Why future supercomputing requires optics

H. John Caulfield and Shlomi Dolev

Could optical technology offer a solution to the heat generation and bandwidth limitations that the computing industry is starting to face? The benefits of energy-efficient passive components, low crosstalk and parallel processing suggest that the answer may be yes.

There are points in time at which a well-known technology, which has been used and gradually improved for decades, reaches its physical limitations, forcing a shift towards dramatically different technology. Automobiles, television, computers, photography and communication are all domains in which such shifts have occurred. The point at which these massive shifts in technology occur is often decades after the initial invention and often requires substantial investment.

The time is fast approaching for such a shift in computer technology, as the frequency (bandwidth) limitations of silicon electronics and printed metallic tracks are reached. The computer industry has already acknowledged the situation by introducing multicore technology which implicitly recognizes the use of a distributed and parallel architecture to scale computer power.

For decades, optics have proved to be the best means of conveying information from one point to another, as can clearly be seen by the growth of the massive fibre optics communications industry. The fact that light beams can be easily transmitted in parallel in free space, and do not suffer from crosstalk, is of great importance. The fabrication of multicore chips requires fast buses to connect cores, and new designs are considering the use of optical interconnects to aid reliable and fast communication[1]. Indeed, the first steps towards commercializing the use of optics for communication among cores were recently reported by firms such as Intel and Avago, among others[2–5].

But optical technology also has the opportunity to go beyond being just a convenient pipe for ultrafast data transmission, and actually perform data processing. Given that the basic computing element of a computer is a transistor — an electronic switching element — and that great progress has been made in all-optical switching and logic, there is an opportunity to avoid light-to-electronics and electronics-to-light conversions in the design of all-optical computers.

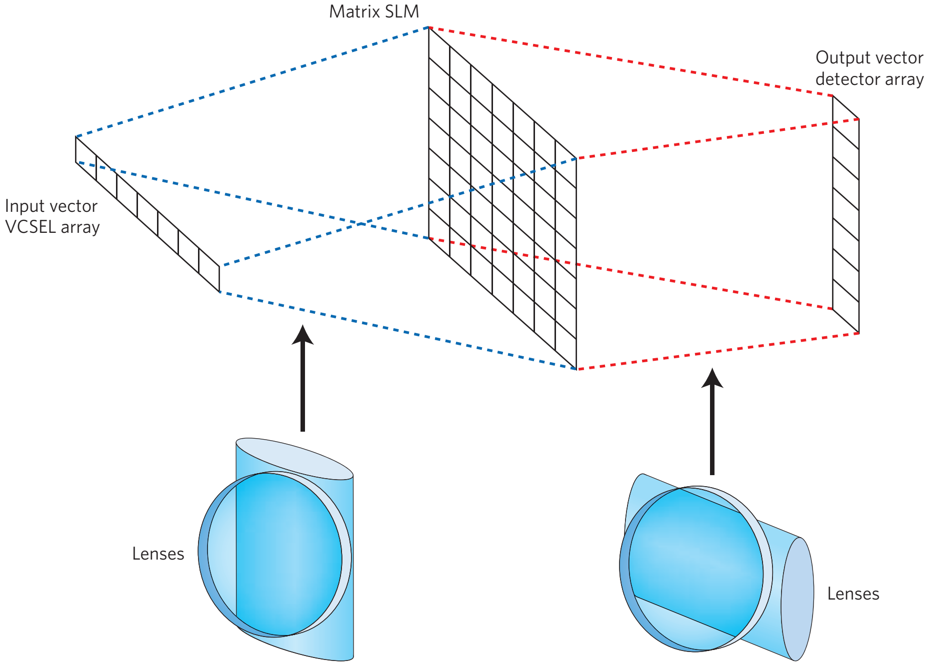

Figure 1 | The Stanford Vector Matrix Multiplier. The use of a parallel approach greatly improves computing performance. VCSEL, vertical cavity surface emitting laser; SLM, spatial light modulator. Figure reproduced with permission from ref. 11, © 2009 OSA.

It is important to note that the design of optical computers does not necessarily have to mimic the design of the electronic computer, just as the design of electronic computers did not mimic the design of mechanical computing devices. Optics implies a new means for the application of parallelism, with many beams in space representing individual computations. Recently, attempts to produce industrial-scale optical processing devices have been partially successful[6]. In addition, recent research activity in academia[7–10], and the United States government contract reviews from the Air Force and from DARPA, are clear signs for great excitement in the field.

