Design and Analysis of a Scalable 3-Dimensional Multicomputer Architecture Using Optical Interconnection for PetaFLOP Computing

Ekpe Okorafor & Mi Lu
Department of Electrical Engineering
Texas A&M University
College Station, TX 77843
{ekpe,mlu}@ee.tamu.edu

Keywords: PetaFLOP, Optical interconnection network, Parallel processing, Multicomputer, Scalable

Abstract
In this paper, we present an initial design and analyses of an extremely scalable multicomputer architecture that has the potential to deliver in the order of petaFLOP performance. Our design takes advantage of free-space optical technologies, harnessing the features inherent in optics, to produce a 3D stack that implements efficiently a large, fully connected system of nodes forming a true 3D mesh. Each node is a complete computer system with both compute and storage units, and six communication interfaces to the optical medium. This packaging greatly improves density and communication performance. In future work, the system will be built to allow failed nodes stay in place without appreciable performance degradation. We investigate in detail the case study for 100 teraFLOP performance and describe the characteristics of the proposed hardware components in our case study.

1 INTRODUCTION
The current trend in multi-computer network design is to pack nodes more densely in such a manner as to efficiently distribute computing resources and interconnect uniformly in three-dimensional space. This has led to a remarkable improvement in communication performance, scalability and density. Ultimately, the demand for ever greater performance by many computation problems pushes the boundaries for the development of such large-scale supercomputers. PetaFLOP (10^{15} floating-point operations per second) performance are required by many applications and they include real-time image processing, artificial intelligence, real-time processing of databases, weather modeling, simulation of neural networks, simulation of physical and biological phenomena, etc.

The functions of such large-scale petaFLOP-performance computer architecture will include data acquisition and transmission, data processing, data management and storage. These functions necessitate a large-scale, cost-effective computing and storage capability to handle the extensive requirements for simulations and analysis of massive amounts of data. They rely on advanced, emerging information technologies to create combinations of hardware and software which will achieve unprecedented increases in numerical processing through parallel computation.

On the other hand, the issue of scalability of such large-scale petaFLOP-performance computer cannot be over-emphasized. Massively parallel systems are required to scale in the sense that their performance should be proportional to the number of nodes. Unfortunately, unlimited scalability is not theoretically possible and worse still even harder to achieve practically beyond some order of magnitude in the number of nodes. The performance objectives of supercomputers is hindered because of the difficulties associated with developing low complexity, high-bisection bandwidth, and low-latency interconnection networks to connect thousands of nodes while still keeping the system scalable. Many interconnection networks have been proposed for the design of massively parallel computers, including hypercubes [1], meshes and tori [2]. Others include fat trees and enhanced meshes. Amongst these, the hypercube has been researched more intensively because of its good topological properties and high interconnectivity. The difficulty posed by the extremely high VLSI complexity incurred, due to very high communication channels needed to implement these interconnection networks, has continued to hinder the use of these topologies to achieve large-scale computer systems. The high VLSI complexity problem is obviously unbearable for any scalability.

It is desirable to have low-dimensional massively parallel computers with full-connectivity in each direction. It is also desirable to make use of a topology that has an extremely small diameter and average inter-node distance, and a large bisection width. The utmost flexibility in exploiting parallelism is afforded by a topology with diameter equal to one, where each processor can directly communicate with any other processor. The most useful properties of a parallel processor interconnection network are high bandwidth (scaling directly with the number of processors), low latency, no arbitration delay, and non-blocking communication. It is apparent that the electronic implementation of such a large-scale system is very difficult.
According to [3], metal interconnects have reached their physical limits and have become a limiting factor because of power, delays and density considerations. The idea of optical interconnection of very large-scale integration (VLSI) electronic was proposed and analyzed in [4]. This notion was the start of the field of optical interconnects. Many advances have been made in the field of optical interconnects to date. Engineering analysis has showed specific energy dissipation benefits of optical interconnects. The physical limits and have become a limiting factor because of electrical line is scaled down on all three dimensions, its resistive-capacitive time constant does not change. This is an undesirable quality, since the wires do not scale to keep up with the transistors. Optical interconnects avoid this problem altogether because they do not have the resistive loss physics that gives rise to this phenomenon. In recent years, extremely fast photonic networks are being developed that have the potential to support very large bandwidth interconnections, with an extraordinarily quick response time and very low latency.

