社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

Arbitration-free Time-Division Permutation Switching suitable for All-Optical Implementation

Alvaro Cassinelli[(1)], Makoto Naruse [(2)], Alain Goulet[(1)] and Masatoshi Ishikawa[(1)]

1: University of Tokyo, Dept. Information Physics and Computing, 7-3-1 Hongo Bunkyo-ku, Tokyo 113-0033, Japan.

2: Communications Research Laboratory, 4-2-1 Nukui-kita, Koganei, Tokyo 184-8795, Japan.

E-mail: alvaro@k2.t.u-tokyo.ac.jp

Abstract. *Pursuing high-bandwidth and transparency for optical communications, we propose here a multistage switching fabric whose internal interconnection patterns are periodically reconfigured. Specifically, by cascading guided-wave modules containing each a small set of interleaved, independent addressable plane-to-plane global interconnections, we experimentally demonstrate arbitration-free time-division permutation routing in a transparent multistage architecture suitable for high-bandwidth inter-processor communications in massively parallel machines. Furthermore -and perhaps more interestingly- computer simulations confirms that the addition of small inter-stage buffering nodes (suitable for being implemented as optical fiber delay lines) along with a simple and local packet flow arbitration rule (adapted to the global, periodic interconnection reconfiguration strategy) would lead to fairly good network performance for packet switching applications.*

Keywords: Guided-wave optical interconnections, multistage interconnection networks, circuit switching, packet switching.

1. Introduction

With the fast growing of Internet, larger and faster core-routers will be required for packet switching. As pointed out in [1], it is the time-consuming electronic scheduler that presently limits throughput, even when the switching fabrics are built upon efficient – but expensive – electronic crossbar switches. So, even if all the ingress/egress optoelectronic conversions could be avoided by the use of an all-optical crossbar, a complex electronic scheduler would still be required, limiting the ultimate performance of the system. Hence the interest in scheduler-free (or simple) routing strategies. In particular, this has triggered a renewed interest on the so-called shuffle-exchange multistage interconnection networks (SEMINs) [2], because these can exhibit *self-routing* capabilities (see Fig.1). These are formed by cascading active stages built upon elemental 2x2 exchange switches (the "exchange" stages), and passive interconnection stages (the "shuffle" stages). Circuit switching on SEMIN architectures have been extensively studied for handling inter-processor communications in massively parallel computers [3]. Global control (or column-control) of switches belonging to the same switching stage leads to what we will call the Global-Stage MIN architecture (or GSMIN for short), which will be simpler to implement and control. Obviously, a GSMIN have even less permutation capacity than the SEMIN from which it is derived, but as suggested in [4],[5], with fast reconfiguring switches and high-bandwidth channels, it is possible to multiplex permutations and eventually produce all the required communication primitives in due time. This approach is particularly well suited for a "transparent" implementation, because of the inherently huge bandwidth of optical channels.

When considered for packet routing purposes, blocking MINs still represent an interesting alternative to the full-crossbar both from the point of view of their implementation cost (which does not grows exponentially with the network size) but also from the point of view of the simplicity (node-locality) of the routing decisions. The price to pay is unavoidable contention of resources (switches and links) or "internal blocking", which can severely degrade performance when the network is confronted with heavy traffic, or even for certain

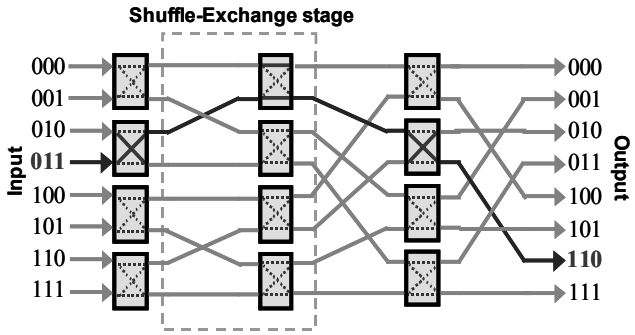

Fig. 1: The Inverse Baseline SEMIN. The stage-distributed control bits (101) result from XORing the input (011) and the output addresses (110).

