

Stage-distributed time-division permutation routing in a multistage optically interconnected switching fabric

Alvaro Cassinelli⁽¹⁾, Makoto Naruse⁽²⁾, Masatoshi Ishikawa⁽¹⁾

1: University of Tokyo, Dept. Information Physics and Computing, 7-3-1 Hongo Bunkyo-ku, Tokyo 113-0033, Japan. alvaro@k2.t.u-tokyo.ac.jp

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

Abstract Two-dimensional fiber arrays containing multiple interconnections are proposed as building blocks for a multistage interconnection network useful for permutation routing as well as packet switching.

Introduction

While many demonstrator systems have been built to illustrate the advantages of free-space optics over electronics for dense plane-to-plane interconnections [1,2,3], there has been relatively little research on the use of *two-dimensional wave-guide-based* interconnections. Yet, these can easily achieve better transmission efficiency than holographic-based interconnections while almost completely cancelling cross-talk, and contrary to the common belief they may be *more volume efficient* than free-space optics for both space-invariant and space-variant interconnects [4]. 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 [5] (see Fig. 1). Recently, we tested this approach by implementing several 4x4 fiber-based modules each integrating a different fixed topology [6]. In this paper we study a more complex module containing a set of *independent addressable permutations*. Addressing can be done by mechanical displacement of the whole *multi-permutation module* (fig.1). Cascading such modules without intermediate optoelectronic arrays gives a transparent "globally-stage switched" multistage interconnection network (GSMIN) that can be used as a circuit-switched permutation network for multiprocessor communications. Most interesting, we found that an inter-stage *buffered* GSMIN architecture may also represent an interesting alternative to the well-known Shuffle-Exchange MINs (SEMINs) for point-to-point packet-routed communications, both from the point of view of its implementation complexity and simplicity of routing protocol.

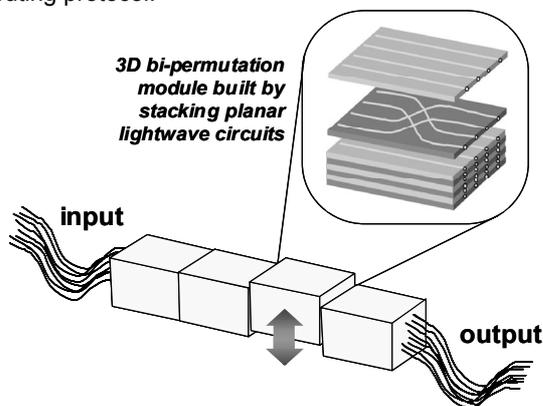


Fig. 1: Schematic representation of the Globally Stage-Switched Multistage Interconnection Network (GSMIN).

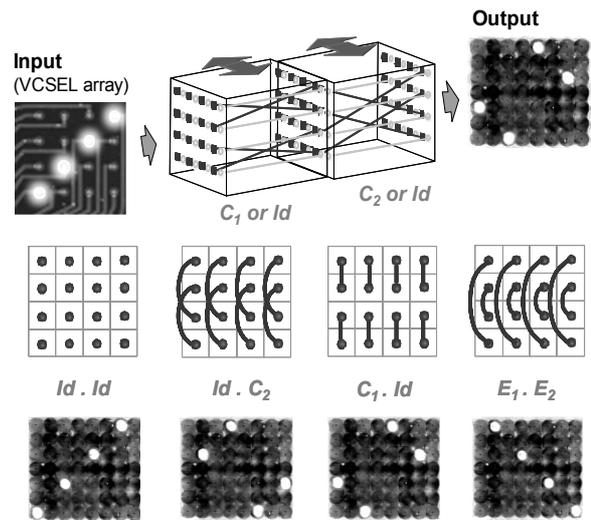


Fig. 2: Stage-distributed permutation switching in a transparent 16x16 input-output GSMIN architecture.