Significant progress both at the device and sub-system levels has been made in Free-Space Optical Interconnects (FSOI) to the point where FSOI can now be considered to push the envelope in computing hardware at the board to board interconnect level [7]. Opto-Electronic (OE) devices including VCSELs, light modulators, and detectors have now been developed to the point that they can enable high speed and high density FSOI [8-10]. We also note here that recent attempts to connect boxes or computer systems with FSOI links have proven practical [11]. System boards usually run at some fraction of the processor clock, usually about half. In the next few years, we would expect the off-board communication to approach 10 GHz. Signals have to be routed at 10 GHz over a small distance at 2.5 or 1.8 V cycles. Cross-talk and reflections on electrical lines have been identified as major problems. It is well known that VCSEL links can provide the interconnection bandwidth thereby replacing the current large edge connectors. This will improve system noise margins because cross-talk and ground noise coupling become more difficult to control in traditional connectors as edge rates increase. Sophisticated CAD tools for free-space optical systems are already in development [12].

From the foregoing, we propose an interconnection network that utilizes free-space optical technology because of its increased connectivity and reduced packaging complexity. We set about formulating a relevant and attainable objective, and present a viable solution to fulfill this objective. To this end, we will present an innovative 3D interconnection network design with feasibility analysis and performance characteristics of our design.

The rest of this paper is organized as follows. In Section 2, we present the basic architecture of our design. Section 3 contains a detailed description of our design for a system capable of teraFLOPS. This section also includes overall performance characteristics of this system and the feasibility analysis. Section 4 presents analyses for the optical interconnection network. Finally, we conclude in Section 5.

2 BASIC STRUCTURE

Our architecture encompasses a 3D optical interconnection network. In Section 2.1, we present the structure of the basic building block and the issues relating to the implementation.

2.1 Basic building block

Our basic design takes advantage of free-space optics technology to produce a fully connected scalable node unit. In our design, we set out with the following objectives in mind: high density and low packaging complexity, reliable, low-cost yet powerful, and above all a robust interconnection network. A root cause for the staggering difficulty of managing large systems is the proliferation of too many building blocks such as processors, disk arrays, switches, communication protocols, etc. [13]. Accordingly, this has lead to combinatorial explosion of complexity. In our design, this explosion is controlled by encapsulating complexity within the basic building blocks or nodes. These nodes have well defined hardware and software interfaces. The basic shape of our node is a cube with six sides.

Figure 1. Diagram showing basic node structure with FSOI couplers capable of gigahertz communication

Figure 1 shows the basic node structure in our design. All nodes have six sides as in a cube structure. On each side is an optical coupler capable of gigahertz communication. Internally, each node consists of 8 multiprocessor units coupled to the optical highway as shown in Figure 2.

Figure 2. Diagram showing logical interconnection within each node

Each node consists of 8 CMOS chips with optical ports as the only means of external high-speed data
communication. Each processor unit in the node has its own local memory. Each processor unit is interfaced with optical transmitter/receiver modules and attached wave guide for inter-processor data transfer.

Figure 3. Schematic diagram of each PE

Figure 3 is a schematic of our design concept of each processor. Each PE includes two CPUs with L1 and L2 cache connected by a high-speed multiport optical switch. The switch connects to an on-chip shared L3 cache and multiple high-speed optical ports. The thermal management of the electrical CMOS and optical ports are separated. The optical port interfaces decode and multiplex signals for all-optical routing. The interface consists of low-power VCSEL transmitters, photodetector receivers and the optical interface. Each optical port is capable of sustaining 40 GB/s (320 Gb/s) data throughput in each direction. Each chip has 8 of such optical ports. One is dedicated to local main memory, another for inter-chip communication and the remaining six are for external IO.

The scalability of the system is dependent on there being enough bandwidth available for the network. The bisection bandwidth of the 8-port optical switch in each chip is 640 GB/s (5.12 Tb/s).

The result of our design is a set of high-performance encapsulated processors serviced by high-bandwidth optic interconnects that form the basic building block of our 3D system.