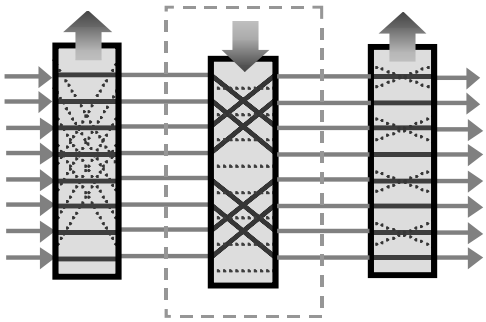

Fig.2: GSMIN paradigm: each bi-permutation module integrates switching and permutation.

particular patterns of requests. Several (more or less complementary) contention-resolution schemes have been proposed for packet switching networks, such as on-the-flight sorting before entering the switching network [6], dilation and replication [7], deflection routing (including *bufferless* hot-potato routing) [8], and of course all sort of buffering techniques. It is important to note that buffering is necessary even in a crossbar-based switching fabric - otherwise it is known that output contention would limit the throughput to a maximum of 63% for a large switch (and for uniform traffic). Inter-stage buffering is very appealing in self-routing MINs, since the node-local routing strategy marries well with the distribution of buffering functions among the different stages of the network. Last, although still unrealistic, an all-optical implementation of a FIFO buffers is somehow made easier on a MIN, since it has been noted that as few as four packets per buffer will approach infinite-buffer performances under uniform traffic [7].

This paper is organized as follows. In Section 2 we propose a possible implementation of a transparent GSMIN network suited for permutation routing, based on cascading fiber-based interconnecting modules that can be mechanically reconfigured (see Fig.2). In Section 3 the results of computer simulations of an interstage first-input-first-output buffered GSMIN architecture are presented, and performance compared with that of a standard SEMIN for packet routing purposes. A very simple reconfiguration mechanism for the switches is proposed and validated, which does not rely on a local or global arbiter. This simple routing protocol is also simulated for a GSMIN architecture where buffers have been replaced by simple delay-lines (whose implementation is within the scope of present optical technology). In the conclusion section, we summarize our results and discuss further research directions.

2. Implementation of a transparent GMSIN

While many demonstrator systems have been built to illustrate the advantages of free-space optics over electronics for dense plane-to-plane interconnections, there has been relatively little research on the use of *three-dimensional wave-guide-based* interconnections. Yet, these can easily achieve better transmission efficiency than holographic-based interconnections while almost completely cancelling cross-talk. Besides, and contrary to the common belief, they may be *more volume efficient* than free-space optics for both space-invariant and space-variant interconnects [9]. Moreover, in the case of "column/row-decomposable" permutations, which happen to be the ones required in most parallel computing algorithms, fiber modules can be easily implemented by stacking layers of printed lightwave circuits [10]. Extending our previous research on these *fixed*, dense, plane-to-plane guided-wave-based interconnections for pipelined optoelectronic systems [11], we propose here a transparent implementation of a GSMIN based on *multi*-permutation modules (Fig.3). A multi-permutation module contains a reduced set of inter-stage global interconnections, that is to say, switching of individual channels is not allowed

Although reduced, the permutation capacity of a GSMIN may accommodate the limited number of communication primitives required during the operation of a massively parallelized algorithm. Indeed, the multistage "spanned" version of most direct network topologies (hypercube, cube-connected-cycles, etc.) can be implemented as a GSMIN architecture. A time-division multiplexing (TDM) technique can then be used to select

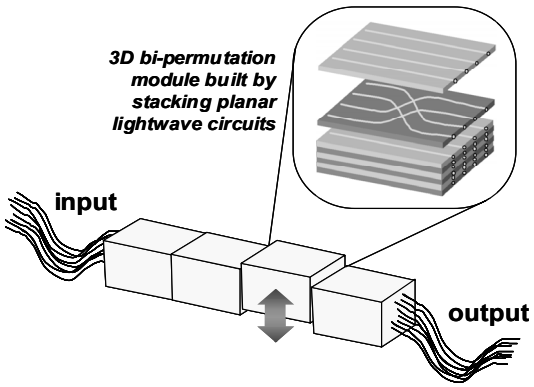

Fig. 3: An implementation example of a GSMIN: mechanically reconfigurable, cascaded guided-wave based interconnections.

