Design and Implementation of a Multi-Terabit Optical Burst/Packet Router prototype

F. Masetti, D. Zriny, D. Verchère, J. Blanton, T. Kim, J. Talley (1)
D. Chiaroni, A. Jourdan, J.-C. Jacquinot, C. Coeurjolly, P. Poignant (2) M. Renaud (3)
G. Eilenberger, S. Bunse, W. Latenschleager, J. Wolde, U. Bilgak (4)

Alcatel Research & Innovation
1000 Coit Rd, Plano, Tx, 75075, USA (1); Route de Nozay, F-91460 Marcoussis, France (2);
OPTO+ Route de Nozay, F-91460 Marcoussis, France (3); Holderaeckerstr. 35, 70499, Stuttgart, Germany (4)
Contact: francesco.masetti-placci@alcatel.com

Abstract: This paper reports the first demonstration of a multi-Terabit IP optical router. A sub-equipped rack-mounted prototype has been designed and assembled, demonstrating all key functions of large, scalable packet router. The design exploits burst switching techniques through to an integrated optical packet switching fab ric.

Introduction

Future routers for IP core networks will soon cross the Terabit capacity barrier, requiring vendors to investigate new architectures and technologies. In this paper, we report implementation of a rack-mounted prototype, assembled as a proof-of-concept for a scalable multi-Terabit IP optical router (TIPOR). The design is based on the new technique of burst switching coupled to an optical packet switching fabric. The technology building blocks, the design and techniques used to implement the optical packet switching matrix have been reported in [1][2]. This paper reports the final implementation and analysis of all functionalities, including burst assembly, switch control and scheduling, framers/transceivers and optical switching matrix. The viability of the approach is assessed showing experimental and simulation results.

Concept description

Current approaches to implement terabit routers consist of using parallel, multi-stage architectures switching small internal cells. We have investigated the feasibility of another approach based on a single-stage optical switching fabric and a larger switched granularity called a burst, which consists of aggregation of IP packets (or ATM cells). This concept has advantages:

- relaxed requirements for processing speed: aggregation into bursts allows increase of data rate and scalability of the fabric and router capacity with limited impact on control;
- relaxed switch arbitration and contention management: it is expected that a single-stage Time-Space-Time swi tch should be simpler to handle. Adapted algorithms have been developed to study scalability and complexity of the process. In particular, the latency can be reduced due to limited buffers in cascade.
- expected higher robustness due to simpler of the fabric structure, limited number of components, simpler monitoring.

To implement a single stage core switch, we adopted an optical switching matrix, as opposed to electronics, for these reasons:

- relaxed interconnection problems, as bit-rate increases, due to the advantage of optics in this respect. In fact, an optical switching matrix represents the natural extension of passive optical interconnects widely used;
- wire-speed switching of packets without parallel demultiplexing, as with electronics, scalable to future higher bit rates;
- availability and acceptable maturity of some optical technologies (optical gates, tunable lasers, ...) to consider their integration and exploitation in a packet switching system.

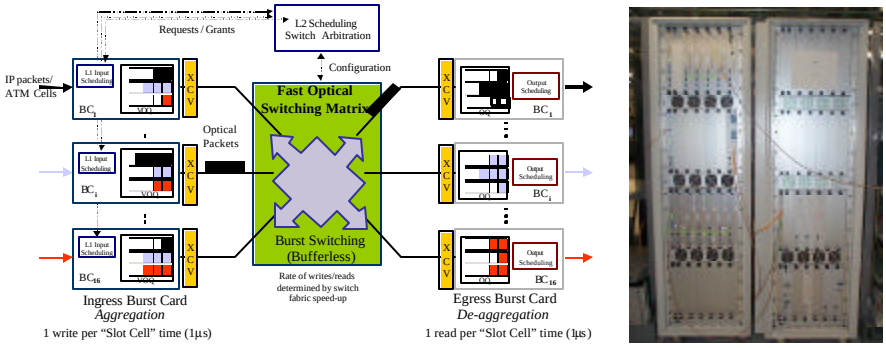


Fig. 1: TIPOR prototype architecture (left) and rack-mounted assembly (right)

Prototype implementation

The four main functions required to design our router are: (1) the aggregation of packets into bursts and contention resolution (burst card (BC)), (2) the scheduling for a single-stage multi-terabit switch, (3) the physical interface to the switching matrix (transceivers (XCV)) and (4) the optical fast switching matrix itself.

