known optimal method that is widely used in computing and manufacturing. The necessary requirement for PF is having common time reference (CTR), which can possibly be derived from Universal Coordinated Time (UTC). In UTC-based PF all packet switches are synchronized, while utilizing a basic time period called time frame (TF). The TF duration ($T_f$) may be derived, for example, as a fraction of the UTC second received from a time-distribution system such as the global positioning system (GPS) and, in the near future, Galileo. TFs are grouped into time cycles (TCs) and TCs are further grouped into super cycles, each super cycle lasts for one UTC second. The TC provides the periodicity of the reserved flow. This results in a periodic schedule for IP packets to be switched and forwarded, which is repeated every TC.

The basic PF operation is regulated by two simple rules: (*i*) all packets that must be sent in TF $t$ by a node must be in its output ports buffers at the end of TF $t-1$, and (*ii*) a packet $p$ transmitted in TF $t$ by a node $n$ must be transmitted in TF $t+d_P$ by node $n+1$, where $d_P$ is an integer constant called *forwarding delay*, and TF $t$ and TF $t+d_P$ are also referred to as the *forwarding TF* of packet $p$ at node $n$ and node $n+1$, respectively. The value of the forwarding delay is determined at resource-reservation time and must be large enough to satisfy (*i*). In PF, a synchronous virtual pipe (SVP) is a predefined schedule reservation for forwarding a pre-allocated amount of bytes during one or more TFs along a path of subsequent PF switches. For example if *b* bytes are reserved every TC in each node along the path, the rate $R$ guaranteed is $R = b/T_C$ where $T_C$ is the time duration of the TC.

A hierarchical resource reservation model can be used to set-up SVPs, which enables multiple component SVPs to be aggregated in larger, possibly pre-provisioned, SVPs in the core of the network.

The forwarding delay may have different values for different nodes. Moreover, two variants of the basic PF operation are possible. When node *n* deploys immediate forwarding, the forwarding delay has the same value for all the packets transmitted by node *n*. When implementing non-immediate forwarding, node *n* may use different forwarding delays for packets belonging to different flows.

Two implementations of the PF were proposed: Time-Driven Switching (TDS) and Time-Driven Priority (TDP).

*Time-driven switching* (TDS) was proposed to realize sub-lambda or fractional lambda switching (FλS) in highly scalable dynamic optical networking [4], which requires minimum optical buffers. In this context, TDS has the same general objectives as optical burst switching and optical packet switching: realizing all-optical networks with high wavelength utilization. TFs can be viewed as virtual containers for multiple IP packets that are switched at every TDS switch based on and coordinated by the UTC signal.

In TDS all packets in the same TF are switched in the same way. Consequently, header processing is not required, which results in low complexity (hence high scalability) and enables optical implementation.

Due to their low complexity and high scalability, TDS switches are suitable for very high speed (possibly optical) backbone core where traffic can be organized in large capacity SVPs.

More flexibility could be required at the edge of the network as offered, for example, by conventional IP destination-address-based routing. *Time-driven priority* (TDP) [1] implements UTC-based pipeline forwarding combined with conventional IP routing: routing and packet forwarding may be based on either conventional IP destination-address-based routing (as mentioned above), or multi-protocol label switching (MPLS), or any other routing technology of choice.

### 2.2. Non-pipelined Traffic

Non-pipelined (i.e., non-scheduled) IP packets, i.e., packets that are not part of a SVP (e.g., IP best-effort packets), can be transmitted during any unused portion of a TF, whether it is not reserved or it is reserved but currently unused. Consequently, links can be fully utilized even if flows with reserved resources generate fewer packets than expected. A large part of Internet traffic today is generated by TCP-based elastic applications (e.g., file transfer, e-mail, WWW) that do not require a guaranteed service in term of end-to-end delay and jitter. Such traffic can be handled as non-pipelined and can benefit from statistical multiplexing. Moreover, any service discipline can be applied to packets being transmitted in unused TF portions. For example, various traffic classes could be implemented for non-pipelined packets in accordance to the “differentiated services model”.

## 3. AN ISP BACKBONE: A CASE STUDY

In order to assess the deployment impact of PF on an ISP backbone, the TI network was considered as a case study. The backbone network is composed of 32 points of presence (POPs) interconnected as shown in Fig. 1. There are three different types of POP: *Inner Core*, *Outer Core*, and *Secondary*, classified according to the total amount of traffic they handle. In each POP there are some *edge access routers* and two *backbone routers* connected to the backbone of capacity 10 Gb/s. As shown in Fig. 2, the *edge access routers* are *network access server* (NAS), *provider edge* (PE) and *media gateway*.