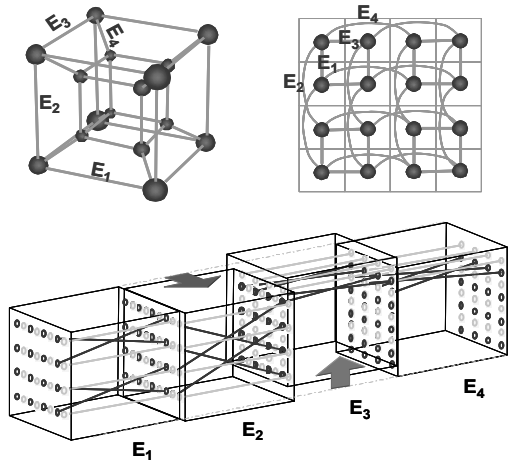

Fig.4: Four-dimensional hypercube mapped on a plane and "spanned" version using bi-permutation modules

the interconnections at each stage. Figure 4 represents a spanned version of a 4-dimensional hypercube using four bi-permutation modules, each providing a cube permutation and the identity permutation (this gives a total of 16 available global permutations). Two of these modules were actually fabricated using interleaved optical fibers, and the resulting four possible interconnections observed (Fig.5, below). The coupling efficiency between modules (without additional optics, index-matching oil nor antireflection coating) was measured to be 1.7 dB, validating the simple optical implementation. A small electro-mechanical switching device (much like a pick-up head, but with independent control in two directions –interleaving of permutations can be done in both directions) has also been fabricated and is currently being tested. Its resonant frequency is around 430 Hz, for a maximal excursion of ±62.5µm. Since the pitch of the interleaved fibers at the surface of the module is 125µm, each permutation is addressed once at each mechanical oscillation. The time slot for a transfer of data (measured as the interval during which the transmission efficiency drops below 3dB from its optimum) has been measured to be around 200µs, and the switching latency is about 0.96ms. Therefore, assuming a typical optical channel bandwidth of 10Gbit/s, the present system would be able to accommodate 2 Mbs bursts of data every millisecond (an average bandwidth of 2 Gb/s). Experiments are being carried out to measure the bit-error rate (BER) of such time-slotted communication channels. The duration of a time slot could be stretched (and the interconnection latency reduced) by using appropriate optical relays such as micro-lenses. Although the reconfiguration latency can be relatively slow, an appealing characteristic of the proposed mechanically reconfigured system is that the switch is inherently cross-talk free. Micro electro-mechanical (MEMS) actuators may also be an interesting alternative when switching latency in the millisecond range is tolerable.

Non-mechanical reconfiguration is also possible, using for instance liquid-crystal based reconfigurable holograms [12], or by combining acousto-optical (AO) beam-steering cells with motionless multi-permutation modules. Instead of actually translating the module, an acousto-optic (AO) cell placed between two fixed multi-permutation modules would globally deflect the two-dimensional array of light beams from the output of one module in order to address the required array of channels at the input face of the following multi-permutation module. Since the array size may be very small (<1mm^2 in our fiber-based prototype), an acousto-optical cell may be able to swap interconnections in the order of tens of microseconds or less.