2.1 3D system

As stated earlier, each node is made up of 8 PEs all connected via guided, planar optical interconnect. The nodes need to communicate with each other and also with the external world. The communication is realized by a network which links all the nodes into a true 3D mesh, shown in Figure 4. This physical architecture leads to very high system density. An important innovation is the elimination of cables and connectors, and instead substituted with free-space optical couplers. This undoubtedly leads to remarkably improved communication hardware cost/performance (magnitudes of about 100 Gbps per interface) compared to conventional, centralized switch solutions. This architecture is able to scale extensively while delivering a large amount of bandwidth. We emphasize that the quest for teraFLOP computing begins by solving the scalability and bandwidth issues.

Two very important considerations in a design of this nature are power and cooling, however, these will not be discussed in details in this paper, as it is beyond the scope.

Figure 4. 3D 4x4x4 mesh interconnection network using optical interconnect

Recall that each processor unit has 6 optical ports for external communication. Each of these ports is capable of sustaining 40 GB/s (320 Gb/s) data throughput in each direction. Each of these ports is also optically coupled to the specified node interface. The free-space optical coupler attached to each face of the node is capable of sustaining 100 GB/s (800 Gb/s) data throughput in each direction. If we assume the communication frequency \( f \) for each PE is about 1GHz (2 Gb/s). The guided planar optical interconnect should be able to sustain 16 Gb/s data throughput in each direction. Similarly, each node optical interface has to sustain this data rate. Obviously, this is quite lower than the capacity of the interface and indeed the guided planar optical interconnect. The number of processors in a node can be increased up to the bandwidth capacity of the inter-processor link. However, due to power and thermal considerations, there will be a limit to how many processors can be packed in a certain volume of space.

3 100 TERAFLOP PERFORMANCE

Innovative circuit design using 0.1 µm CMOS technology have produced clock speeds in GHz. The resulting peak performance of a single processor is about 10 gigaFLOPS. Thus for 100 teraFLOPS performance, we need approximately 10,000 processors. Each node in our design with multiple processors is capable of peak performance \( px 10 \) gigaFLOPS, where \( p = 8 \), we have 80 gigaFLOPS. An \( n x n x n \) 3D mesh has a total of \( 8n^3 \) processors. Since each processor is capable of 10 gigaFLOPS, to achieve 100 teraFLOPS or more we need at least 10,000 processors. This gives a value of \( n \geq 11 \). In the next few subsections, we take
a look at the feasibility analysis for the optical components in our design.

3.1 Feasibility analysis for optical components

Our analysis here is for 3D mesh architecture of consisting of 1331 (11 x 11 x 11) nodes. This gives a performance of roughly 106.5 teraFLOPS. The optical link we analyze is integrated in a bi-directional free-space interconnect between two adjacent faces of two nodes, separated by a distance ranging from 0 to 25 cm. This system is able to sustain a ±1-mm lateral misalignment, and a ±1° angular misalignment between the adjacent faces. The system uses arrays VCSELs and photodetectors (PDs). We consider the optical coupler interface, transmitters/receivers, power consumption, and the optical planer wave guide.

3.1.1 Free space optical coupler interface

As stated before each FSOI is capable of sustaining 100 GB/s. FSOI can provide high bandwidth with no physical contact, however it suffers from poor tolerance to misalignment. Therefore, a key implementation objective is to use an active alignment scheme in conjunction with an optimized optical design. The optical link is implemented using both passive and active alignment techniques. The system is aligned mechanically under no lateral misalignment. When misalignment is introduced, redundancy is used to guarantee proper optical performance. A schematic of the optical link is shown in Fig. 5.

![Figure 5](image)

Figure 5. Optical link assemble with built-in redundancy

3.1.2 Optical transmitters/receivers

The optical system provides a maximum power coupling efficiency between a 3 x 3 array of single-mode 960-nm 3-µm diameter VCSELs with 250-µm pitch, and 3 x 3 array of 70-µm diameter PDs with a 125-µm, under any degree of lateral or angular misalignment within the specified limits. Each VCSEL in the array emits -2.22 dBm of optical power. In [14], the performance of single- and multimode VCSELs intended for high capacity free space optical interconnects at 10 Gb/s is presented. The receiver sensibility at about 2 Gb/s is results in a requirement of at least -25 dBm of optical power. The optical link system for the transmitter consists of a planar microlens (PML) array to collimate the VCSELs and a macro lens to relay beams. The receiver part of the link uses only macroptics.