The widely accepted Strong Church–Turing thesis asserts that anything can be computed by a digital computer. Accepting this suggests that optics offers no new computational capability. To be useful, the best that optics can do is to accomplish the same goals as digital electronics, but in a more efficient or advantageous manner. The advantage could be in speed, energy consumption, heat generation or another metric.

Speed is a frequently suggested advantage for optical computing. Ironically, the speed of optical processing is ultimately limited by the speed of electronic input and output. In essence, optical processing is inextricably tied to Moore's law for electronic computing, and the maximum benefit that optics can offer today is to replace electronic functions that slow system speeds. It is not the speed of components that counts; it is the speed of the supercomputer as a whole. Optics can help in this respect. An algorithm's computational complexity is normally the way algorithms are compared. Electronics has two ways to handle complexity: the use of space (parallelism) and time. Optics has a third way: fan-in and fan-out, in which many independent beams may be modified by a single pixel in an optical processor.

In principle, supercomputers based on optical technology may help to mitigate the heat generation that plagues electronic computers. The heat problem of electronic devices becomes much worse as speed increases and density decreases, both of which are scaling rapidly according to Moore's law (the two-year doubling in the number of transistors on a silicon chip). Early workers envisaged optical computers as rivals of electronic computers. It now appears that they are allies in keeping Moore's law alive. Faster operations require more power, while smaller components provide less area through which to remove the resulting heat from each element. The increased use of optical components that require little or, in the case of passive devices, no heat dissipation may prove very important. There are several approaches that use the special characteristics of light for information processing. They are described in turn below.

Vector matrix multiplier

A vector matrix optical multiplier (Fig. 1) uses a lens to split the light received from each entry of an input vector to the corresponding entries of the vectors of the matrix. When a beam traverses the matrix, the intensity of the beam is changed according to the value of the entry that the beam reached. Another lens is used to sum the resulting light intensity of each resulting vector. A high-performance, power-efficient, optical digital signal processor (DSP) has recently been described[11]: when compared with current technology, it provides a gain in processing speed of two to three orders of magnitude while consuming significantly less power. Its operation can be understood by the architecture presented in Fig. 1, the Stanford multiplier. The input is a vector of light sources that are projected in parallel by lenses on the rows of a slide or on a spatial light modulator, and the output that encodes the multiplication product is obtained by gathering light from each row to a sensitive detector. Loading data in parallel to the input vector and the spatial light modulator leads to a great improvement in computing speed. The architecture supports principal DSP primitives such as vector-by-matrix and matrix-by-matrix multiplications, convolution, correlation and discrete Fourier transforms. Functioning as a co-processor of a legacy processor, the architecture can efficiently process integers as well as complex numbers and perform extended-precision arithmetic. These building blocks enable DSP-intensive applications, and the finding of optimal solutions to several non-deterministic polynomial-time complete problems using exponential space. The device operational rate, of the order of at least 16 tera eight-bit integer operations per second, and power rating, of more than 16,000 giga eight-bit instructions per watt, show great promise for enhancing existing applications and introducing new optical computer applications such as video compression or a three-dimensional geometry engine. The applications of such a device include motion estimation, string matching or solution to more instances of NP-hard (non-deterministic, polynomial-time) computing problems.

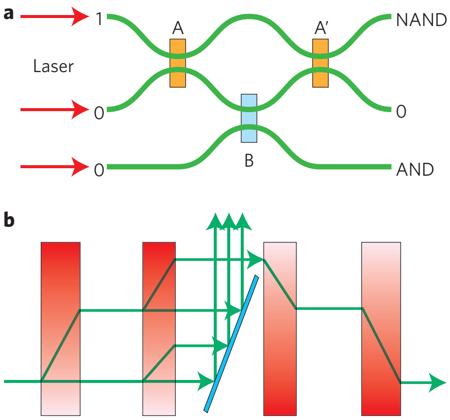

Figure 2 | The concept of 'zero-energy' logic gates using optical passive elements. **a**, Powered electronic elements to perform AND and NAND boolean functions. A, A′ and B are the inputs of the device. **b**, Rochon prisms.

Zero-energy logic

In the 1960s, the study of the physics of computing rapidly focused on the fundamental minimal operation required — the Boolean logic gates. Rolf Landauer[12] showed that the energy needed to operate such a gate stems from the need to erase one of the output bits, as the Boolean logic gates have two binary inputs but only one binary output. Shortly after that, Bennett, Landauer, Keyes and others realized that if the extra bit was kept, there would be no minimum energy per gate operation. Aside from the energy needed to input data and readout results, the logic itself could be free.

Recently, the concept of logic systems that consume zero energy has been proposed. Two independent ways to do this were recently discovered by Caulfield, Shamir, Hardy, Soref and others[13–15], accomplishing the theoretical thermodynamic 'permission' to do so.

