

## Bending Light for Multi-Chip Virtual PRAMs?

Uzi Vishkin<sup>1</sup> and Igor Smolyaninov<sup>2</sup>

**Abstract--A new paradigm for an all-to-all optical interconnection network is presented. An interesting modeling aspect is that (limited) bending of optical communication channels is allowed. In a computer system, the paradigm could provide part of an interconnection fabric between several tens (e.g., 64) of chips comprising parallel processing elements and the first level of the cache. An optical interconnection network raises an intriguing possibility: obtain both improved performance and significant cost reduction with respect to standard serial computer system models.**

### I. INTRODUCTION

Optical interconnection networks (“interconnects”) inside computers are getting increasing attention [S02]. Assuming that processing elements will continue to mostly be electronics-based, the closer the optical interconnect is to the processing elements the more challenging the introduction of optics becomes. The need to operate at high speeds and power requirements are some of the issues. A more recent recognition of the need for stronger off-chip interconnection in high-end computing and the appeal of optical interconnects appears in [R04] (see section 2.1.4).

Modern computer design puts processing elements and the highest level of cache memories on the same large computer chip. A motivation for using a most recent VLSI technology is to allow for larger memories and higher bandwidth interconnects to be included. The use of an optical interconnect between processing elements and the first level of the cache, could replace altogether the need for a large VLSI chip based on the most advanced technology. Processing elements and caches could instead reside on several chips. These

chips could be much smaller; they could be based on older and cheaper chip technologies. If properly packaged with the optical interconnect, they could still provide good performance, but a significant reduction of the manufacturing cost. For example, rather than put 64 processing-plus-memory modules, as well as interconnect fabric, on a single expensive cutting-edge .09 (or soon .065) micron chip, one could go a few generations back and use .13, or older, micron technology for 64 (very inexpensive) chips packaged with the optoelectronic component comprising the interconnect. The optoelectronic component and the overall packaging will have to be relatively inexpensive. The manufacturing of VLSI-based products is based on first producing a mask (which can be quite expensive), followed by mass production. Often the result is a relatively low price per unit. It is quite conceivable that the interconnect concept presented below will lend itself to the same manufacturing model, reducing the price for the optoelectronic component. We have nothing to say at this stage about prospects for inexpensive packaging, but we hope that our approach will stimulate work towards this goal.

### II. A SINGLE-LAYER WAVEGUIDE MODEL

A new paradigm for an optical interconnect is presented. It could serve any level of the memory hierarchy, including between parallel processing elements and the first level of the cache. Optical interconnects are attractive since optical communication channels can cross in the same plane, and they need not be implemented using straight lines. The interconnect allows all processing of data to continue to be done in electronics. Optics is only used to transport data.

Given a plurality of modules, each comprising processing and memory elements, the interconnect provides a system of optical communication channels between every module and every other module. If the optical communication channels are implemented in the plane the following considerations are important (note that explanations appear in italic fonts):

(i) the bending of each optical communication channel must be limited *because of the finite difference in refractive index between the core and the cladding of any optical waveguide. Guiding properties of a typical waveguide are based on the phenomenon of total internal reflection (complete*

---

<sup>1</sup> U. Vishkin is with the University of Maryland Inst. for Advanced Computer Studies and Electrical and Computer Engineering Dept., College Park, MD 20742 (email: [vishkin@umd.edu](mailto:vishkin@umd.edu)). Partially supported by NSF grant 0325393.

