Comparison of Optical and Electronic 3-Dimensional Circuits

Charles W. Stirk and Demetri Psaltis
Caltech 116-81, Pasadena, CA 91125

1 Abstract

Optical and 3D electronic implementations of a rearrangeably non-blocking permutation network are compared using complexity models. The electronic complexity model is based on the Thompson model for VLSI, while the optical complexity model is new and based on the characteristics of optoelectronic devices and holographic optical elements. The analysis holds the manufacturing cost, size, and BER of the two implementations equal while comparing their latency. Whether optics is better than electronics depends on the message length.

2 Introduction

Under what conditions should free-space optical interconnects be used in a computer instead of electrical wires? The way one chooses to address this question can be as significant as the answer itself. For example, detailed engineering analysis of a single communication link shows that optical interconnects consume less power than electrical wires over large distances and at high speeds. [1, 2] In such situations, and where power consumption is the only concern, one should use optical interconnects. In practice, however, what this type of analysis does not address is just as important. The manufacturing cost, reliability and performance of the system are equally important considerations and must be included in a comparative evaluation of optical versus electrical interconnects.

In this paper we attempt a system-level analysis of the utility of free-space optical interconnects. The analysis uses abstract complexity models which are based on a small set of assumptions about device behavior. These assumptions are general enough to make the complexity analysis simple and the results valid even if the device behavior changes to some extent. The weakness of system complexity analysis is that it only points the way to a solution; the results are valid if the system's size is infinite. It does, however, reveal where an engineering analysis might be fruitful.

The foil we use to compare optics and electronics using complexity theory is a rearrangeably non-blocking permutation network. In other words, it can realize any of the $N!$ permutations of N inputs by rearranging the settings of its switches. The network is circuit-switched, and messages can be pipelined through it. We choose this network because it makes the method of analysis clear, not because we believe it to necessarily be the optimum way to interconnect

parts of a system. On the other hand, this particular network is representative of a family of networks that computer and telecommunication switch designers use extensively. We show that for a network with the same size, cost and reliability, the best technology to implement it depends on the message length.

The following section introduces a complexity model for each technology and shows how each technology can construct a permutation network. Then, we attempt a fair comparison between two dissimilar technologies like optics and electronics. Finally, we highlight the results of the comparison and conclude with a discussion of the salient features and utility of the results.

3 Complexity Models

The comparison framework for optical and electronic interconnects is economic: composed of assumed preferences, costs and a method to measure utility. We focus our attention on only commercially available technologies. For electronics, this includes VLSI, multi-chip modules (MCM), and multilayer printed wire boards (PWB). For optics, the technology is fixed planar holographic optical interconnecting elements and planar arrays of optoelectronic switches which may be integrated with electronic devices.

The costs are those of the user rather than of the system designer: the manufacturing cost, which impacts the purchase price; the size of the system as measured by the number of items we wish to permute (N); the reliability of communication, as measured by the bit error rate (BER); and the performance of the system, as measured by the time it takes for a message to traverse the network (latency).

When measuring the utility of the implementations, we compare electronic and optical systems with the same size, reliability and manufacturing cost in order to derive their relative performance for large N. Optics is a potential alternative to electronics if its performance is better. Whether or not that potential can be realized in practice depends on the constants and lower order terms that we neglect. This is the subject of an engineering analysis which has yet to be attempted.

3.1 Electronic Model

The assumptions of the electronic model are based on the characteristics of systems made by microelectronic lithography. Figure 1 illustrates a cross-section of a wire arrangement in which the intervening volume is filled by an insulator. The limits on delay are primarily electrical, mechanical and process dependent. The manufacturing cost of a circuit is process and system dependent. The numbered assumptions are necessary for the analysis, but the justification which follows them is not. They are only used to motivate the assumptions and are not an inherent part of the model.