If switching times orders of magnitude faster are required, it is always possible to combine the control lines of individual 2x2 integrated electro-optical switches as proposed in [4]. The functionality of the resulting column-controlled SEMIN is equivalent to that of a (bi-permutation based) GSMIN; however, the switching modules would not merge permutation and switching functions in the same module, which may result in a more complex system implementation.

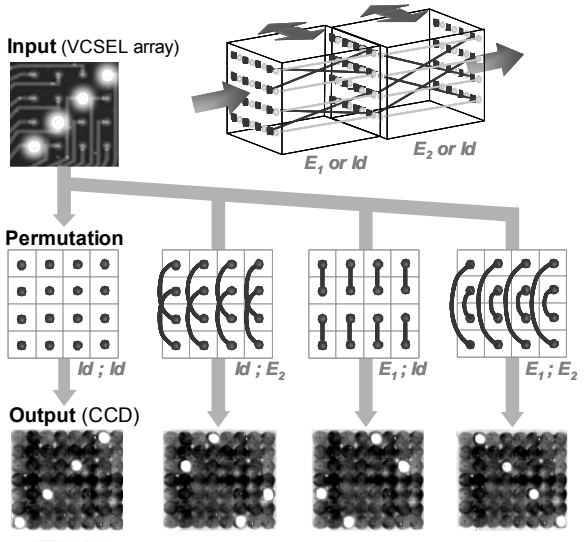

Fig. 5: Demonstration of transparent permutation switching using a pair of fiber-based modules.

3. A packet switching GSMIN

As said in the introduction, an unbuffered GSMIN architecture may be of interest as a circuit-switched permutation network, but it presents too much packet loss for packet switching purposes. However, we will show next that a proper routing strategy combined with a moderately buffered GSMIN architecture may lead to performance competitive with most standard buffered MIN architectures under random traffic load.

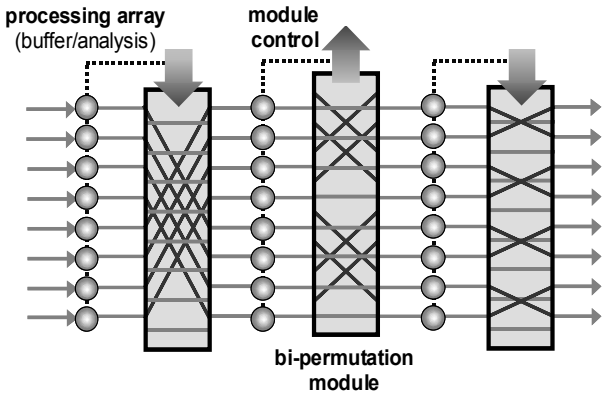

Fig. 6: Buffered GSMIN architecture (compare with fig.2).

3.1. GSMIN with inter-stage FIFO buffers

Figure 6 represents an inter-stage *buffered* GSMIN suited for packet routing. Routing conflicts are not resolved individually at the switch level, as is the case in the standard SEMINs, but *globally* at the stage level by a "tournament" between all the incoming requests to that particular stage. Provided that these request are uniformly directed to any possible output, "votes" leading to the adoption of one of the two possible states of the global-switch will be evenly distributed. Such behaviour takes place for all stages of the network, so that at each stage, half of the requests will be dropped and half will be able to pass to the next stage. This means there is an enormous number of discarded packets, certainly much bigger than that occurring by internal blocking in a standard SEMIN; however, if one considers a buffered architecture, then presumably there will be no need to provide it with a large buffer memory, because the packets that have been retained in the buffers are very likely to go forward in the following tournament (since if they are made to participate in that tournament, they will certainly bias the evenly distributed requests of the new arriving packets to "their advantage"). We can go even further and conjecture that in the particular case of truly random traffic, analysis of packet headers for selecting the global-switch state may be unnecessary: a continual "blind" alternation of switching states may perform just as well. Computer simulations have verified both conjectures. Figure 7 shows