The edge access routers are interconnected by an overdimensioned Ethernet network with 1 Gb/s links, while the media gateway is connected directly to the backbone router as depicted in Fig. 2.

MPLS is deployed in the backbone network. Using the MPLS terminology, the PEs and the media gateway act as

Edge-LSR while the backbone routers act as LSR. In more details the backbone routers act as LSRs for the traffic coming from the PEs and from the media gateway and act as Edge-LSRs for the traffic coming from the NASs.

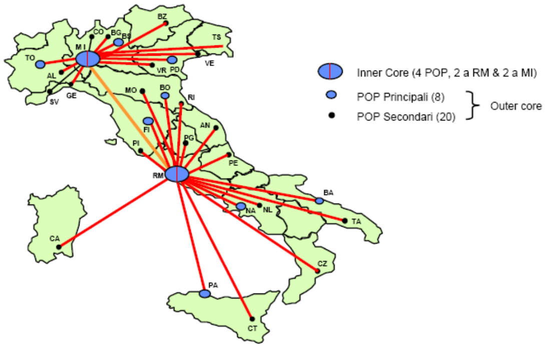

**Fig. 1 Telecom Italia Backbone Network**

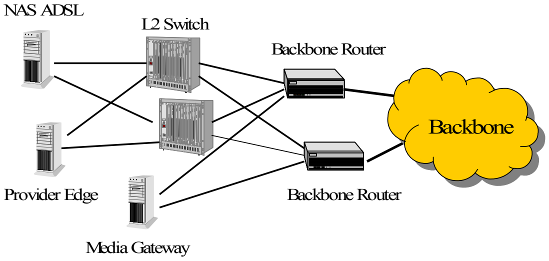

**Fig. 2 Architecture of a Point of Presence**

The NASs act as concentrator of the ADSL traffic of the residential users. The PEs act as concentrators of the VPN traffic coming from the customer edge routers (CEs) and the ADSL traffic of the business users. The media gateway acts as concentrator of the voice traffic coming from the PSTN network. This traffic is carried on the backbone as VoIP traffic.

### 3.1. Data plane architecture

Both IP and MPLS traffic are carried in the backbone network, thus the backbone routers implement common destination-based forwarding and label switching. The ADSL traffic coming from the NASs, with destination outside the TI network, is routed as IP traffic in the backbone, whereas the ADSL traffic, with destination inside the TI network, the VPN and the VoIP traffic are routed as MPLS traffic.

The VoIP traffic is handled with high priority in the backbone following the MPLS/DiffServ model implemented with traffic shaping at the network edge and strict priority FIFO scheduling in the backbone routers.

### 3.2. Control plane architecture

The routing architecture is divided in two levels. Inside the backbone each backbone router announces the address of its neighbors utilizing OSPF algorithm providing full connectivity among all the backbone routers.

The edge routers using BGP propagate the prefixes of the internal destinations and learn Internet destinations from edge routers of the other ISPs having peering agreements with TI. The edge routers maintain also iBGP sessions among them, in order to propagate internally the routes acquired with the BGP. An example of the routing architecture is depicted in Fig. 3 where routers A and D act as edge routers.

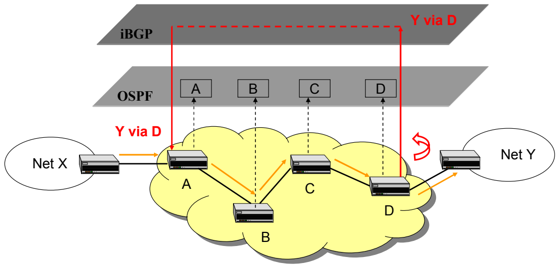

**Fig. 3 Routing architecture in the Telecom Italia Network**

The architecture deployed to support VPN services is compliant to the BGP/MPLS VPN solution.

Two different signaling protocols are deployed in network: (*i*) the Label Distribution Protocol (LDP) is deployed to build up the Label Switched Path (LSP) between PEs or backbone routers and (*ii*) the ReSource Reservation Protocol in its Traffic Engineering version (RSVP-TE) to build up the LSP between the media gateways carrying the VoIP traffic.

The RSVP-TE has been adopted to implement links protection policies providing circuit switching network services like fast reroute (i.g. less than 50 ms) of traffic in case of failure either of a link or of the routers linecard.

## 4. PIPELINE FORWARDING DEPLOYMENT

The next paragraphs describe a low complexity solution to introduce the PF in a complex network environment, fully exploiting its benefit in term of QoS without changing the operational model adopted by TI.