3.1.3 Power

As already mentioned each VCSEL in the array emits -2.22 dBm of optical power. Each node transmits information to each of the 1330 other nodes in the system via approximately 48(6x8) dedicated VCSELs, and radiates, on the average, 1330x48x2.22 dBm = 14W of optical power.

3.1.4 Guided planar optical interconnect

Plastic optical fibers (POFs) are used as the optical pathways within a node. POFs are preferred over glass fibers because of their lower cost, their smaller bending radius and their large numerical aperture (NA). The pitch of the POFs is designed to meet that of the active devices (250-µm). These optical pathways have been fabricated using Toray's PGR-FB125 fiber. In [15], an interconnect demonstrator using multimode POF fiber ribbon is presented. The fibers are butt-coupled to the VCSELs and detectors. The light from each processor is coupled into the POF.

To connect all the PEs, an 8 x 8 POF arrays (pitched at 250 µm in the two dimensions) is developed. The optical pathways for connecting the different PEs have been fabricated. They consist of two arrays of 8 x 8 POF ribbons. The optical pathway uses an approach where 1-D arrays of POF-fiber plates are stacked, which makes it easier to manufacture.

4 ANALYSES OF THE OPTICAL INTERCONNECTION NETWORK

In this section we present some results of the analysis, simulation, and feasibility study for the optical interconnection network of our design.

4.1 Modeling the free-space optical interconnect

The minimum lens diameter for each interconnection length $L$ will be determined by the diffraction of the VCSEL beam which has a waist $w_0$ and a divergence $\theta$. From the minimum lens diameter we can then calculate the maximum channel density as a function of the distance traveled in the optical path, assuming that the pitch of the channels equals the lens diameter.

4.1.1 Cross-talk and transmission efficiency

If we assume that in the middle of the optical path (at $z = 0$) the beam waist is $w'(0) = w_0$ then the beam radius at the lens ($z = L_{\text{max}}/2 = L$) is given $w'(L) = w_0 \sqrt{1 + \frac{\lambda^2 L^2}{\pi^2 w_0^4}}$.

If we now apply the rule that the laser beam must always be smaller than 2/3 of the lens diameter so that more than 99% optical energy throughput through the lenses is achieved and cross-talk is absent in the system, we also have
that \( w'(L) = \frac{2}{3} \frac{\phi_{\text{lens}}}{2} \) and \( L = \frac{\pi}{\lambda} \sqrt{\frac{w_0^2 \phi_{\text{lens}}^2}{9} - w_0^4} \). We now calculate the beam waist \( w_0 \) such that the optical interconnection distance \( L \) is at its maximum. For \( \frac{dL}{dw_0} = 0 \) we find \( w_0 = \frac{\phi_{\text{lens}}}{3 \sqrt{2}} \) and \( L_{\text{max}} = 2L = \frac{\pi \phi_{\text{lens}}^2}{\lambda} \frac{9}{9} \).

So due to diffraction of the laser beam the minimum lens diameter \( \theta_{\text{lens}} \) for an interconnection length \( L \) is limited to

\[
\phi_{\text{lens}} = \frac{3}{\pi} L_{\text{max}}
\]

(1)

### 4.1.2 Bit error rate

The bit error rate (BER) indicates the required source power and signal-to-noise levels necessary to achieve the desired signal fidelity, and represents an important measure of system performance. With Gaussian statistics we find that the probability of error (POE) is given by

\[
\text{POE} \equiv \frac{1}{(2\pi Q)^{1/2}} e^{(-Q^2/2)}
\]

(2)

where \( Q \) is a normalized number that qualifies the quantity of the current signal. We calculate that in order to achieve a BER of \( 10^{-17} \), we need \( Q = 8.5 \).

### 4.1.3 Simulation results and discussion

In Figure 6, we calculate the allowed channel density as a function of the interconnection length in our design using an 8 x 2 array of VCSELs. Using the earlier derived expression (1), the minimum lens diameter to achieve at least 99% transmission efficiency while avoiding cross-talk is approximately 130 \( \mu \text{m} \) with wavelength (\( \lambda = 960 \text{ nm} \)).

We also calculated the transmission efficiency of the optical interconnection system (the ratio between the powers of the emitted light and the light impinging on the detector area) for different focal numbers of the lens and for different working distances between the sides of two adjacent communicating nodes and the results are shown in Figure 7.