E1. *Wire dimensions and separations must be at least some constant distance.*

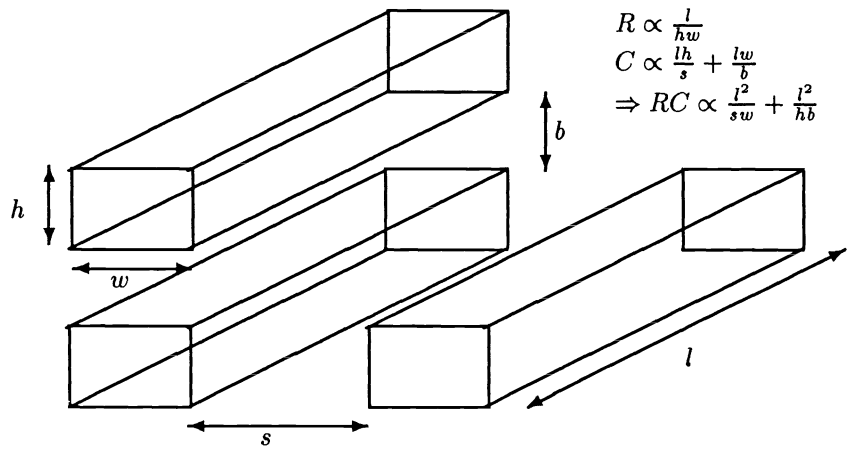

Figure 1: Wire Separation and Dimensions

The lower limit on horizontal separation (s) and dimension (w) comes from the linewidth of the lithographic process. Vertically, the lower limit on separation (b) comes from the desire to maintain planarity, and, thus must be greater than the vertical height (h). The separations also have lower limits that come from signal dependent effects like capacitive and inductive coupling and dielectric breakdown of the insulators, but we will neglect them and their effect on BER. The lower limit on vertical height (h) is due to the thickness of atomic layers. The wire cross-section dimensions also have a signal dependent lower limit from electromigration. The specific values of the limits are not critical for our purposes because they will be lost in the complexity analysis.

E2. *Wire cross section dimensions and wire separation have a constant upper bound.*

The upper limit on wire width (w) and wire height (h) is primarily due to the mismatch between the wire's thermal expansion coefficients and those of the insulator, which could lead to fractures during processing. The upper limit on vertical wire separation is due to the built-in stress of the dielectric; if it gets too thick, the stress could lead to fractures. There is no upper limit on horizontal wire separation other than wafer size. Again the specific values of the limits are not important in this instance, but they must be constants.

E3. *The time it takes to send a signal through a wire is either linear or quadratic in the length of the wire.*

If the size of the wire drivers stays constant as the length of the wires increase, then the active device propagation and switching delays are much smaller than the passive wire propagation delays. In this case, we must measure the wire length to determine delay.

Linear delay arises in superconducting wires or conventional wires if the bit period is much less than the wire's RC

time constant. The linearity in length is due to electromagnetic wave propagation along the wire.

Quadratic delay occurs in conventional conductors when the bit period is greater than the conductor's RC time constant. Note that quadratic delay requires that RC scale quadratically with the length of the wire. This is why we need upper bounds on the wire's width, height and vertical separation (assumption E2); linear delay does not require this assumption.

E4. *The manufacturing cost of a monolithic electronic system is at most* $Ahe^{\lambda Ah}$, *where* A *is the area of each layer,* h *is the number of layers, and* λ *is the defect density.*

Equipment and time spent on making a circuit determines its manufacturing cost. Assume that the manufacturing cost for a single lithographic process step is fixed. Then the cost to produce a circuit is Ah.

Producing a circuit, however, does not guarantee that it works. For a circuit to work, it must have no fatal defects. If λ is the independent and identically distributed probability that a fatal defect occurs in a unit area of wafer, then the probability of getting a working circuit is $e^{-\lambda Ah}$. [3] The manufacturing cost per circuit is inversely proportional to the probability that the circuit works.