<sup>2</sup> I. Smolyaninov is with the Maryland Optics Group (MOG) at the UMD Electrical and Computer Engineering Department.

reflection from the interface between two media having different refractive indices). A typical waveguide usually consists of a core with refractive index  $n_1$ , which is surrounded by a cladding with somewhat lower refractive index  $n_2$ . In the simplest ray optics picture of a waveguide the angular range  $\alpha$  of a guided mode propagation (due to the total internal reflection over the core-cladding interface) is given by Snell's law:  $\sin\alpha_1/\sin\alpha_2=n_2/n_1$ , where  $\alpha_1$  and  $\alpha_2$  are the incidence angles in the respective medium. Thus, total internal reflection occurs when  $\sin\alpha_1$  exceeds  $n_2/n_1$  [O82] (since for the equation to hold  $\sin\alpha_2$  needs to be larger than 1). As a result, the bending angle of an optical waveguide can not surpass  $\alpha$ , where  $\sin\alpha = n_2/n_1$ ;

(ii) if two optical communication channels cross, their angle must not be too acute (i.e., close to 90 degrees), otherwise the crosstalk between the involved channels will increase dramatically;

(iii) only two optical communication channels can cross at the same point. Otherwise the losses in each crossing channel and the crosstalk between the channels will be prohibitively large;

(iv) the distance between any two crossing points must not be too small because of the channel crosstalk issues: each crossing point gives rise to light scattering in the waveguides involved, and hence increased crosstalk between the waveguides involved; and

(v) unless near their crossing point, the distance between two optical communication channels must not be too small because of the substantial width of the waveguide cladding, which typically must be at least 50-75  $\mu\text{m}$  wide [O82] because of a typically small difference between the refractive indices of the core and the cladding.

As a first approximation, Figure 1 depicts an all-to-all straight-line geometric interconnect among 16 processor-plus-memory modules: 15 lines connect each module to the other modules. Figure 2 depicts an idea for turning Figure 1 into an interconnect. Suppose that:

(i) the diameter of Figure 1 was 20 centimeters, which corresponds to a typical wafer size,

(ii) it is implemented as a single-layer waveguide. Although multilayer geometries are possible in principle, the fabrication issues are much more complicated and expensive in a multilayer waveguide geometry,

(iii) a waveguide does not have to be a straight-line; the waveguide can be bent, but to reduce radiation losses the bent part will at no point have a radius of curvature less than 50 micrometers, which is the typical width of the waveguide cladding [O82],

(iv) two waveguides can cross in the plane, preferably with a right (90 degree) angle, which drastically decreases the undesired coupling between the waveguides; one alternative is to bend a waveguide over the other to avoid crossing in the same plane; however, such out of plane bending leads to much more complicated fabrication procedures and increased production costs;

(v) only two waveguides can cross at the same point and the distance between two crossing points is at least 100 micrometer; which is equal to two characteristic cladding widths,

(vi) unless near their crossing point, the distance between two waveguides is never less than 100 micrometer.

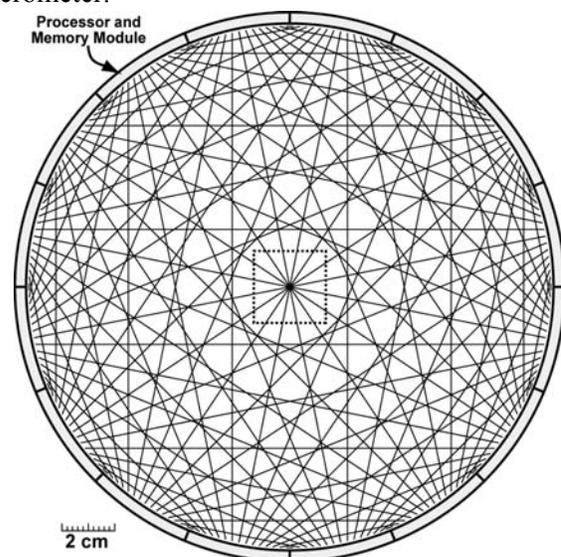


Figure 1

### III. THE ALL-TO-ALL INTERCONNECT

Figure 2 provides a simple way to satisfy all these constraints for 16 processor-plus-memory modules for the region around the center point of Figure 1, where 8 lines intersect. Combined with ad-hoc bending of lines, this scheme can be used to satisfy all these constraints for 16 processor-plus-memory modules everywhere else in Figure 1. This is done without lengthening the waveguides significantly. Although not detailed here, all these techniques

