Can Optical Interconnects Lead to Cheaper High-Performance Multiprocessors?

Uzi Vishkin[1]
The University of Maryland Inst. for Advanced Computer Studies and Elecrical and Computer Engineering Department, College Park, MD 20742

ABSTRACT

A new paradigm for an all-to-all optical interconnect is presented. It could be part of an interconnection fabric between parallel processing elements and the first level of the cache in a computer system. Parallel processing has traditionally aspired to improve performance of such systems. An optical interconnect raises a new possibility: obtain both improved performance and significant cost reduction with respect to standard serial computer system models.

Keywords: Multiprocessor interconnection, optical interconnections.

I. INTRODUCTION

Optical interconnection networks ("interconnects") inside computers are getting increasing attention [8] and [7]. 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.

Modern computer design puts processing elements and the highest level of cache memories on the same large computer chip. A motivation for using 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 .065 micron chip, one could go a few generations back and use .25 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.

II. THE ALL-TO-ALL INTERCONNECT

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: (i) the bending of each optical communication channel must be limited, (ii) if two optical communication channels cross, their angle must not be too acute (i.e., close to 90 degrees), (iii) only two optical communication channels can cross at the same point, (iv) the distance between any two crossing points must not be too small, and (v) unless near their crossing point, the distance between two optical communication channels must not be too small.

[1] vishkin@umd.edu; phone 301 405-6763; fax 301 314 9658; partially supported by NSF grant 0325393.

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 (ii) it is implemented as a single-layer waveguide, (iii) a waveguide does not have to be a straight-line; the waveguide can be bent, but to reduce radiation loses the bent part will at no point have a radius of curvature less than 50 micrometers, (iv) two waveguides can cross in the plane, preferably with a right (90 degree) angle; one alternative is to bend a waveguide over the other to avoid crossing in the same plane, (v) only two waveguides can cross at the same point and the distance between two crossing points is at least 100 micrometer, (vi) unless near their crossing point, the distance between two waveguides is never less than 100 micrometer.

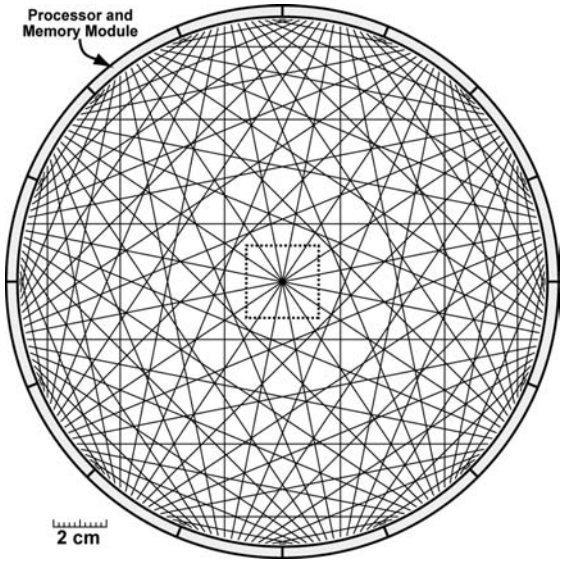

Figure 1

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 line 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.

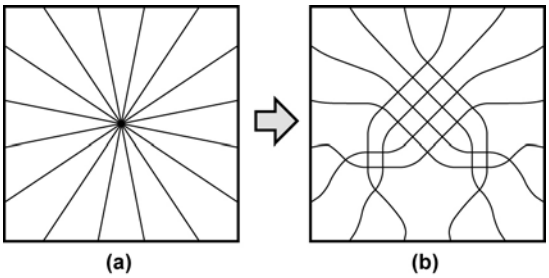

Figure 2

III. 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, see e.g. [10], 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? (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 .05dB per crossing, which requires special attention. (g) How big will radiative/scattering loss be? (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 or fiber optics work better? (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.

Our motivation came from *parallel computing*. 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 [3]. To many users programming them is "as intimidating and time consuming as programming in assembly language" [1]. 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 in [4]. 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.) [4] is based on supporting simultaneous requests by pipelining throughout a powerful all-electronic interconnection network [5] 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 [5], [2] and the current paper provide.

A variety of applications on an XMT architecture simulator were studied in [4]. Their simulation results are applicable here. Assuming similar throughput to the all-electronic interconnect, the change to an optical interconnect will affect performance only marginally. A brief review of [4] 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 [4]. 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, are reported in Figure 3.

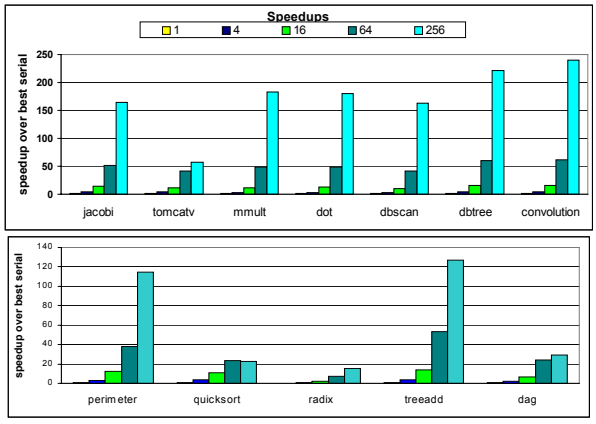

Figure 3

IV CONCLUSION

This paper envisions parallel computer systems where optics is generally responsible for communication, but where electronics continues to dominate processing. Written by a computer scientist, it aims 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.

A more elaborate computer architecture context and a more elementary introduction to optics, as would be appropriate for most computer scientists, are provided in [11].

ACKNOWLEDGEMENT

Helpful discussions and comments by I. Smolyaninov, C.C. Davis, M. Dagenais, A. Iliadis and T. Murphy are gratefully acknowledged.

REFERENCES

[1] D.E. Atkins (Chair). *Revolutionizing Science and Engineering Through Cyberinfrastructure*. NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure. 1/2003.
[2] A. Balkan, G. Qu and U. Vishkin. "Arbitrate-and-move primitives for high throughput on-chip interconnection networks". *Proc. IEEE International Symposium on Circuits and Systems (ISCAS), Volume II, pages 441-444, SoC Design Technology lecture session*, Vancouver, May 23-26, 2004.
[3] *Information Technology Research: Investing in Our Future*. President's Information Technology Advisory Committee, 1999, www.ccic.gov/ac/report/.
[4] 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).
[5] J. Nuzman and U. Vishkin, "Circuit architecture for reduced-synchrony on-chip interconnect". US Provisional Patent Application 60/0297,248, 6/2001. .Allowed: May 2004.
[6] T. Okoshi, *Optical Fibers*, Academic Press, New York, 1982.
[7] 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. See Section 2.1.4.
[8] N. Savage, "Linking with light", *IEEE Spectrum*, August 2002, 32--36.
[9] U. Vishkin, "Spawn-join instruction set architecture for providing explicit multithreading (XMT)", US Patent 6,463,527, October 8, 2002.
[10] U. Vishkin. "Prefix sums and an application thereof". US Patent 6,542,918, April 1, 2003.
[11] U. Vishkin and I. Smolyaninov. "Bending light for multi-chip virtual PRAMs?" *Proc. 3rd Workshop on Non-Slicon Computation, held in conjunction with the 31st International Symposium on Computer Architecture (ISCA 2004)*, June 19-23, 2004, Munich, Germany, 32-38.
[12] 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.