3.2 3-Dimensional Electronic Circuits

3-dimensional electronic circuits (3D) have active devices and wires which are distributed throughout the volume. Whereas monolithic 3D-VLSI is not yet a reality, vertically hybrid constructions are. The cost model holds for hybrid constructions if they are not tested before they are assembled.

A non-blocking rearrangeable permutation network for $N = 8$ is shown in figure 2. The figure could represent a 2D layout of the network. The straight lines represent wires, and where they terminate there are 1x2 or 2x1 switches organized into $2lgN - 1$ stages of N switches. There exists a setting of all the switches for each permutation. The settings are calculated and distributed by a central controller.

There is a 3D construction of the network shown in figure 2. [4] The construction has volume $\Theta(N^{\frac{3}{2}})$, where Θ signifies an exact bound. In other words, the upper bound is equal to the lower bound. The maximum length for a single wire in the 3D construction is $O(\sqrt{N})$.

3.3 Single Active Layer Electronic Circuits

Not all 3-dimensional technologies can have active devices which are distributed throughout the volume. Single active layer electronic circuits (SAL) have only one active device layer, usually the bottom, and the wires run throughout the volume. This model represents commercially available technologies like MCM and PWB.

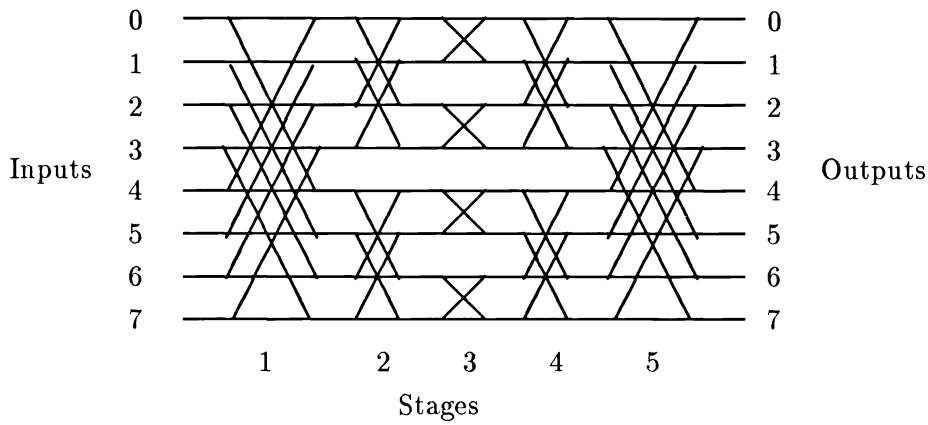

Figure 2: Non-Blocking Rearrangeable Permutation Network

There exists an SAL construction for any N node graph whose nodes have up to two distinct inputs and one distinct output which requires volume $O(N^{\frac{3}{2}})$. [5] A slight modification to the construction allows it to apply to graphs which have nodes with two distinct outputs. The construction is shown in figure 3, where r and s are two nodes connected by an edge.

$O(N)$ area is sufficient for the active plane. Since one layer routes a wire in the x direction and another in y, and since there are at most $O(\sqrt{N})$ nodes that can conflict with this rout, $O(\sqrt{N})$ layers are sufficient to route any degree 2 graph of N nodes.

Using the method of figure 3 for each layer, we can construct a SAL embedding of the rearrangalbe permutation network (shown in figure 4). The volume of this construction is $O(N^{\frac{3}{2}} lg N)$, and the maximum length of a wire is $O(\sqrt{N})$.

3.4 Optical Model

Recall that we prefer optoelectronic planes which are connected through free-space by planar holographic optical elements (HOE) as in figure 5. We assume that the process technology for electronics, lithography, manufactures the optoelectronic planes and the HOE's. The limits on dimensions, separations and delays are primarily optical. Similar to electronics, the manufacturing cost is both process and system dependent.