could be extended to 32, or even 64, processor-plus-memory modules. Figure 2 illustrates the main idea which enables modifying Figure 1 into an interconnect, where the interconnect satisfies the limited bending, not-acute angle, not-too-near crossings and the not-too-near channel requirements. Figure 2(a) depicts an enlarged view of the central square in Figure 1. Figure 2(b) shows how the intersection of 8 lines may be replaced by an equivalent configuration in which all crossing points are between two lines in a 90 degree angle; no two crossings are too near, no two lines are too near, except near their crossing point, and bending is limited. The figure shows how to bend the 4 lines that come from the North-West quadrant so that they all run parallel to one another; the 4 lines that come from the North-East quadrant also run parallel to one another; the former 4 lines form a grid with the latter 4 lines providing all the crossings between them where no two crossings are too close. The crossings within each group of 4 lines are obtained by recursively repeating a similar grid for each group. Figure 2(b) depicts the crossings within the 2 groups of 4 lines, and then within the 4 groups of 2 lines. The point at the center of Figure 1 is most problematic. By generating similar all-to-all straight-line interconnects among 32, as well as 64 processor-plus-memory modules and then zooming on them, one can illustrate that the situation elsewhere is much easier to handle, since no more than 3 lines intersect at the same point, and there is sufficient space for combining ad-hoc bending of lines with the solution of Figure 2 to satisfy all these constraints.

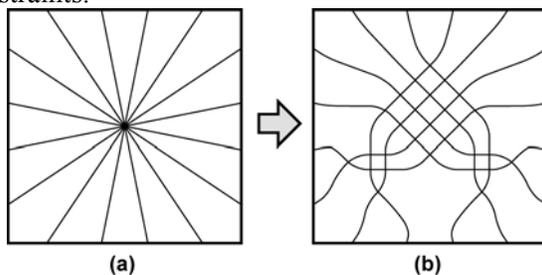


Figure 2

#### IV. DISCUSSION AND FUTURE DIRECTIONS

Depending on the exact optoelectronic technology used, the following issues, which are beyond the scope of the current paper, will need to be addressed:

(a) How to get communication rates that fit the needs of the application?

(b) The communication rates for each channel will be limited not only by the capacity of the channel but also by the capacities of the sending and receiving ends which would need to temporarily store the transmitted data; one way for regulating the aggregate rate for all channels with the same receiving end is as follows. Each channel that needs to send data to a common destination will communicate the size of the data. Using special (electronic) hardware at the destination, e.g., see the prefix-sum hardware proposed in [V03], future time slots for the transmissions on each of the channels to ensure that the amount of data received at any point in time can be safely handled will be computed and sent back the sending ends.

(c) Thermo-modeling: translation of optics-to-electronics and back and driving optical signals to accomplish our performance objective requires considerable power; how to evaluate the resulting heat and minimize it?

Overall the power issues are quite different than in an all-electronic solution, where such translation is not needed, but, on the other hand, much more power is required to drive the signal over wires. An overall “apple-to-apple” comparison of power requirements would not be a trivial task.

(d) Spacing: what is the correct stacking density of processor-plus-memory modules in view of this thermo modeling? The larger the heat load, the larger the diameter of the interconnect has to be to facilitate cooling; since the speed-of-light is 30cm/ns, a too large diameter could increase latencies by too much for the application.

(e) If waveguide technology is used, what would be the most appropriate waveguide technology? Will it be silica-on-silicon? *While such technology would be the most convenient from the fabrication point of view, progress in achieving silicon-based light emitting devices has been slow. Thus, alternative gallium arsenide based, etc. technologies may be considered.*

(f) How many crossings can we allow for each waveguide and still meet performance objectives? for a 64 module interconnect, a waveguide may cross up to 1000 others; this seems to allow a loss of no more than 0.1% per crossing assuming equal losses in each crossing point, which requires special attention. Under such conditions about 36% of the original optical power in a channel will reach the receiver, which is enough for the channel operation. *Such a low limit on the amount of*

channel crosstalk at a single intersection may require the use of somewhat more expensive out-of-plane bending, in which the amount of crosstalk may be reduced to zero. This fact may indicate ultimate desirability of the out-of-plane technology in the waveguide-based optical interconnects.