Introduction of the PF requires considering some aspects: (*i*) which asynchronous routers to be replaced with PF routers (*ii*) when to replace the asynchronous routers, (*iii*) which PF deployment options to be adopted in the edge and core network and (*iv*) which kinds of traffic must be handled with priority in accordance to the PF principles.

### 4.1. PF-based network architecture

The PF capable routers can be introduced gradually in the network, coexisting and interoperating with asynchronous packet routers and with asynchronous routers of other ISPs having peering agreements with TI.

A key parameter in discriminating and justifying the introduction of the PF is the number of nodes traversed by a generic path in the network because the benefits in term of QoS are higher when a high number of nodes are traversed.

The deployment of the PF in all the routers of TI backbone is suggested because of the few number of nodes traversed.

Fully exploiting the performance benefits of TDS requires a pre-emptive establishment and dimensioning of the SVPs based on the traffic characteristics. This approach would require changing the normal destination based paradigm of traffic management commonly adopted. Thus, in order to minimize the impact of the PF, TDP deployment option seems to be the best choice due to its high flexibility in traffic management.

Therefore in the resulting network architecture the edge access routers become *TDP edge routers* and the backbone routers become *TDP routers*.

The TDP edge routers, implementing the *SVP Interface,* are responsible of (*i*) classifying incoming traffic (*ii*) labeling the incoming traffic handled as MPLS traffic based on the association FEC-to-LABEL, (*iii*) policing and time-shaping the incoming traffic requiring the TDP services (e.g. the operation of time-shaping is based on the association FEC-to-SVP build up during the resource reservation phase) (*iv*) marking the non-pipelined traffic based on the DiffServ model and (*v*) policing and shaping the non-pipelined traffic.

The TDP backbone routers are responsible of PF packets belonging to the pipelined flows and forwarding the non-pipelined packets based on the DiffServ model.

### 4.2. Traffic Management

In the early deployment phase the entire traffic can be handled as non-pipelined without changing the TI operational model. Obviously the service experienced by the traffic flows is exactly the same they receive in the current DiffServ network. Thus, PF allows ISP to gradually plan which kind of traffic to be handled in a TDP basis. In the following we present a possible traffic management plan suitable for the initial TDP deployment phase based on the analysis of the kinds of traffic carried in the TI backbone and the service provided to them.

First of all, we propose to handle the QoS sensitive traffic in a TDP basis. Therefore, the VoIP traffic, VPN traffic and the multimedia broadcasting traffic are three main candidates.

In particular we envision the possibility of provisioning SVPs between the Voice Gateway and PE routers belonging to different POPs, in order to carry the VoIP and VPN traffic, and provisioning SVPs from the multimedia content servers toward the network access, in order to carry the multimedia content with high level of QoS as near as possible to the end users. The advantages of handling these kinds of traffic in a TDP basis will be described in the paragraph 4.6.

The ADSL traffic of the residential and business users can be handled as non-pipelined traffic in accordance to the DiffServ model, for example handling ADSL business traffic at higher priority than the residential traffic.

Moreover, it is suggested to handle the traffic generated by the routing and signaling protocols as non-pipelined because the SVPs dedicated to these kinds of traffic could be potentially underutilized due to ON-OFF nature of such a traffic. Obviously, the reserved bandwidth left empty is used to forward the non-pipelined traffic, but the reservation needed could potentially decrease the probability of finding schedules for other flow requiring TDP service.

However, such a traffic is considered critic for the network maintenance, thus it is suggested to handle it with the highest priority in accordance to the DiffServ model.

### 4.3. The Control Plane Architecture

The SVPs on backbones can be used to provision bandwidth reservation to single flows or to an aggregate of traffic. Setting up an SVP requires solving a distributed scheduling problem on the links on the route from source to destination [2]. The main scheduling and resource reservations data structure is the called *link availability vector* associated with each link of the network and containing the amount of bits that have not yet been reserved during each TF. When the scheduler processes an arriving resource reservation request, it builds up an availability vector containing the amount of bits that can be reserved on the whole path (*availability vector*) per each TF of the time cycle.

Resource allocation will be performed by selecting the TFs to be reserved based on the information carried within the availability vector. An extension of the protocols previously elaborated in the Integrated Services (IntServ) and MPLS context is needed to implement the above scheduling and resource reservation algorithm. The aforementioned protocols are Label Distributed Protocol (LDP), Resource reSerVation Protocol (RSVP) and traffic engineering variant (RSVP-TE).