O1. *The minimum separation of optoelectronic devices is a constant.*

The wavelength of light in a material $(\frac{\lambda}{n})$, where n is the index of refraction, places a lower limit on a resolvable

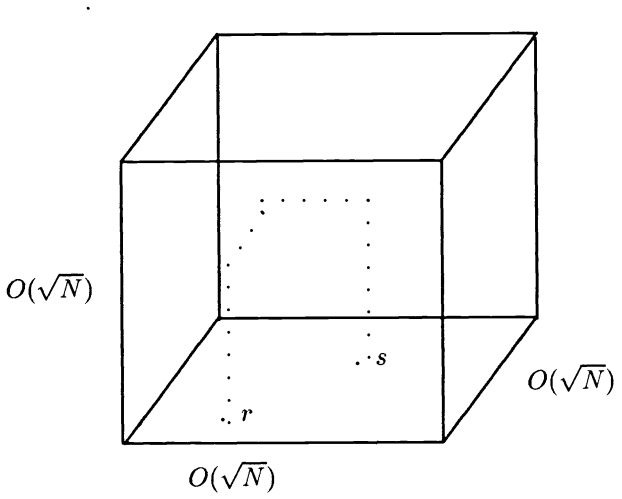

Figure 3: Degree 2 Graph Construction

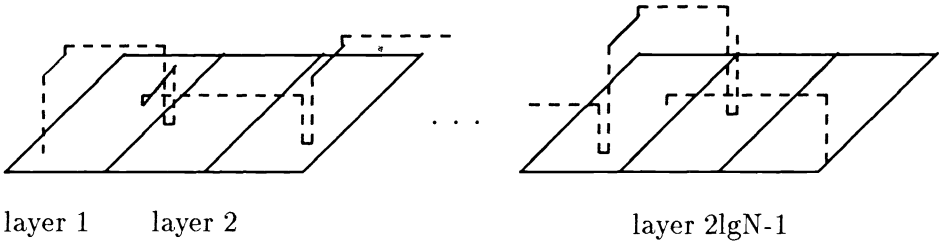

Figure 4: SAL Permutation Network

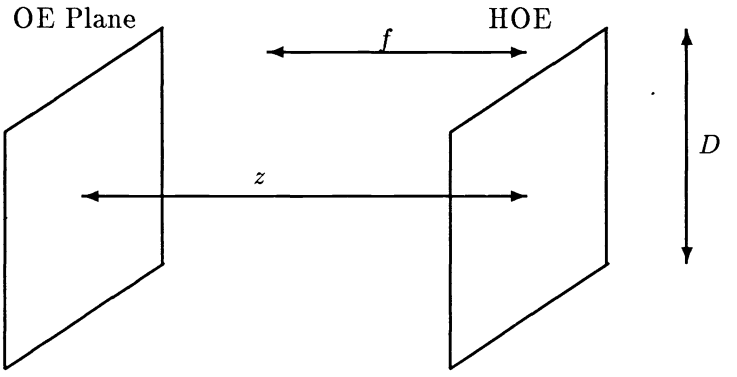

Figure 5: Optical Arrangement

spot's size. Optical detectors and modulators for different channels must be separated by at least this distance for acceptable SNR. Optoelectronic devices can be manufactured closer than this but they cannot be resolved by an optical imaging system.

O2. *Optoelectronic devices can be within a constant factor of the minimum size.*

Many devices have been both manufactured and tested that operate close to the minimum size.

O3. *The numerical aperture of a HOE has a constant lower bound.*

The numerical aperture, or f-number, of an HOE is the focal length divided by the aperture. Typically, aberrations dictate that the numerical aperture of a HOE be greater than 1.

Since the condition for real image formation requires that the image plane to HOE distance, z, must be greater than the focal length, O3 effectively forces the optical axis to scale with the aperture.