(g) How big will the radiative/scattering loss be? This question requires a detailed experimental study. From the waveguide theory [O82] and the simple geometrical arguments (Fig.3) it may be argued that the scattered energy will be proportional to  $(n_2-n_1) \times (\text{waveguide crossing area}) / (\text{core area affected by the crossing}) = (n_2-n_1)(\pi d^2/4) / (2\pi d^2/\cos\alpha) = (n_2-n_1)\cos\alpha/8 \sim 3 \times 10^{-4}$ , where  $\alpha$  is the angular propagation range of a guided mode described above,  $d$  is the core diameter, and  $(n_2-n_1) \sim 3 \times 10^{-2}$  is the characteristic refractive index difference between the core and the cladding (e.g., note that there will be no scattering if  $(n_2-n_1)=0$ ). This estimate is below the maximum allowed limit. Note that it may be further increased by the various imperfections of the crossing point. Thus, further experimental study of this problem is needed.

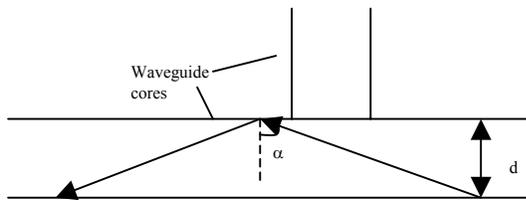


Figure 3

(h) Will the waveguide approach, or any other approach, lend itself to low-cost mass production similar to mask-based VLSI? Next, recall that we seek a substitute for a large on-chip design. The cost for 64 modules that are much smaller is going to be minimal, as they could rely on older VLSI technologies. So, if the interconnect and its overall packaging become affordable, the whole approach becomes affordable as well.

(i) Will approaches other than using waveguides, such as free-space optics [DSM03] or fiber optics work better? Free space optical interconnects work by point-to-point transmission of light signals via a system of lenses and/or mirrors. Such systems may have zero crosstalk between the communication channels and unlimited number of crossing points between the channels. However, the fabrication costs of such optical wireless systems on a chip may be quite high, and at the moment unpredictable. Fiber optics systems in which individual dedicated fibers provide point-to-point connections have basically the same advantages. However, such technology does not exist now, and the costs of developing such technology are unpredictable too. The main advantage of the proposed waveguide optical interconnects is the fact that this proposal is based on the existing well-known technology.

(j) An alternative optoelectronic implementation approach could rely on a 2-layer implementation. In this case, only two optical communication channels could cross at the same vertical point but they have to be in a different layer (i.e., same X, Y coordinates, but different Z coordinate). Limited vertical bending of an optical communication channel in order to advance from one layer to another is allowed. The same design as above could be used, but where: (1) unless near a crossing the optical communication channels are all in the same layer, and (2) near each crossing one of the communication channels bends vertically into the other layer, and then bends back again into the first layer.

(k) Difficulties in integrating emitter and detectors on silicon may call for putting emitters (and possibly all optical components) on a separate (GaAs) plane, since at least at the moment GaAs-based optical transceiver and waveguide technology is much more mature and better developed.

(l) Flight time including conversion between electrical and optical signals: we expect that this will be dominated by the time to cross the interconnect at the speed of light (which will be below a nano-second if the diameter is 20 cm). Note that standard 40Gb/s optical transmission rates are implemented using faster than 40GHz devices for conversion from optical to electronic signals (using a single photo diode) and for conversion from electronic to optical signals (using

a single laser diode). As the conversion is done serially, the conversion latency per bit is at most .025 nano-seconds.

