Fast and scalable optical packet switch architecture for computer communication networks

S. Di Lucente[1], Y. Nazarathy[1,2], O. Raz[1], N. Calabretta[1] and H.J.S. Dorren[1]

[1] Eindhoven University of Technology, Dept. of Electrical Engineering, P.O. Box 513, 5600 MB Eindhoven, the Netherlands
[2] EURANDOM, Eindhoven University of Technology

We present a novel low latency, high throughput and scalable optical packet switch (OPS) capable to optically interconnect hundreds of input/output ports. We focus on a strictly non-blocking Spanke architecture with contention resolution based on wavelength conversion. Highly distributed control of the OPS reduces the switching time to few nanoseconds regardless the amount of inputs/outputs. Queuing node analysis (mean values analysis) of input buffers in a computer communication network with window flow control confirms that the new architecture, unlike rearrangeable non-blocking (i.e. Benes) architecture, can operate with low latency and high throughput with a very large amount of input/output ports.

Introduction

The next generation High Performance Computer Systems and data-centers require PetaFlop/s processing speed and Petabyte storage capacity [1] with thousands of low-latency short link interconnections between computers nodes [2]. Switch matrices that operate transparently in the optical domain are a potential way to efficiently interconnect 1000's of inputs/outputs, complying the harsh transmission quality requirements (BER<10^{-21} after retransmission and FEC) and the end-to-end latency (~ 1 μs including encoding/decoding and processing time) of these systems [3].
We consider the general system depicted in Fig. 1a. It consists of an electronically controlled optical switch matrix interconnecting a large number of ingress and egress nodes. Typical switch architectures that are used in such system are rearrangeable non-blocking architectures (Benes, Banyan, Omega, etc…). Figure 1b shows the time it takes to reconfigure such switch matrices expressed in clock-cycles for several routing algorithms [4]. Assuming that a clock cycle takes one ns, it is visible that the fastest algorithm is the looping algorithm (which scales as $N\log_2(N)$, N being the numbers of nodes), it follows that latency requirements cannot be met. During this time, the switch is not able to handle data, thus incoming packets are either lost or need to be stored in electronic buffers, limiting the maximum load of the system. A part from that, the switches matrices discussed above allow data packets to be simultaneously routed to the same output port. This leads to packet collisions that degrade the system performance.
In this paper we investigate the load and the latency for systems as shown in Fig. 1a. We firstly compute the load and latency for a Benes architecture and show that there is little potential to scale such architecture to a large number of nodes. We present an alternative architecture that has distributed control. The architecture we present has as an advantage that the control time is independent of the number of nodes and therefore the architecture has an advantage that latency can be very small while operated at high number of nodes.

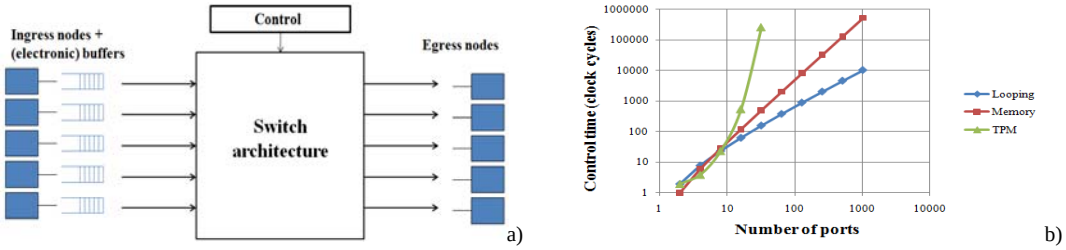

Fig. 1: (a) abstract system under investigation and (b) matrix reconfiguration time according to different algorithms.

Switch architecture with distributed control and contention resolution

Fig. 2 shows the OPS architecture with distributed control to interconnect the input/output nodes. Window flow control is commonly used to govern the data transmission. In a system with window flow control copies of the transmitted packets are stored in electronic input buffers until they are acknowledged. New traffic is refused when a certain number of packets in the node is unacknowledged, the upper bound of unacknowledged packets is called Window Size (WS). Thus we deal with packet contention by simply retransmitting of data.