O4. *The dimensions of an optoelectronic or holographic plane scales with the dimensions of the previous plane in the optical path.*

Besides the co-scaling of the planes, this assumption has two important effects: scaling of the optical axis with the dimensions of the optoelectronic planes due to O3 and constant SNR. If the number of HOE's and the diffraction efficiency for each channel is constant in an optical path through the imaging system, then O4 implies that the optical efficiency of the system has a constant lower bound. Neglecting cross-talk due to scattering and side-lobes, this ensures that the SNR will be independent of N. Note that O4 is not physically necessary but will be satisfied by the system described in the next section.

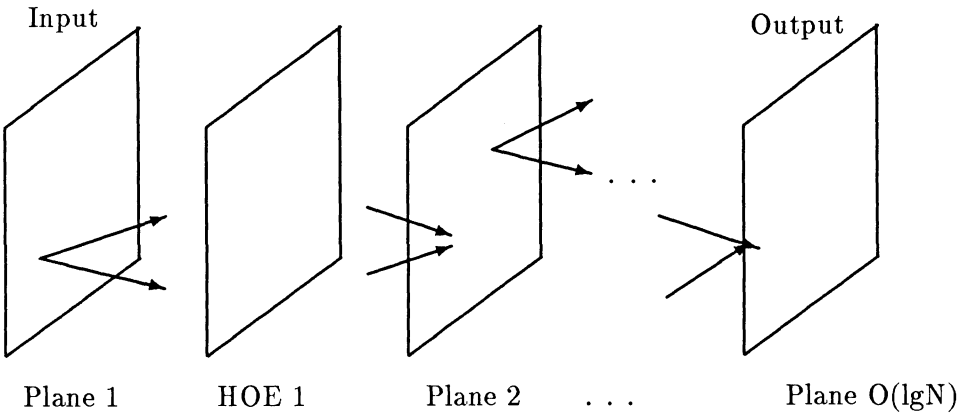

Figure 6: Optical Permutation Network

O5. *The manufacturing cost of an optoelectronic device plane with area A is* $Ae^{\lambda A}$.

Because the same process makes both optoelectronic and microelectronic devices, they have the same manufacturing cost.

O6. *The manufacturing cost of an HOE with aperture D is* D^2.

Lithographic methods also form the basis of manufacturing HOE's. Assume that we use holograms with a constant number of phase levels. Point defects introduced during processing, unlike those for microelectronics, do not destroy the utility of the HOE; they just lower the SNR. In the following analysis, you will note that this assumption can be tightened to the form of O5 without changing the results. We mention it only as an illustration of the different effect of point defects on HOE's and active devices.

3.5 Optical Permutation Network

There is a permutation network composed of $O(lgN)$ stages of $\frac{N}{2}$ 2x2 crossbar switches in which successive stages are connected by perfect shuffles. [6] If we allow each stage of switches to lie on separate optoelectronic device planes, we can use the folded perfect shuffle [7] with the lenses replaced by HOE's for the interstage connections (as in figure 6).

Since each plane can be built and tested separately, the manufacturing costs add, rather than multiply in the assembled structure. The separation between planes grows as $O(\sqrt{N})$.

Implementation	Cost	Modified Cost
3D	$N^{\frac{3}{2}}e^{\lambda N^{\frac{3}{2}}}$	$(Nw_{3D})^{\frac{3}{2}}e^{\lambda(Nw_{3D})^{\frac{3}{2}}}$
SAL	$N^{\frac{3}{2}}lgNe^{\lambda N^{\frac{3}{2}}lgN}$	$(Nw_{SAL})^{\frac{3}{2}}lgNe^{\lambda(Nw_{SAL})^{\frac{3}{2}}lgN}$
Optics	$NlgNe^{\lambda N}$	$Nw_OlgNe^{\lambda Nw_O}$

Table 1: Cost and Modified Cost of Implementations

4 Comparison Methodology

How can we compare the performance of the network's optical and electronic implementations while holding size, reliability and manufacturing cost constant? The reliability question is simple if we limit ourselves, as we have in the previous section, to optical systems with a constant number of passive optical elements per stage, each with a constant diffraction efficiency. Because the manufacturing cost depends on the size, the hard part is to get them to be equal for both implementations.