#### IV.1 A COMPUTER ARCHITECTURE QUESTION

The motivation for this work came from the area of parallel computing and in particular from our PRAM-On-Chip research project. Although massively parallel processors (MPPs) provide the strongest available machines, recent studies demonstrate that, due to their coarse-grain parallelism, MPPs have not been a success for some general-purpose applications and in particular applications having irregular parallelism [ITR]. To many users, programming them is “as intimidating and time consuming as programming in assembly language” [A03]. Achieving programmable, high-performance general-purpose parallel computing has been the objective of the explicit multi-threaded (XMT) fine-grained parallel on-chip computer architecture framework (of the PRAM-On-Chip project) in [N+03]. A substantial challenge for an XMT design is to provide connectivity between the many execution units and the many cache modules, on-chip. While the capacity for sending signals increases with technology shrinkage, the latency for propagating signals down a fixed-length wire is increasing. Due to the memory model supported, memory requests can travel to any memory location on the chip. A latency cost for such memory accesses cannot be avoided. Fortunately, the “independence of order semantics (IOS)” of XMT threading allows for such latency to be tolerated. (IOS does not inhibit progress irrespective of the order in which parallel memory requests are satisfied. Also, using high bandwidth interconnects to minimize memory stalls due to higher latencies is a known idea in parallel computing; this is key to understanding why the latencies due to the distances in the presented optical interconnect do not inhibit high performance.) [N+03] is based on supporting simultaneous requests by pipelining throughout a powerful all-electronic interconnection network [NV01], which overcomes two problems: (a) Providing a centralized scheduling resource to coordinate communication would be costly for a large design. (b) Driving a fast global clock across a deep submicron chip is also very difficult and power consumptive. The

solution was to use a decentralized routing scheme. The hardware cost of tagging and local switching structures is justified by the benefits of such an asynchronous or loosely synchronous structure, as both [NV01] and [BQV04] as well as the current paper provide. A variety of applications on an XMT architecture simulator were studied in [N+03]. Their simulation results are applicable here. These results are reviewed in Appendix A.

One example for a computer architecture question would be comparing performance of an all-electronic 0.09 micron XMT chip design versus a design comprising 64 chips in 0.13 micron technology packaged with the optoelectronic interconnect. Below, we derive a preliminary answer to this performance question from the experimental data provided in Appendix A. More detailed and informative answers will be the subject of future studies.

The largest XMT configuration in Appendix A comprises 32 clusters, each with 8 thread control units (TCUs). Appendix A reports speedups of 16-240X (i.e., by factors ranging between 16 and 240) relative to fastest serial solution for the same problem over a range of applications. It is reasonable to expect that for a configuration of 64 clusters each with 16 TCUs these speedups will scale up by a factor of at least 2. Now, for the sake of the current example, let us assume that the clock rate of a .09 micron TCU will be up to twice as fast as the .13 micron one; so, we are back with the same range of 16-240X speedups relative to the fastest serial version for the same problem.

#### V CONCLUSION

This paper envisions parallel computer systems where optics is generally responsible for communication, but where electronics continues to dominate processing. It should be recognized that while there has been a huge multi-decade technology investment in all-electronic approaches, there had been no motivation to develop some of the technologies required for the current paper.

One of our goals will be to reach out to optoelectronics researchers, inviting them to think about the best way to implement the interconnect presented and its packaging, and/or come up with competitive alternatives. Co-author I. Smolyaninov and 4 other colleagues in the area have already helped a lot in presenting this work.

## ACKNOWLEDGEMENT

Helpful discussions and comments by the following optics and optoelectronics specialists are gratefully acknowledged: C.C. Davis, M. Dagenais, A. Iliadis and T. Murphy.