These four functions have been implemented in a rack-mounted prototype, which includes BCs with 10 Gbit/s throughput (designed scalable up to 160 Gbit/s), burst framer/transceiver boards with 10 Gbit/s throughput, and an optical fast packet switching matrix designed for 640 Gbit/s (scalable to 2.56 Tbit/s), to demonstrate all functionalities. The adopted burst size is 9 kbit/s, easy to implement, but also a reasonable trade-off between routing constraints and aggregation efficiency. The

framing includes guard-bands of 50 ns, to absorb physical constraints in the transceivers and optical switching matrix.

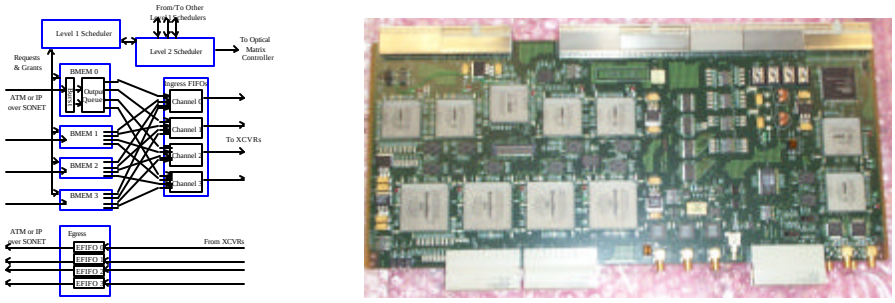


Fig.2: Burst card schematic (left) and burst card implementation (right)

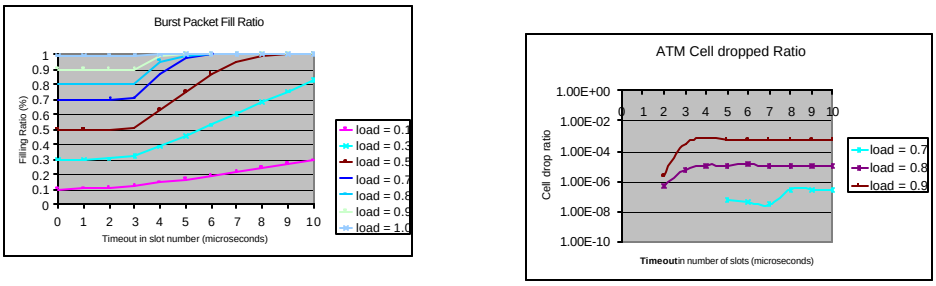
Burst card and scheduler

The BC has been implemented as follows (Fig. 2): ATM cells (16x622 Mbps), generated from line cards are injected in the BC and then are aggregated into 9 kbit/s bursts using FIFO memories. The bursts are made up of multiple cells of data that have similar classes of service and egress ports. An efficient two-level scheduling algorithm on multi-server architecture provides simultaneously the maximal matching between input port servers and output port servers during each switch slot as well as limiting the time of cells in the queues.

Virtual output queues (VOQs) are filled with the bursts. At the beginning of each switch slot, a request to the level 1 (L1) scheduler is made to reserve matrix resources, to send fully filled bursts (when a queue is full) or partially filled bursts that reached the timeout. The L1 scheduler aggregates all the requests from a module and sends them to the level 2 (L2) scheduler that has the algorithm to optimize the matrix usage. The L2 scheduler determines a conflict-free maximal matching (not necessarily a maximum matching) between input servers and output servers, by granting or not the requests. The selected flows enter aggregation ingress FIFOs going out to the transceivers. On the egress side, burst are disassembled, errors are checked and cells are reconstructed and sent out. The 2 levels of scheduling reduce complexity and decrease latency. The Binary Tree Arbiters (BTA) are multi-server version with the possibility of shuffling mechanisms to eliminate the inherent unfairness of polarized traffic. Statistics gathered on the module include cells transmitted and received, lost cells, errored cells, and queue overflows.

The number of available servers in the BC impacts the throughput performance, due to statistical multiplexing. The prototype has been designed for up to 16 BCs, each consisting of 16 VOQs and 16 servers, each serving one burst of 12 ATM cells. The BC ingress capacity is 160 Gbit/s without link speed-up between the BCs and the switching matrix.

For a simulation time of 10 ms, Fig. 3 (left) shows the burst filling ratio and (right) the ATM cells dropped at the ingress BC buffers. For a non-polarized traffic, the cell dropped ratio reaches the 10^{-6} order at 80%-input load per link for a timeout at 10



µs. A long timeout increases the fill ratio and a short timeout reduces the drop ratio.

Fig. 3: Burst filling ratio = f (timeout, load) (left) and dropped cells/packets = f (timeout, load) (right)

Framer/transceiver board

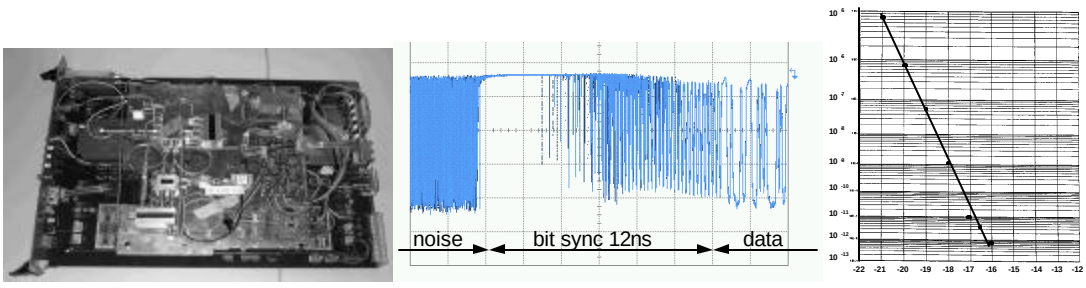


Fig. 4: The framer/transceiver board (left); the packet power and phase recovery after the packet-mode receiver (center); performance of the burst mode receiver (right)

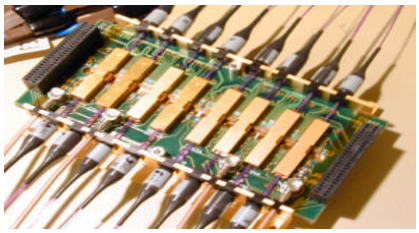
This board (Fig. 4 (left)) receives a 2.5 Gbit/s burst flow, speeds it up to 10 Gbit/s, and generates the optical packet frame (including guard band and necessary overheads – Fig. 4 (center)), and then performs the electrical to optical conversion (and vice versa on the receiving side). However, an important feature of this prototype is the use of packet-mode receivers. The

clock and data recovery (CDR) has been optimized to receive RZ formatted optical bursts at 10 Gbit/s data rate: Fig. 4 (center) shows the behavior of the CDR when a new burst arrives. In a timeframe of 12 ns after the first pulse of a leading bit synchronization sequence, the CDR adapts to the bursts signal power level and clock phase and starts to deliver the correct burst data. The performance of the CDR has been tested previously by an optical pattern containing bursts of roughly 9000 bit data (a PRBS 2^{11}-1 string), 512 bit gap between bursts, 128 bit synchronization sequence: these latters were masked for bit error rate measurement (Fig. 4 (right)).