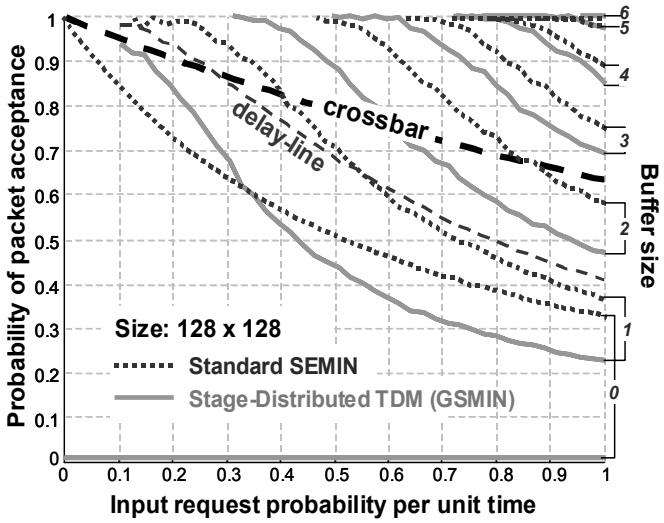

Fig.7: Performance comparison between the SEMIN and the GSMIN architectures.

performance (normalized amount of satisfied requests) as a function of the input load (computed as the probability of a request being issued at any input per unit time) for the GSMIN with stage-distributed "blind" switching, and for the standard equivalent SEMIN with individual control of switches (both 128x128 large networks). Observing the figure, we see that, as buffer size increases, GSMIN and SEMIN performance disparity rapidly decreases. Therefore, individual control of switches as well as arbitration may be unnecessary on a standard SEMIN for buffer sizes larger than three. Also, when buffer size is equal to three, the 128x128 GSMIN already outperforms a 128x128 full-crossbar for any input load.

3.2. GSMIN with inter-stage delay-Lines

Optical FIFO buffers are not feasible nowadays; we simulated then a network with simple delay-lines instead of buffers. Figure 8 represents the elemental routing node, comprising a 2x2 interconnection switch (whose state blindly alternates with each network cycle[1]), and two smaller switches that can divert the input towards a delay-line, if required. At each network cycle, the packet in the delay line is considered first. If the delay line is empty or the delayed packet still cannot go trough the switch, then the packet in the input line is considered for transfer in that cycle. If it cannot go forward, it is diverted towards the delay line. This scheme does not represent an optical memory, since a packet cannot wait more than one cycle in the "buffer". Figure 7 also shows the performance of a 128x128 delay-line based, arbitration-free MIN. As can be seen, the

[1] Individual control of switches has not been studied for this configuration.

performance of the delay-line "buffered" network falls somewhere between that of a two-size and a one-sized buffered network (this is because the delay-line and the input line can be seen as "store-and forward" buffers if the network cycle corresponds exactly to the to delay-line and the transfer delay).

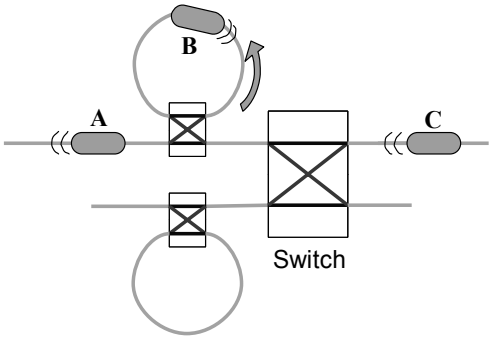

Fig.8: Schematic representation of an elemental 2x2 switch with two selectable input delay-lines.

4. Conclusions

Joint operation (column-control) of elemental switches belonging to the same stage in a standard Shuffle-Exchange Multistage network certainly reduces its overall interconnection capacity, but if things are properly designed, the architecture can still accommodate the required communication primitives of most static interconnection networks. The interest of such an arrangement lies in the ease of control and straightforward implementation. We presented here preliminary experimental results demonstrating a simple optical architecture using cascaded fiber-based bi-permutation modules. An electro-mechanical system has been developed providing stage-switching times on the order of milliseconds, making this architecture suitable for reconfigurable, high-bandwidth inter-processor communications.