One solution to this problem is to let the width of a channel vary. [8] What we mean by variance is to replace every wire with w_e wires and every optical channel with w_o parallel optical channels. Clearly, the changes affect the cost of each implementation as shown in table 1 (all table entries from this point on are in big-O notation). The size of the network remains the same, however, and by setting the costs equal to one another, we can derive an expression that relates w_o to w_e.

Now, we must measure the performance, or latency. If our message length is L bits, and if we can only send w bits through each set of channels per unit time, then the latency is

$$\text{latency} = T_d\left(\frac{L}{w} + D\right) \tag{1}$$

where T_d is the delay and D is the diameter of the network: the number of wire traversals through the network (lgN). The delay is the elapsed time when information propagats to the next stage. Because we consider only synchronous systems, the propagation delay is the same for all wires and determined by the delay model and the maximum wire length. For optics the propagation delay is determined by the length of the optical axis divided by the speed of light. The delays are in table 2.

Implementation	Linear Delay	Quadratic Delay
3D	$(Nw_{3D})^{\frac{1}{2}}$	Nw_{3D}
SAL	$(Nw_{SAL})^{\frac{1}{2}}$	Nw_{SAL}
Optics	$(Nw_O)^{\frac{1}{2}}$	—

Table 2: Time Delay per Stage

Now, we have all the pieces together: set the costs equal and derive the relationship between w_o and w_e; then plug w_e and the cycle time into equation 1. There are four different cases based on which term of equation 1 dominates for electronics and optics.

First, let's consider optics and 3D electronics. In the low yield limit, which is typical of immature technologies, the cost equation is dominated by the exponential term. In this case the relationship between the channel widths increases with N (as shown in equation 2).

$$w_o = O(w_{3D}^{\frac{3}{2}}\sqrt{N}) \tag{2}$$

Table 3 lists the ratio of the optical to electronic latency for both linear and quadratic delay. One set of conditions on the relative magnitude of equation 1's terms is inconsistent.

When the latency ratio is less than 1, optics has an advantage over electronics. From table 3 we see that 3D is best when the delay is linear and the message length is short to moderate—in this case optics cannot use its advantages of low delay or wide channels. Furthermore, optics has lower latency in all other conditions except for sometimes winning the tradeoff of w and N at quadratic delay and moderate message length.

In the high yield limit the linear term in the manufacturing cost equation dominates. In this case the relationship between the channel widths decreases with N (equation 3).

Condition	Linear	Quadratic
$L << w_{3D} lgN$	$w_{3D}^{\frac{1}{4}}$	$\frac{1}{w_{3D}^{\frac{1}{4}}\sqrt{N}}$
$L >> w_{3D} lgN$ $L << w_{3D}^{\frac{3}{2}} N^{\frac{1}{4}} lgN$	$w_{3D}^{\frac{5}{4}} lgN$	$\frac{w_{3D}^{\frac{3}{4}} lgN}{\sqrt{N}}$
$L >> w_{3D}^{\frac{3}{2}} N^{\frac{1}{4}} lgN$	$\frac{1}{(Nw_{3D})^{\frac{1}{4}}}$	$\frac{1}{(Nw_{3D})^{\frac{3}{4}}}$

Table 3: Ratio of Optical to 3D Latency: Low Yield

$$w_o = O\left(\frac{w_{3D}^{\frac{3}{2}}}{lgN}\right) \tag{3}$$

Table 4 lists latency ratio for 3D vs optics at high yield. In this case electronics only has an outright latency advantage for linear delay and message lengths that are large with respect to N but short with respect to w. Optics either has a tradeoff of w for N or an outright latency advantage in all the other cases.