## REFERENCES

- [A03] D.E. Atkins (Chair). *Revolutionizing Science and Engineering Through Cyberinfrastructure*. NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure. 1/2003.
- [BQV04] A. Balkan, G. Qu and U. Vishkin. Arbitrate-and-move primitives for high throughput on-chip interconnection networks, to appear in Proc. IEEE International Symposium on Circuits and Systems (ISCAS), Volume II, pages 441-44, SoC Design Technology lecture session, Vancouver, May 23-26, 2004.
- [DSM03] C.C. Davis, I.I. Smolyaninov and S. Milner. Flexible optical wireless links and networks. *IEEE Communications Magazine* 42, 51-57 (2003)
- [ITR] *Information Technology Research: Investing in Our Future*. President's Information Technology Advisory Committee, 1999, [www.ccic.gov/ac/report/](http://www.ccic.gov/ac/report/).
- [N+03] D. Naishlos, J. Nuzman, C-W. Tseng and U. Vishkin. Towards a First Vertical Prototyping of an Extremely Fine-Grained Parallel Programming Approach. *TOCS* 36,5 pages 521-552, Springer-Verlag, 2003 (Special Issue for the 13th ACM Symposium on Parallel Algorithms and Architectures, SPAA 2001).
- [NV01] J. Nuzman and U. Vishkin, Circuit architecture for reduced-synchrony on-chip interconnect. US Patent Application 60/0297,248, 6/2001. Allowed: May 2004.
- [O82] T. Okoshi, *Optical Fibers*, Academic Press, New York, 1982.
- [R04] D.A. Reed (editor). Summary of the Workshop on The Roadmap for the Revitalization of High-End Computing, Washington, D.C., 6/2003, commissioned by the White House Office of Science and Technology, Computing Research Association, 1/2004. [http://www.nitrd.gov/hecrtf-outreach/20040112\\_cra\\_hecrtf\\_report.pdf](http://www.nitrd.gov/hecrtf-outreach/20040112_cra_hecrtf_report.pdf)
- [S02] N. Savage, Linking with light, *IEEE Spectrum*, August 2002, 32--36.

[V02] U. Vishkin, Spawn-join instruction set architecture for providing explicit multithreading (XMT), US Patent 6,463,527, October 8, 2002.

[V03] U. Vishkin, Prefix sums and an application thereof. US Patent 6,542,918, April 1, 2003.

[V04] U. Vishkin, Optical interconnect structure in a computer system and method of transporting data between processing elements and memory through the optical interconnect structure. US patent application, March 2004.

## APPENDIX A: SIMULATION RESULTS FROM [N+03]

Assuming similar throughput to the all-electronic interconnect, the change to an optical interconnect will affect performance only marginally. A brief review of [N+03] follows. To increase resource utilization and to hide latencies, a set of thread control units (TCUs), which can be thought of as stripped down processing elements, can be grouped together to form a cluster. The TCUs in a cluster share a common pool of functional units, as well as memory access and prefix-sum access resources. The clusters can be replicated on a given chip. The simulations assumed 8 TCUs per cluster. Assumptions regarding various memory and inter cluster communication latencies and the number of functional units per cluster are reported in [N+03]. Configurations were simulated with 1, 4, 16, 64, and 256 TCUs (namely the largest number of clusters simulated was 32; the 1 and 4 TCU configurations obviously had fewer than 8 TCUs per cluster). The number of TCUs per cluster indicates the number of simultaneous execution contexts. It does not imply hardware functionality equivalent to the same number of standard microprocessors. Applications considered were: jacobi (a 2D PDE kernel), tomcatv (mesh generation), mmult (matrix multiply), dot (dot product), image convolution, and two database kernels – dbscan from SQL and dbtree from MySQL. These programs feature regular computations that operate on different entries of a data structure independently of one another. Irregular, and more challenging, applications included: Quicksort, Radix sort, graph traversals: dag (searching a directed acyclic graph) and treeadd, and perimeter (computing the total perimeter of a region. The speed-ups obtained, relative to the best serial version, were:

