Ultra Scalability Switching without "Stopping" the Serial Bit Stream

Mario Baldi[•], Renato Lo Cigno[◆], Yoram Ofek[◆], Achille Pattavina[■]

[•] Computer Engineering Department, Politecnico di Torino
Corso Duca degli Abruzzi, 24, 10129 Torino, Italy
[◆]Department of Information and Communication Technology – DIT
University of Trento, Via Sommarive 14, I-38050 Povo, Trento, Italy
[■] Department of Eletronics and Information, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italy

***Abstract* — For achieving ultra scalable switching of IP packets it is essential to minimize "stopping" of the serial bit streams. In our recent experimental work we demonstrated how such switch can be realized with and ultra-scalable switching architecture reaching multi-terabits per second switching in a single chassis. Additionally, we are experimenting end-to-end support of bufferless bit-streaming including a wireless access network enabling cheap large-bandwidth, low latency services.**

***Index Terms*—Optical Network, Terabit Switch, UTC-based Pipeline Forwarding, Sub-lambda Switching, Optical-Wireless Interconnection.**

I. Introduction

The Internet has been growing steadily in the past few years. One likely scenario is that the future Internet will be dominated by applications such as HDTV (perhaps in 3D), video on demand, high quality videoconferencing, distributed gaming, (3D) virtual reality, and many more. These applications are likely to generate traffic that either is by nature a streaming one or can be efficiently mapped into and handled as such (e.g., very large file transfer).

In order to achieve ultra scalable switching of IP packets for the above applications, it is essential: (1) to minimize "stopping" the serial bit stream, and (2) to minimize buffering requirements. In our recent experimental work we demonstrated that converting to electronics (without "stopping" the bits) in order to utilize off-the-shelf cross point switches it is possible to construct the most scalable switching system. We have demonstrated that today sub-lambda switching is best performed in electronics, while interconnection and transmission are best performed in optics. The experimental switching solution, which can scale to tens of Tb/s and beyond in a single chassis, utilizes global time (i.e., UTC – coordinated universal time) and pipeline forwarding, and ensures continuous serial bit stream flow with minimum buffers.

This paper discusses a method known as *pipeline forwarding* (see, for example, [1][2][3][4]) that is particularly suitable to carry various streaming media applications over the Internet since it offers:

1. High scalability of network switches (multi-terabits per second in a single chassis, which is this paper focus,
2. Quality of service guarantees (deterministic delay and no loss) for (UDP-based) constant bit rate (CBR) and variable bit rate (VBR) streaming — as needed, while
3. Preserving the support of elastic, TCP-based traffic, i.e., existing applications based on "best-effort" services are not affected in any way.

In order to put this work in a wider perspective, note that the per-chassis switching capacity of Cisco's top-of-the-line router with a novel network processor design, CRS-1, is only 640 Gb/s [the 92 Tb/s announced switching capacity should be divided by 2 (as the announcement refers to the sum of the overall input and output capacity) and then by 72 chassis's]. This capacity represents a factor of 2 improvement over the Cisco 12000 after 5 years of development with a 500 million dollars investment. So, if Internet traffic is instead doubling, say, every 18 months there is a real switching bottleneck on the horizon.

Section II focuses on the scalable switch design using off-the-shelf components. The architecture and the implementation of the switching system (fabric and switch controller) are presented in Section III, while Section IV describes the prototype and some testing results, most significantly obtained with six nodes and 100 km of single mode fiber (four 25 km fiber segments).

II. Scalable Design with Off-the-shelf Components

A. Why it is Scalable

On the one hand existing asynchronous IP switching seems to be limited to about one Tb/s in a single chassis due to memory access scalability. On the other hand, scalability of all-optical switching was successful only for whole lambda (optical channel) switching. This implies capacity provisioning of the entire (whole) optical channel capacity or nothing. Thus, in order to maximize scalability the following three design principles were used (the motivations and justifications will be presented in the following sections):

1. High-speed electronic switching fabric, with
2. Optical interconnection, and with
3. Global pipeline forwarding (PF) with time-driven switching (TDS) and control.

Today, *single-chip, low-cost, high-capacity* electronic cross-connects are available on the market: for example, a 144-by-144 switch matrix with up to 11 Gb/s per input/output port with more that 1.5 Tb/s aggregate switching capacity costs just a few hundreds dollars. However, constructing a large switching matrix based on such cross-connects requires short-range/high-density optical interconnects. Optical interconnects allow, at least in principle, any desired

interconnection topology, while minimizing noise and interference sources. Finally, by using global time ([1][2][3][4]) it is possible to minimize the following switching complexity components:

1. Input buffer size (while eliminating the need for output buffers all together) —which impacts the space domain complexity,
2. Number of switching elements in the electronic switching fabric —which impacts on complexity in the space domain— and
3. Number of operations required by the electronic switch controller when continuously configuring switching fabric input ports to output ports permutations —which impacts on complexity in the time domain.

To summarize the above discussion, our design was guided by understanding the clear limitations of both "all-optical" and "all-electrical" switching approaches. Consequently, our current conclusion is that a hybrid or "best-of-breed" switching solution is needed in order to achieve 10-100 Tb/s switching capacity in a single chassis. Specifically, our design approach is based on electronic switching and (very limited) buffering with optical interconnects.

B. Optimal switch design

Figure 1 shows the switch model used in this work with the following optimized components:

1. Input port with optimal memory access speedup of 1 (s=1), where speedup is defined as the ratio between the link bandwidth and the memory access bandwidth;
2. Input port with a single queue/buffer (typical input buffer switches, without any quality of service support, have N queues – one queue per output in order to prevent head-of-the-line blocking);
3. Optimal output port design without buffers, i.e., data packets are forwarded from the switching fabric directly onto the out-going link;
4. Minimal switching fabric complexity by using a multistage Banyan interconnection topology, with switching complexity of O(a*N*lg_aN), where N is the number of input/output and a is the size of each switching element or block;
5. Switch control complexity which is equal to the fabric complexity, O(a*N*lg_aN), i.e., the number of switching elements that the switch controller has to configure.

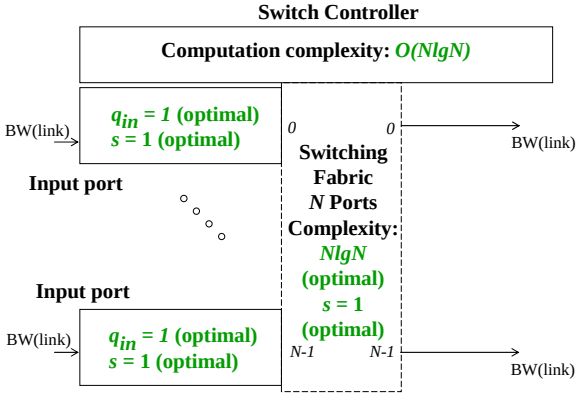


Figure 1. Optimal switch design

C. Optimality and pipeline forwarding

Pipeline forwarding is a known optimal method that is extensively used in manufacturing (e.g., automotive assembly line) and computer architecture (e.g., reduced instruction set computer). In this work pipeline forwarding is used as the basic operating principle for time-driven switching (TDS) (also known as sub-lambda or fractional lambda switching, FλS). As in other pipeline forwarding implementations, the necessary operating requirement for realizing TDS is having a *common time reference* (CTR). In the context of a global network the CTR is effectively realized by using UTC (coordinated universal time) that is globally available via GPS (or Galileo in the near future).

D. Optimal switching scalability

Figure 2 and Figure 3 are two possible Banyan-based switching fabric configurations. The first configuration, in Figure 2, is based on Mindspeed M21151/6 cross-connects that was used in our current prototype, as described in the following sections. The Banyan design using M21151/6 has aggregate switching capacity of 10 Tb/s. The switch design in Figure 3 is based on a state-of-the-art cross-connect, VSC3040, made by Vitesse. This cross-connect has 144 serial input/output ports operating at up to 11 Gb/s (the maximum rate of M21151/6 ports is only 3.2 Gb/s). A Banyan-based switching fabric with aggregate switching capacity of 160 Tb/s (128*128*10 Gb/s) can be built using Vitesse's cross-point switch, as shown in Figure 3.

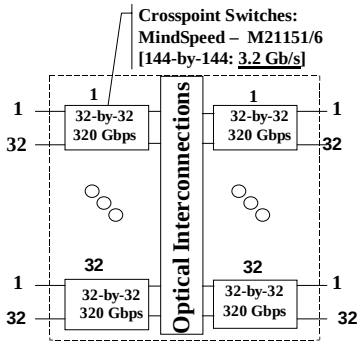


Figure 2. 10 Tb/s Switching with M21151/6 cross-point

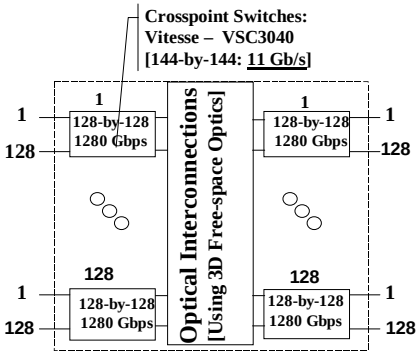


Figure 3. 160 Tb/s Switching Module (for future testbeds) based on off-the-shelf VSC3040 cross-point switches

III. TESTBED AND TESTING

The current testbed implemented in the Electronics Lab at the University of Trento is shown in Figure 4. The major components are the streaming media sources, a pipeline forwarding router for time shaping (i.e., scheduling) packet asynchronously generated by the sources, and a 25 km single mode optical fiber connecting two TDS switching nodes.