Optics, when compared with electronics, has two important advantages — superior fan-in and fan-out, and signal transmission without the need to apply a voltage. Once generated, light propagates by itself according to Maxwell's equations, and today we can manipulate the apparent speed of light in some striking ways, without violating Maxwell's equations or relativity. We also have at our disposal numerous passive elements for manipulating light that require no energy to be supplied and do not generate heat, for example waveguides, fibres, mirrors, prisms and beamsplitters.

Figure 2 presents two designs for optical implementation of an energy-free logical gate. Both designs are zero-energy devices. Figure 2a is an example of using a new kind of logic (called directed logic) that gives the same result as conventional Boolean logic but does so in a way that is optics-friendly[13]. This device produces both A AND B and also its complement A NAND B. Note that NANDs can be combined to implement any other kind of Boolean logic gate. Figure 2b shows a generalized optical logic element[14,15]. It is essentially a physically embodied lookup table that computes A AND B, with NAND obtained by the use of additional Rochon prisms. By choosing the beams to deflect or not to deflect, we can implement any Boolean logic gate using only passive components (Rochon prisms of two different displacements and deflectors, shown in the figure as mirrors).

Optical pre-processing

Designs based on optical pre-processing (repeated copying and doubling the result iteratively) yield an exponential growth in the number of results with the number of iterations. The idea is that an intermediate result is recorded as an optical pattern on a film and then projected and (optically) modified to gain the next multiplicative step over a larger film, thus gaining exponential size results within a linear number of steps. The doubling process can be executed once, as a pre-processing stage, to support any future input. For example, all the Hamiltonian paths in a clique can be represented in a very large matrix, with the transparent points in each line of the maxtrix representing a single Hamiltonian path. When an actual weighted graph is given, the weight vector of the graph is multiplied in parallel by the pre-processed matrix to obtain the best Hamiltonian path.

Pre-processing by modifying and copying previously obtained results has proved to be useful in architectures that are based on the representation of Turing machine configurations as a location in three-dimensional space[16]. Traditional microprocessors support a set of computation primitives, such as addition and multiplication. A new set of primitives that fits the optics capabilities, and is based on the Turing machine mapping to configurations, is suggested in ref. 17. One such representation is designed to solve the Hamiltonian path problem, or in fact the 'travelling salesman' problem. Intuitively, if we allow many light beams to traverse all possible Hamiltonian paths in parallel, and design the traversal length to be proportional to the edges' weights, the beam that succeeds in doing so first must have traversed the shortest Hamiltonian path. In some cases, the number of beams that arrive simultaneously at a location representing a city may grow exponentially with the number of cities, creating a number that can overwhelm any physical device. Moreover, the traversal history of a beam can be lost. The suggested apparatus uses a distinct location for each possibility of beam traversal[16].

Figure 3 shows three columns, with each column representing a city. The arrival of a beam at a certain slot in a column corresponds to a particular traversal path; the higher the slot is, the longer the path it represents. This allows detection of whether there exists a Hamiltonian path. The architecture is designed for a fully connected graph that allows the representation of any input graph by the use of barriers that block the light traversing from one city to another. The barriers prevent light beams from traversing edges that do not exist in the input graph. Some locations in a column represent a path that has traversed a certain city more than once; pre-processing by copying masks block the propagation of light from these locations.

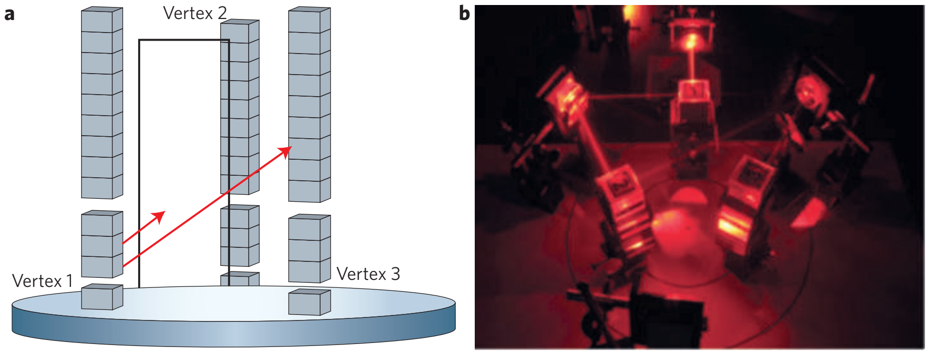

Figure 3 | Architecture for solving Hamiltonian path problem. **a**, Each column represents a city, and each box in a column represent the history of a light beam that may arrive there. Red arrows represent light beams; the left one is blocked by a barrier that represent a non-existing edge from vertex 1 to vertex 2. The lowest box is a path that starts in the vertex represented by the column; the second group of three boxes represents a path that started in another vertex and continued to the vertex represented by the column; the last group of nine boxes represent paths of length two. **b**, Photo of a typical set-up, reproduced with permission from ref. 17, © 2007 Springer.