In order to minimize the latency and to maximize the load of the system, queues at the input buffers should be as small as possible. In such system queues at the electronic buffers are related to three factors: the Round Trip Time(RTT), the contention probability and the reconfiguration time of the switch. While RTT is determined by the physical link, contention probability and reconfiguration time depend on the architecture and on the amount of input/output ports of the switch. Considering a short link (50 m), in long links the latency associated with the RTT is simply not acceptable, and rearrangeable non-blocking architectures, like Benes, the reconfiguration time of the switch is the predominant term and limits the scalability.

The proposed new switch is a strictly non-blocking architecture based on parallel 1xN optical switches and Contention Resolution Blocks (CBRs) as shown in Fig. 2. This architecture is self-configured. Each 1xN switch can manage the forwarding of the packets with a control that is independent of the other 1xN switches. This results in a highly parallel and distributed control. The only latency introduced to reconfigure the switching matrix is the time to process the packet labels.

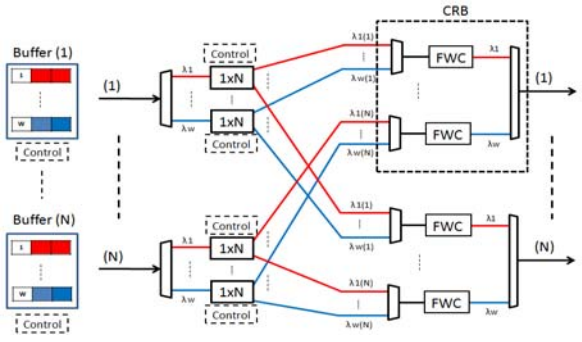

Fig. 2: New OPS architecture with highly distributed control and contention resolution.

The system in Fig. 2 operates as follows. Each of the N input buffers consists of W queues. Packets can leave the input buffers using any of the W available wavelength channels. Up to W packets can leave each input buffer in a single time slot if they are addressed to different egress nodes; in case that two or more packets have the same

destination, one or more packets need to be stored and forwarded in next time slots. The W signals, indicated by λ_1 to λ_W, are then multiplexed and represent one of the N input of the switch. Thus, the total number of logical inputs/outputs is NW.
At each switch input, packets at different wavelengths are demultiplexed and fed into 1xN optical switches. The packets consists of the payload and in-band labels that identify the destination. Fig. 3a shows the packets with the labelling scheme in both time (left) and spectral (right) domains [5]. The 1xN switch with fast label processor is depicted in Fig. 3b. Such switch has been implemented in [5]. Labels can be extracted by using a passive filter (demux in Fig. 3b). Key is the fact that the extracted parallel labels, once opto-electrically converted, can be parallel processed by the control logic with very low latency (< 3 ns) regardless the number of input/output ports. The outputs of the control logic drive the 1xN space switch to forward the packet to the right destination. An integrated electro-optical 1xN switch that can support the architecture is described in [6].

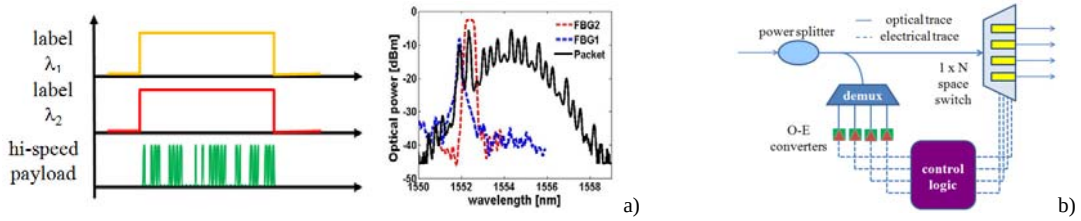

Fig. 3: (a) packets with two encoded labels in time (left) and spectral (right) domains; (b) 1xN switch with parallel labels processor.

Packets at the output of 1xN switches are fed in the CRB. The purpose of this block is to perform contention resolution by using the wavelength domain, employing Fixed Wavelength Converters (FWCs). Packets coming from different input fibers are converted into different wavelengths by the FWCs in order to solve contentions. The CRB is then always able to solve contentions between different input fibers.