Most interesting, simulations confirmed that a *FIFO buffered* column-controlled architecture would not require excessive buffer size to achieve respectable performances under uniform traffic. Moreover, it was found that the path-selection mechanism could be further reduced to simple alternation of the available permutations per stage, without degrading the performance. It is interesting to note that under such stage-distributed, time-division (permutation) multiplexing strategy, the SEMIN and GSMIN fabrics become strictly equivalent routing architectures; hence, provided that buffer size is chosen to be larger than three, this analysis-free strategy will provide a very simple arbitration mechanism for *standard* SEMIN networks. This is an interesting result on its own.

There are a number of technical reasons why an all-optical FIFO buffers may not be feasible anytime soon. Therefore, we studied the effects of replacing a buffer by a single delay-line to handle conflicts and resource contention, and showed that the performances of the resulting arbitration-free architecture compares well with a one or two-packet-sized FIFO buffered network. Performance of such architecture are not good enough for real applications, but this preliminary study tends to show that simple delay-lines and arbitration free (TDM-like) reconfiguration of the interconnection switches can be contemplated as building blocks for a real, highly scalable switching fabric.

References

[1] N. McKeown, "Optics inside Routers", Proc. of ECOC 2003, Rimini, Italy, Vol. 2, pp.168-171, 2003.

[2] A. Varma and C.S. Raghavendra, "Interconnection Networks for Multiprocessors and Multicomputers", IEEE Comp. Soc. Press, Los Alamitos, California, 1994.

[3] Y. Yang, J. Wang and Y. Pan, "Permutation Capability of Optical Multistage Interconnection Networks", J. of Parallel and Distributed Comp., Vol 60, No.1, pp.72-91, 2000.

[4] R.A. Thompson, "The Dilated Slipped Banyan Switching Network Architecture for Use in an All-Optical Local-Area Network", J. of Lightwave Tech., Vol.9, No.12, pp.1780-1787, 1991.

[5] C. Qiao and R. Melhem, "Reconfiguration with Time Division Multiplexed MINs for Multiprocessor Communications", IEEE Trans. on Parallel and Distributed Systems, Vol.5, pp. 337-352, 1994.

[6] M. Narasimha, "The batcher-banyan self-routing network: universality and simplification," IEEE Trans. Comm., Vol. 36, No.10, pp. 1175-1171, 1988.

[7] C. P. Kruskal and M. Snir, "The Performance of Multistage Interconnection Networks for Multiprocessors", IEEE Trans. Comput., Vol. C-32, No. 12, pp.1091-1098, 1983.

[8] Ching Fang Hsu, Te-Lung Liu, and Nen-Fu Huang. "Performance analysis of deflection routing in optical burst-switched networks", Proc. of IEEE Infocom, pp. 66-74, 2002.

[9] Y.Li and J. Popelek, "Volume-Consumption Comparisons of Free-Space and Guided-Wave Optical Interconnections", Appl.Opt. Vol. 39 (11), pp. 1815-1825, 2000.

[10] A. Cassinelli, M. Naruse, M. Ishikawa and F. Kubota, "*A modular, guided wave approach to plane-to-plane optical interconnects for multistage interconnection networks*", JSAP-OJ Conf., Koganei, Tokyo, pp.124-125, 2002.

[11] M. Naruse, A. Cassinelli and M. Ishikawa: "*Two-dimensional fiber array with integrated topology for short-distance optical interconnections*", Proc. Of the IEEE-LEOS Annual Meeting, Glasgow, pp.722-723, 2002.

[12] N. McArdle, M. Naruse, H. Toyoda, Y. Kobayashi and M. Ishikawa, "Reconfigurable Optical Interconnections for Parallel Computing", Proc. of the IEEE, Vol .88(6), pp.829-837, 2000.