An outlook for the future

Ultimately, electronics and optics in computing are more complementary than competitive, and all applications of optics will involve some degree of electronics. The development of faster and smaller electronics makes optical processing ever more useful and attractive in reducing the heat load and in increasing the net speed of a computing system. What is important is how to partition a task so that the net effect of both electronics and optics is optimized in a well-defined manner. Major companies, universities and government agencies are now studying the role that optics can have in obtaining the advantages of faster operation at lower heat burden.

Although quantum entanglement is a hot topic among physicists and computer scientists, it seems premature to expect it to solve our computing problems in the near future, as it is still at the stage of basic research. Currently, classical optics alone seems to be ready to join electronics on the silicon chip to realize operations more powerful than either technology alone can deliver. Optical interconnects are already used in practice to help remove electronic bottlenecks[18–20], and optical-based zero-energy logic may relieve some of the heating problems that Moore's law implies.

The appropriate question today is no longer 'why bother with optical computing?' but rather 'what is the best way to incorporate optics into future supercomputers?' There are great challenges in the actual prototyping for mass production of optical devices, in a way similar to the difficulties in first prototypes for electronics-based processing units. This is the beginning of an alliance between optics and electronics. It is an exciting time. ❒

John Caulfield is in the physics department at Fisk University, 1000 17th Avenue, North Nashville, Tennessee 37208, USA; Shlomi Dolev is at the Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel.
e-mail: hjc@fisk.edu; dolev@cs.bgu.ac.il

References

1. http://web.mit.edu/newsoffice/2009/optical-computing.html
2. http://www.acreo.se/templates/Page____2215.aspx
3. http://www.avagotech.com/pages/home/
4. http://techresearch.intel.com/articles/Tera-Scale/1419.htm
5. Assefa, S. Xia, F. & Vlasov, Y. *Nature* **464,** 80–84 (2010).
6. http://www.thirdwave.de/3w/tech/optical/EnLight256.pdf
7. Dolev, S., Haist, T. & Oltean, M. (eds) in *Proc. 1st Int. Workshop on Optical Supercomputing* Vol. 5172 (Springer, 2008).
8. Dolev, S. & Oltean, M. (eds) in *Proc. 2nd Int. Workshop on Optical Supercomputing* Vol. 5882 (Springer, 2009).
9. Caulfield, H. J., Dolev, S. & Green, W. M. J. (eds) *J. Opt. Soc. Am. A* **26** (*Optical High-Performance Computing* feature issue) (2009).
10. Caulfield, H. J., Dolev, S. & Green, W. M. J. (eds) *Appl. Opt. A* **48** (*Optical High-Performance Computing* feature issue) (2009).
11. Tamir, D., Shaked, N., Wilson, P. & Dolev, S. *J. Opt. Soc. Am. A* **26,** A11–A20 (2009).
12. Landauer, R. *IBM J. Res. Dev*. **5,** 183–191 (1961).
13. Hardy, J. & Shamir, J. *Opt. Express* **15,** 150–165 (2007).
14. Caulifield, H. J., Soref, R. A, Zavalin, A. & Hardy, J. *Opt. Commun.* **271,** 365–376 (2007).
15. Caulfield, H. J. in *Proc. 2nd Int. Workshop on Optical Supercomputing* Vol. 5882, 30–36 (Springer, 2009).
16. Dolev, S. & Nir, Y. US patent 20050013531 (2005).
17. Dolev, S. & Fitoussi, H. in *Proc. 4th Int. Conf. FUN 2007*, 120–134 (Springer, 2007).
18. Bozhevolnyi, S. I. *Plasmonic Nanoguides and Circuits* (World Scientific, 2008).
19. Brongersma, M. L. *et al.* in *Plasmonic Nanoguides and Circuits* (ed. Bozhevolnya, S. I.) (World Scientific, 2008).
20. Fainman, Y., Ikeda, K. & Tan, D. T. H. in *Proc. 2nd Int. Workshop on Optical Supercomputing* Vol. 5882, 2–4 (Springer, 2009).

Acknowledgements

We thank William Green, Shaya Fainman, Joseph Rosen, Nati Shaked and Hen Fitoussi for helpful inputs.