In the high yield limit for optics versus SAL, the relation between channel widths increases with N as shown by equation 4.

$$w_o = O\left(w_{SAL}^{\frac{3}{2}} N^{\frac{1}{2}}\right) \tag{4}$$

Table 5 lists the latency ratio for SAL vs optics at high yield. As we found for 3D and low yield, SAL has lower latency when delay is linear and message length is short because optics cannot use its advantages of wide channels and linear delay. Optics is superior or has a tradeoff in all other cases.

In the low yield limit for optics vs SAL, the relation between channel widths is

$$w_o = O(w_{SAL}^{\frac{3}{2}} N^{\frac{1}{2}} lgN) \tag{5}$$

which also increases with N. Table 6 lists the latency ratio for SAL vs optics at low yield. In this case optics has a tradeoff at short message lengths and quadratic delay and for long message lengths. Electronics has better latency

Condition	Linear	Quadratic
$L << w_{3D}^{\frac{3}{2}}$ $L << w_{3D} lgN$	$\frac{w_{3D}^{\frac{1}{4}}}{lg^{\frac{1}{2}}N}$	$\frac{1}{\sqrt{w_{3D}^{\frac{1}{2}} N lgN}}$
$L >> w_{3D} lgN$ $L << w_{3D}^{\frac{3}{2}}$	$w_{3D}^{\frac{5}{4}} lg^{\frac{1}{2}}N$	$\frac{1}{w_{3D}^{\frac{7}{2}} N lgN}$
$L >> w_{3D}^{\frac{3}{2}}$ $L << w_{3D} lgN$	$\frac{1}{w_{3D}^{\frac{5}{4}} lg^{\frac{1}{2}}N}$	$\sqrt{\frac{w_{3D}^{\frac{3}{2}} lgN}{N}}$
$L >> w_{3D}^{\frac{3}{2}}$ $L >> w_{3D} lgN$	$\frac{lg^{\frac{1}{2}}N}{w_{3D}^{\frac{1}{4}}}$	$\sqrt{\frac{lgN}{Nw_{3D}^{\frac{3}{2}}}}$

Table 4: Ratio of Optical to 3D Latency: High Yield

Condition	Linear	Quadratic
$L << \sqrt{Nw_{SAL}^3} lgN$	$(Nw_{SAL})^{\frac{1}{4}}$	$\frac{1}{(Nw_{SAL})^{\frac{1}{4}}}$
$L << \sqrt{Nw_{SAL}^3} lgN$ $L >> w_{SAL} lgN$	$N^{\frac{1}{4}} w_{SAL}^{\frac{5}{4}} lgN$	$\frac{w_{SAL}^{\frac{3}{4}} lgN}{N^{\frac{1}{4}}}$
$L >> \sqrt{Nw_{SAL}^3} lgN$	$\frac{1}{(Nw_{SAL})^{\frac{1}{4}}}$	$\frac{1}{(Nw_{SAL})^{\frac{3}{4}}}$

Table 5: Ratio of Optical to SAL Latency: High Yield

Condition	Linear	Quadratic
$L << N^{\frac{3}{4}} w_{SAL}^{\frac{3}{2}} lgN$	$w_{SAL}^{\frac{1}{4}} \sqrt{NlgN}$	$\frac{lg^{\frac{1}{2}} N}{w_{SAL}^{\frac{1}{4}}}$
$L << N^{\frac{3}{4}} w_{SAL}^{\frac{3}{2}} lgN$ $L >> w_{SAL} lgN$	$N^{\frac{1}{4}} w_{SAL}^{\frac{5}{4}} lg^{\frac{3}{2}} N$	$w_{SAL}^{\frac{3}{4}} lg^{\frac{3}{2}} N$
$L >> N^{\frac{3}{4}} w_{SAL}^{\frac{3}{2}} lgN$	$\frac{lg^{\frac{1}{2}} N}{w_{SAL}^{\frac{1}{4}} N^{\frac{1}{4}}}$	$\frac{lg^{\frac{1}{2}} N}{(N w_{SAL})^{\frac{3}{4}}}$

Table 6: Ratio of Optical to SAL Latency: Low Yield

for all other cases.