Performance of the new architecture and comparison with Benes

We investigated the performance of the new architecture by queueing node analysis of input buffers in a 40 Gb/s link. In the numerical analysis packet duration was 1500 bits (37.5 ns, 40 ns including 2.5 ns guard band between packets). The link length was assumed to be 50 m. We followed the mean values analysis described in [7]. Throughput and latency can be computed recursively with the starting condition that there are no packets in the buffers and increasing the number of packets in the system from 0 to WS.
We compared NxN Benes and new NWxNW architectures (N, NW=2, 4, …, 1024). It is important to say that the new OPS is better suited for WDM applications, anyway we considered also the case with W=1 to highlight the differences in performance with the Benes architecture. From the analysis is clear that in Benes architecture the delay associated with the configuration time of the switch limits the throughput of the switch and it is the dominant factor for the latency of the system. Thanks to the highly distributed control this factor is negligible in the new architecture and it is reduced to few nanoseconds (3 ns in our model) regardless the number of input/output ports. In Fig. 4 (a, b) are depicted the results of the queueing node analysis, it's evident that only the new architecture can operate with a large amount of inputs/outputs with high throughput and low latency (its behavior is independent of ports count and almost constant).

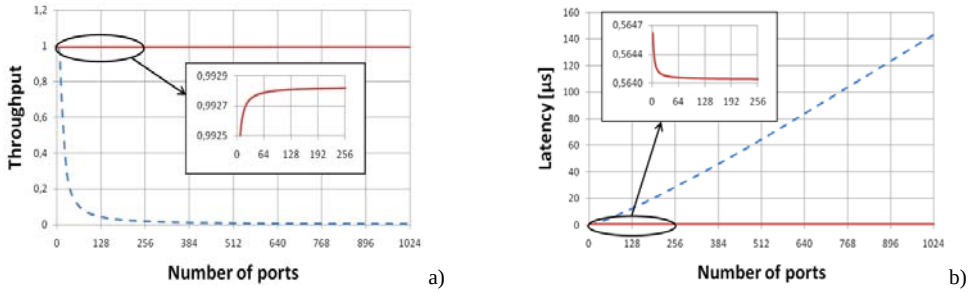

Fig. 4: (a) throughput and (b) latency of Benes (dashed lines) and architecture with distributed control (N=1024, W=1) (solidlines).

In Tab. 1 are reported the throughput and the latency of a 1024x1024 (NWxNW) OPS for different values of the number of input fibers N and of the number of channels per fiber W. In the configuration with N=16 and W=64, the probability of contention is much higher than in the other cases considered, thus the delay related to it becomes the dominant factor for the latency of the switch and limits the throughput of the system.

1024x1024	Throughput	Latency [µs]
N=1024, W=1	0,9928	0,5645
N=32, W=32	0,9933	0,5637
N=16, W=64	0,0928	6,0351

Tab. 1: Throughput and latency of a 1024x1024 OPS based on the architecture with distributed control in three configurations.

Conclusion

We presented a novel OPS architecture with highly distributed control. The advantage of this strictly non blocking architecture is that the control time of the switch matrix does not depend on the number of input/output ports. This leads to great scalability in terms of port count of the architecture.
Mean values analysis of queues at ingress buffers confirms that the distributed control allows the architecture to meet the latency requirements of computer communication networks (< 1 µs) also operated at large number of ingress and egress nodes (> 1000). Instead classical rearrangeable switch architectures, as Benes, with centralized control already cannot meet these latency requirements when the number of ports exceeds 32.
The latency and the throughput of the architecture are strictly related to the ratio N/W. In computer communication networks scenario, where the number of nodes N is close to the number of wavelengths per node W, the throughput of the architecture with distributed control is close to 100%.

References

[1] G. Bell et al., “Petascale computer system”, Computer 39, 110-112 (2006).

[2] G. Gilder, “The rise of exaflood optics”, ECOC 2009, Vienna, Austria, Plenary Talk (2009).

[3] R. Hemenway et al., “Optical-packet-switched interconnect for supercomputer application”, Journ. of Optical Networking, 3, 900-913 (2004).

[4] D.C. Opferman and N.T. Tsao-Wu, “On a class of rearrangeable switching networks”, Bell Syst. Tech J, 50, 1579-1618 (1971).

[5] N. Calabretta et al., “Scalable optical packet switches for multiple data formats and data-rates packets”, IEEE PTL 22, 483-485 (2010).

[6] N. Calabretta et al., “1x16 optical packet switch sub-system with a monolithically integrated InP optical switch ”, OFC 2010, TuC3.4.

[7] M. Reiser, “A Queueing Network Analysis of Computer Communication Networks with Window Flow Control”, IEEE Transaction on Communications, COM-27(8), 1199-1209 (1979).