Optical packet switching matrix:

Alcatel has designed and developed an optical packet switching matrix adopting a broadcast-and-select structure, a two-stage switch design (fiber and wavelength) and using semiconductor optical amplifiers (SOA) as switching technology (space and wavelength switching). This work was already presented in [1][2].

A prototype, representative of a sub-equipped 640 Gbit/s (8 fibres, 8 wavelengths, 10 Gbit/s) optical matrix has been realized, with particular effort devoted to limit the foot print in order to make such a matrix very compact. The architecture is scalable



up to 2.5 Tbit/s and beyond [2]. The basic switching module integrates up to 32 SOAs, grouped in 8 arrays of 4 ([1] Fig. 5 (left)). This enables to build optical switch boards (Fig. 5 (right)) of potential capacities up to 2.56 Tbit/s x 40 Gbit/s.

Fig. 5: Optical switch module (left) and board (right)

Pseudo-random sequences were programmed in the payload of the optical packets for bit-error-rate tests of the optical fabric. Since the experimental hardware is only capable of loading the optical matrix at a maximum of 25%, cell loss due to queue overflow or switching conflicts is not measurable. However, what could be measured was the cell loss rate on a given path through the matrix due to bit errors on that optical path. A path that had been measured to have a bit error rate of 10^{-11} displayed a cell loss rate 2 orders of magnitude worse, as expected. These results were expected for a noisy path due to the burst of data being discarded when random bit error occurring in the optics is detected. Since it can cause the loss of ATM cells already at a loading factor of 25%, we assessed the need to have, as accomplished on most of the paths, a bit error rate of 10^{-13} or better (Fig. 6). For this reason, monitoring and statistics graphical interfaces have been built to better collect system results (Fig. 7).

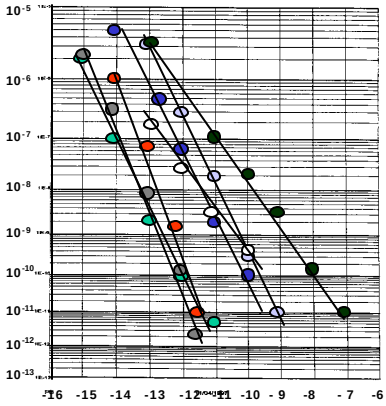


Fig. 6: Sample of BER curves of the optical switching fabric (matrix and burst transceivers)

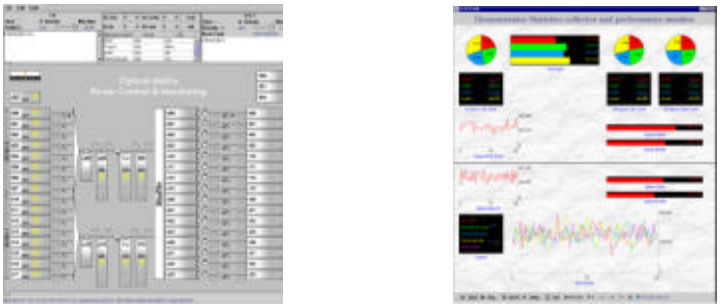


Fig. 7: Optical Matrix control & monitoring (left) and Statistics Collection Screen (left)

Conclusion

We have designed a multi-Terabit-class router exploiting burst switching and an optical packet switching matrix in its core. We have then implemented and tested a rack-mounted prototype, and this paper reports the results and assessment of the key functions required to build such a router: the burst assembly, the scheduling mechanism, the framing and burst-mode optoelectronic interfaces, and the optical fast switching matrix. This is the first such comprehensive implementation and demonstration of an IP optical router based on these principles.

References

/1/ N. Sahri et al " A highly integrated 32-SOA gates optoelectronic module suitable for IP multi-terabit optical packet routers ", Postdeadine paper, OFC 2001.

/2/ D. Chiaroni et al., “ First demonstration of an asynchronous optical packet switching matrix prototype for MultiTerabit - class routers/switches”, Postdeadline paper, ECOC 2001.