The paradigm of building up a SVP is similar to the one of building up a Label Switched Path (LSP) in a MPLS cloud. The difference resides in the specification of the required bandwidth (e.g. an LSP is a path from $x$ to $y$ while a SVP is a path from $x$ to $y$ with a guaranteed bandwidth $B$).

Therefore, the RSVP protocol or the TE variant are well suited to implement the steps required to find a schedule through a PF network as explained above. However, as well explained in [2] the PF does not share scalability issues of the IntServ reservation model, proved not scalable, because it is simpler since resource information consists in arrays of counters (or bits in TDS) accounting for reserved or available capacity during TFs and processing reservation requests involves simple addition and subtraction (or logical OR operations in TDS).

### 4.4. Modifications to the existing operational model

In the following we discuss the variation to the TI operational model during the PF deployment.

During time-based scheduling operations a PF capable router starts serving the priority queue storing the TDP traffic at the beginning of each TF. The TDP scheduling procedure is completely transparent to the forwarding operations. The PF backbone implementing the common destination-based forwarding and the label switching is able to carry both IP and MPLS traffic. Both forwarding paradigm could be implemented for pipeline and not pipelined traffic. Thus the provisioning service model adopted does not change; allowing TI to provide scalable IP services with different priority.

In a PF network, routing paths must not change for the entire connection duration. Most of large IP backbones nowadays deploy connection-oriented technologies like MPLS to implement traffic engineering and fast fault protection. Consequently, PF can be introduced without requiring major changes in the ISPs network operation model.

The routing model, the traffic engineering policies and the traffic protection mechanism previously adopted can be maintained in the PF architecture. The forwarding and resource reservation operations of the PF are completely transparent to the high level protocols; consequently any routing and TE models previously defined could be adopted without changing the TI operational model.

The VPN/MPLS model adopted to provide VPN services is heavily based on routing protocols and on extended addressing plan. Thus also the provisioning service model and managing model of the VPN services do not change.

Moreover, the ISP is able to build up *explicit SVP* and *Hop-by-Hop SVP* both on-demand or statically exactly as done in a MPLS based network. Note that the creation time of a SVP and a LSP are comparable. In case of explicit SVP the signaling protocol, modified to operate in a PF capable network will drive the route selection and reservation operations, while in case of Hop-by-Hop SVPs the route selection will be driven by the routing protocols. The ISP must specify the source and destination nodes and the bandwidth required by the SVP in both cases, then the route selection and resource reservation process will be completely transparent to the ISP and the SVP will be established over any network route with enough capacity available.

The changes to the signaling protocol needed to operate in a PF capable network, are transparent to the ISP and not involve the way the ISP manages and operates in the network.

In conclusion two differences in the operational model must be remarked: (*i*) dimensioning and configuration of the CTR and (*ii*) specification of the required bandwidth before opening a SVP in the network.

### 4.5. Limitations

In a PF network in order to provide deterministic QoS, resources must be reserved in the form of transmission capacity during specific TFs. Reserving resources for a call requires solving a *scheduling* problem to find a feasible sequence of TFs, called *schedule*, on links on the route from source to destination. When a new call is being started and resources are being looked for, the reservation can be denied even though enough capacity is available on all the links on that call's path. This happens if the identity of the TFs on the various links does not match the timing resulting from PF shaping, thus not satisfying the requirements imposed by PF forwarding. The session is said to be *unschedulable*.

Unschedulability does not exist on asynchronous packet networks because resource reservation is based on various heuristic procedures that are called *admission control*. Asynchronous admission control maintains link and network utilization well below 100%, otherwise it is impossible to guarantee the QoS. Unschedulability in PF networks can be compared to *blocking* in digital circuit switching, where it results from a path from an idle inlet to an idle outlet not being available. Thus, PF *blocking probability* is defined as the probability for a resource reservation on a PF network to be denied because of unschedulability. The blocking probability depends on many parameters, such as the size of the packet and the number of bits that can be sent during a TF. Moreover, the choice of the schedule when there is more than one possibility affects the utilization achievable on the network, and thus, the blocking probability.