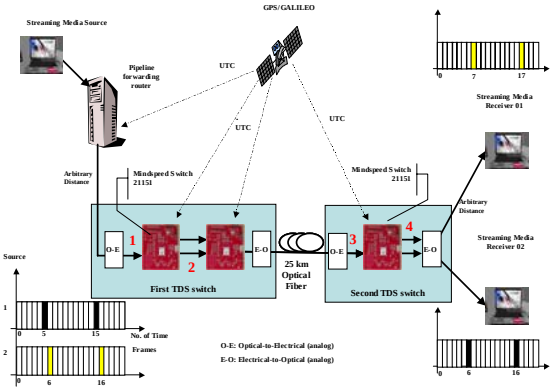
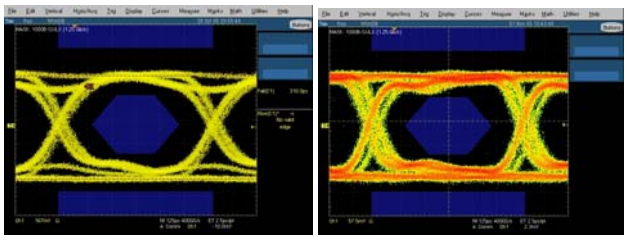


Figure 4: Initial testbed with two TDS nodes

The eye pattern test is quick method for visually examining the quality of serial signals, Figure 5 shows actual eye diagrams as captured at test points 1 and 4 in Figure 4.



(i) (ii)

Figure 5: Eye Pattern at points: (i) 1 and (ii) 4 of Figure 4

To further evaluate the robustness of our initial testbed setup and in order to know its limits the testbed was extended in two manners: (1) using four segments of 25 km fiber – total of 100 km single mode fiber and (2) using nine stages of cross-point switches, as shown in Figure 6. The nine stages of cross-point switches constitute six TDS nodes.

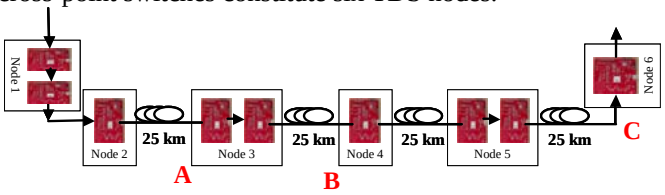
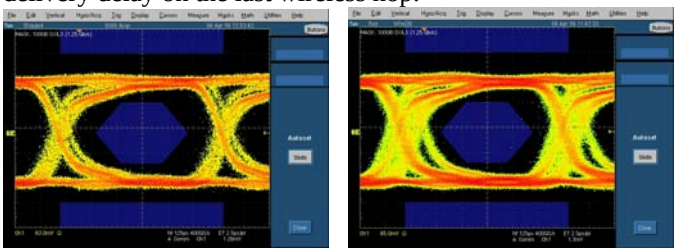


Figure 6*: Metro network topology: 4*25=100 km fiber and 9 stages of cross-connects (or six TDS nodes)*

Standard eye pattern test was performed in various location of this advanced testbed configuration. Figure 7 shows the eye diagram taken at locations B and C of Figure 6.

The final goal of communications is always providing high quality services to end-users. Many access networks are based on wireless LANs. UTC synchronization of clients and Access Points (AP) is not convenient for both technical (e.g., problems in receiving GPS indoor) and economical reasons. Yet, providing a smooth, bufferless, end-to-end datapath as provided by TDS is extremely appealing for real-time applications. Through proper synchronization protocols between the backbone switching domain and the APs of the wireless, we plan to experiment the possibility of optimizing wireless resources management in order to minimize the delivery delay on the last wireless hop.



(i) *(ii)*

Figure 7: Eye Pattern at (i) location B and (ii) location C

The key idea is synchronizing the beacon transmission and 'super-framing' of 802.11e. Naturally, once the AP is synchronized, the Hybrid Coordination Function (HCF) hooks the terminals to the same synchronization and allows the proper management of resources avoiding channel conflicts and contentions.

IV. CONCLUSION

Implementing UTC-based pipeline forwarding in a real testbed that is scalable to multi-terabit/s switching capacity has been a rewarding experience. The implementation success is a direct outcome of the simplicity of the pipeline forwarding method. The beauty is that the simplicity of this realization did not compromise two most desired performance properties for the future Internet: (1) switching scalability to 10 and 160 Tb/s and (2) predictable QoS performance for streaming media and large (content, Grid computing) file transfers.

Furthermore, although the presented prototype is based on optically interconnected electronic switches, the serial bit streams are never "stopped"; i.e., the transmitted bits are neither digitized nor digitally stored. In recent experiments it was possible to transmit the serial bit streams through six nodes and 100 km of fiber without "stopping" them. This is significant as it demonstrates that the deployed technology is suitable for providing ultra-scalable switching capacity in metropolitan area networks with minimum cost. Recent work includes initial studies for extending UTC synchronization over a WLAN access networks.

REFERENCES

[1] M. Baldi and 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.

[2] D. Grieco, A. Pattavina and Y. Ofek, "Fractional Lambda Switching for Flexible Bandwidth Provisioning in WDM Networks: Principles and Performance", *Photonic Network Communications*, Issue: Volume 9, Number 3, Date: May 2005, Pages: 281 – 296.

[3] M. Baldi, Y. Ofek, "Fractional lambda switching," *Proc. of ICC 2002*, New York, USA, Vol.5, pp 2692-2696.

[4] A. Pattavina, M. Bonomi, Y. Ofek, "Performance evaluation of time driven switching for flexible bandwidth provisioning in WDM networks," *Proc. of Globecom 2004*, Texas, USA, Vol. 3, pp1930-1935.