Multi-stage transparent GSMIN network

A multistage "spanned" version of most direct network topologies (hypercube, cube-connected-cycles, etc.) can be implemented as an *unbuffered* GSMIN architecture. A time-division multiplexing (TDM) technique can be used to select the interconnections at each stage. Figure 2 represents the first two modules of a spanned version of a 4-dimensional hypercube. The complete system would use four bi-permutation modules, each providing a cube permutation and the identity permutation, giving a total of 16 available global permutations. Two of these modules were actually fabricated using interleaved optical fibers [7], and the resulting four possible interconnections observed (Fig. 2, 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 hardware approach. Automatic alignment of modules, both dynamically [8] and statically (pre-aligned "plug-and-play" exchangeable blocks [9]), is a critical issue now being studied. A small electro-mechanical switching device has also been fabricated and is currently being tested. The switching speed seems to be limited to the millisecond range. Micro electro-mechanical actuators (MEMS) may also be an interesting alternative when switching latency in the millisecond range is tolerable.

Buffered GSMIN for packet switching

Figure 3 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 [10], but *globally* at the stage level by a "tournament" between all the incoming requests to that particular stage. Provided that these requests 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 4 shows 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 (both 64x64 large networks). Observing the figure, we see for instance that individual control of switches as well as arbitration may be unnecessary on a standard MIN if the buffer size is larger than three. Also, when the buffer size is equal to three, the 64x64 GSMIN already outperforms a 64x64 full-crossbar.

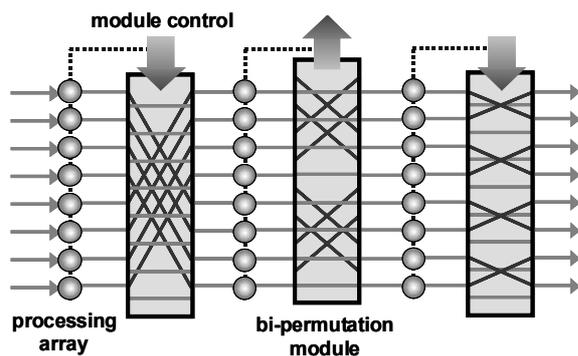


Fig.3: Three stage buffered GSMIN architecture.

Conclusions

Joint operation 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

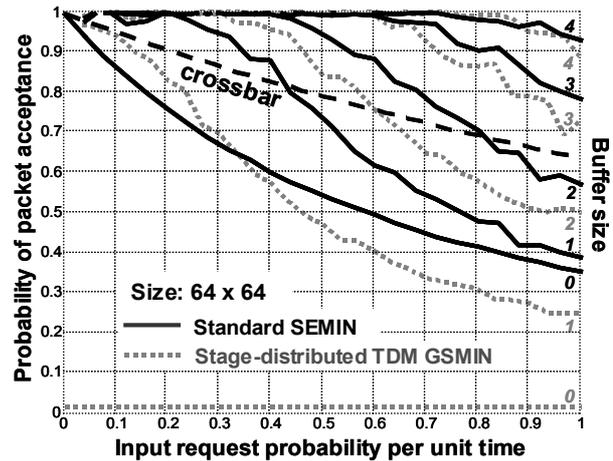


Fig.4: Performance comparison between the SEMIN and the GSMIN architectures.

interconnection networks. The interest of such an arrangement lies in the ease of control and straightforward implementation. We present 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 *buffered* GSMIN 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. Under such stage-distributed time-division permutation multiplexing, 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. Also, using an optical module-based GSMIN architecture, this paradigm may be very appealing for all-optical networks, if optical buffering functions can be integrated on the cascaded modules themselves, an issue worth exploring.

References

1. M. Ishikawa et al., IEEE Comp., **31**(1998) 61.
2. MEL-ARI OPTO REPORT, H.Neefs Ed. (2000).
3. H.M. Ozatkas, Appl. Opt. **36**(1997) 5697.
4. Y.Li et al., Appl. Opt. **39**(2000) 1815.
5. A. Cassinelli et al., JSAP Conf., (2002) 124.
6. M. Naruse et al., IEEE LEOS (2002) 722.
7. A. Cassinelli et al., JSAP OJ Conf. (2003) 1256.
8. M. Naruse et al., IEEE PTL, **13**(2001) 1257.
9. A. Goulet et al., Appl. Opt. **41**(2002) 538.
10. J.Duato et al., Int. Net. IEEE Comp. Press (1997).