The PF blocking probability can be reduced by using *non-immediate* forwarding since the TFs to be reserved by each node on its outgoing link are not uniquely determined by the TFs reserved on the incoming link; it can be chosen from a set of *D* TFs. Note that when *D* is equal to the time cycle size *k*, scheduling is always possible when resources are available, i.e., there is no blocking and full link utilization is possible. The blocking probability has been extensively studied in the literature on both TDP [3] and TDS [4] networks. The main results are: (*i*) the efficiency of PF increases - because blocking probability decreases -

as the link bandwidth increases. This is indeed a strong *scalability* property of PF, (*ii*)fixed the links capacity, higher the number of TF in a TC lower is the blocking probability, (*iii*) packet size affects the blocking probability because of the *fragmentation* of the capacity due to PF. However this phenomenon is negligible on high capacity links, and finally (*iv*) the PF shows higher efficiency with balanced scheduling than with unbalanced scheduling.

In conclusion the Call blocking does not compromise the efficiency of the network because, under immediate forwarding that provides more limited scheduling choices than non immediate forwarding, the link utilization is above 90% if TDP is deployed and above 80% if TDS is deployed.

### 4.6. Expected Benefits

The main advantages of deploying a PF capable backbone are related to the capability of providing services with deterministic QoS without overdimensioning the network infrastructure to meet the QoS requirements.

In details, a PF network allows an ISP to provide voice services and multimedia broadcasting services with high quality and deterministic QoS, in terms of end-to-end delay and loss, while achieving high link utilization. This because the PF completely avoids interactions among different SVPs carrying different kinds of traffic also at the bandwidth mismatch point. Moreover, also the VPN services will benefit from the PF because they could be provided with deterministic bandwidth guarantees. In general a PF network allows an ISP to provide scalable CDN-like connectivity services on a packet network infrastructure.

Besides the advantages on network services already provided a PF network allows TI to provide new valuable services with high QoS to the end users.

Although PF SVPs can be used to provision bandwidth on backbones, they enable bandwidth provisioning on-demand to the single user applications. In order to fully benefit from PF a fraction of a SVP must be reserved to single applications before they start generating packets. Since it seems not realistic changing the current applications in order to support the resource reservation procedure, in a possible deployment, as described in [2], an access bandwidth broker (ABB) at the edges of an SVP can handle signaling procedure on behalf of the applications resident on the end system.

The ability of a PF capable network to provide guaranteed bandwidth in a scalable fashion will allow provisioning of new valuable services with guaranteed QoS to the end users. We envision the following services: Video on-demand (VoD), Streaming Peer-to-Peer, Distributed interactive Gaming, Videoconference and Telepresence services. Nowadays, high quality VoD services and experimental 3D-HD streaming which allows users to explore and interact with 3-dimensional environments are offered in U.S.A. and in Germany. VoD services require high bandwidth ranging from 2 to 15 Mbps while 3D-HD streaming services require even more bandwidth up to hundreds Mbps per flow.

Moreover audio and video conferencing solutions based on a peer-to-peer overlay to forward media streams, such as Skype, are gaining significant importance and diffusion as they are capable of working around firewalls and network address translation (NAT) functionalities. In general, when a peer-to-peer overlay is deployed, resource reservation might follow or accompany the procedures that lead to establishing peering between nodes. PF might be particularly beneficial when the peer-to-peer paradigm is exploited. In fact, media might travel along a sub-optimal path through a number of relay nodes. Given the delay introduced by relay nodes and the non-minimal length of the path, end-to-end delay becomes an even more critical issue and deploying PF might be the only viable solution in certain operating conditions to keep it within acceptable bounds.

## 5. CONCLUSIONS

Pipeline forwarding can be gradually deployed in a ISP backbone without modifying the operational model and without loosing traffic management flexibility. Furthermore deterministic end-to-end QoS can be granted per flow and per aggregated of traffic. Even utilizing resource reservation approach link utilization is very high due to bandwidth reuse for best-effort traffic. Adopting pipeline forwarding an ISP can provide next generation multimedia services needing high bandwidth resources in a simple and low cost manner.

## 6. REFERENCES

[1] C-S. Li, Y. Ofek, A. Segall, k. Sohraby, “Pseudo-Isochronous Cell Switching in ATM Networks,” *Computer Networks and ISDN systems*, no.30, 1998.

[2] M. Baldi, G. Marchetto, Y. Ofek, "A Scalable Solution for Engineering Streaming Traffic in the Future Internet," *to appear in Computer Networks (COMNET),* 2007.

[3] M. Baldi, Y. Ofek, "Blocking Probability with Time-driven Priority Scheduling," *SCS Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2000),* Vancouver, BC, Canada, July 2000.

[4] M. Baldi, Y. Ofek, "Fractional Lambda Switching - Principles of Operation and Performance Issues," *SIMULATION: Transactions of The Society for Modeling and Simulation International*, Vol. 80, No. 10, Oct. 2004, pp. 527-544.