5 Discussion

The answer to the question, "Under what conditions is optics better than electronics for interconnects?" depends on factors like the particular delay or cost models. The main advantage of optics is that delay is linear and air is free; the volume of the system is important only as it affects the delay and the area. Another interesting, although not necessary, feature of the optical cost model is that holograms degrade gracefully with point defects.

For the rearrangeable non-blocking permutation network which we considered in this paper, the relative magnitude of the message length (L), network size (N), manufacturing cost (w) and process maturity determine whether optics or electronics has better latency. For large systems, optics tends to have an advantage when it has larger channel widths or lower delay than the electronic system.

We chose both linear and quadratic delay models for electronics and the VLSI yield cost model. It is important to note, however, that not all electronic and optical technologies satisfy the assumptions of our models. In particular, we did not consider redundant or fault tolerant techniques in the cost model. Moreover, we did not consider cost models for constructions that use more than one type of packaging. And we did not consider cross-talk, electromigration or switching device failure in either the cost or reliability models.

An implicit assumption in equation 1 is that the inverse of the bandwidth is equal to the delay. This is typical of quadratic delay in balanced parallel computer architectures but is by no means necessary with the linear model. As

they are in optics, bandwidth and delay are independent variables within the electronic linear model.

Another limitation is the dominance of wire delay. Propagation delay dominates switching delay only when the the line lengths, as a function of the size of the system, grow asymptotically. In practice, propagation delay dominates only for very long interconnects. For other constant dependent effects, a detailed engineering analysis may be necessary.

Finally, the model did not include power dissipation. As the bandwidth increases, so does power dissipation. It also increases in the quadratic model with the length of the wire. In order to be more realistic, we would have to include power dissipation as a cost in both the electronic and optical models.

All of these caveats and many other effects can be included in more complicated models of optical and electronic systems than the ones we have considered here. The underlying philosophy of this work is still valid, however, that only the externally observable characteristics of a system, like manufacturing cost and reliability, are significant when measuring performance. The internal characteristics, like the model assumptions, are better left to designers for circumvention or improvement.

This work is supported by funding from AFOSR and DARPA. One of the authors, C. W. Stirk, would like to thank TRW for a graduate fellowship.

References

[1] R. Kostuk, J. W. Goodman and L. Hesselink, Optical imaging applied to microelectronic chip–to–chip interconnections, *Appl. Opt.* 24(17), 2851–2858 (1985).

[2] M. R. Feldman, S. C. Esener, C. C. Guest and S. H. Lee, A Comparison of Optical and Electrical Interconnects based on Power and Speed Considerations, *Appl. Opt.* 27, 1742–1751 (1988).

[3] W. J. Bertram, Yield and Reliability, in *VLSI Technology*, S. M. Sze ed., McGraw Hill, New York, 599–637 (1983).

[4] A. L. Rosenberg, Three-dimensional integrated circuitry, in *VLSI Systems and Computations, H. T¿ Kung, b. Sproull and G. Steele, eds.*, Computer Science Press, Rockville, 60–80 (1981).

[5] A. N. Kolmogorov and Ya. M. Bardzin, On Realiztion of Sets in 3-dimensional Space, Problems in Cybernetics, 261–268 (March 1967).

[6] C. L. Wu and T. Feng, Universality of the Perfect Shuffle Interconnection Network, *IEEE Trans. Comput.* C–30, No. 5, 324–332 (1981).

[7] C. W. Stirk, R. A. Athale and M. W. Haney, Folded Perfect Shuffle Optical Processor, *Appl. Opt.* 27(2), 202–203 (1988).

[8] W. J. Dally, *A VLSI Architecture for Concurrent Data Structures*, Ph.D. Thesis, Caltech (1986).