![Figure 7. Transmission efficiency of the free-space of lens with different focal numbers](image)

The angular tilt of the optical beam presents a major constraint in our design. Proper alignment of the optical system is of utmost importance. A value of \( 0.1^\circ \) optimizes the system, whereas a larger angular tilt will require an increase in the lens radius, and other system dimensions as a consequence. It is also not quite practical to compensate for the tilt by increasing the laser power due to the exponential nature of the curve. In figure 8, we show the BER for a laser power of 5mW as a function of the angular tilt, while the BER as a function of laser power is shown in Fig 9. Some of the optimized parameters in our design are shown in table 2.

![Figure 8. The plot of BER as a function of the angular tilt](image)

![Figure 9. The plot of BER as a function of the laser power](image)
Table 1. Optimized parameters

<table>
<thead>
<tr>
<th>Maximum efficiency (%)</th>
<th>98.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal number</td>
<td>2.9</td>
</tr>
<tr>
<td>Propagation distance (µm)</td>
<td>1500</td>
</tr>
<tr>
<td>Alignment tolerance (degrees)</td>
<td>0.1</td>
</tr>
<tr>
<td>Reflective power loss (dB/cm0)</td>
<td>0.25</td>
</tr>
<tr>
<td>Wavelength (µm)</td>
<td>0.96</td>
</tr>
<tr>
<td>Detector diameter (µm)</td>
<td>130</td>
</tr>
<tr>
<td>Q parameter of receiver</td>
<td>8.5</td>
</tr>
<tr>
<td>RMS current noise by receiver (nA)</td>
<td>789.6</td>
</tr>
</tbody>
</table>

Table 2. Characteristics of sources/detector

<table>
<thead>
<tr>
<th>Substrate thickness (µm)</th>
<th>VCSEL</th>
<th>POF</th>
<th>Detector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter (µm)</td>
<td>d&lt;sub&gt;source&lt;/sub&gt; = 7</td>
<td>120</td>
<td>75</td>
</tr>
<tr>
<td>NA</td>
<td>θ&lt;sub&gt;FWHM&lt;/sub&gt; = 12°</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>Pitch (µm)</td>
<td>250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Working Distance</td>
<td>L</td>
<td>L</td>
<td></td>
</tr>
</tbody>
</table>

4.2 Modeling the POF guided planar optical interconnect

For modeling purposes, we have divided the interconnection block schematically into two main parts: the emitter and the receiver side. This allows us to investigate the two optical sub-systems on efficiency, cross-talk and tolerances individually. In table 1 we have listed the parameters that affect the cross-talk and the efficiency. The characteristics of the VCSEL sources and of the InP photodetectors can also be found in table 2. We consider here small diameter POFs with a core diameter of 120 µm and a device pitch of 250 µm.

4.2.1 Cross-talk and transmission efficiency

We have derived an analytic expression for the maximum working distance L<sub>max</sub> from the emitter or receiver to the POF, below which no cross-talk between neighboring fibers will occur. L<sub>max</sub> is given by equation (3) at the emitter side and by equation (4) at the receiver side. Here θ represents the divergence angle θ<sub>FWHM</sub> of the micro-emitters as long as θ<sub>FWHM</sub> is smaller than the acceptance angle of the POF. If the latter condition is not satisfied θ takes the value of the acceptance angle θ<sub>POF</sub> of the POF:

\[
L_{\text{max}} = \frac{P - d'_{\text{source}}}{2 - D/2} \frac{\tan \theta}{\tan \theta_{\text{POF}}} \quad \text{where}
\]

\[
d'_{\text{source}} = d_{\text{source}} + 2T \left( \tan \theta_{\text{FWHM}} \sin \frac{\theta_{\text{FWHM}}}{\eta_{\text{GaAs}}} \right) \quad (3)
\]

\[
L_{\text{max}} = \frac{P - d'_{\text{source}}}{2 - D/2} \frac{\tan \theta}{\tan \theta_{\text{POF}}} \quad \text{where}
\]

\[
d'_{\text{source}} = d_{\text{source}} + 2T \left( \tan \theta_{\text{FWHM}} \sin \frac{\theta_{\text{FWHM}}}{\eta_{\text{GaAs}}} \right) \quad (4)
\]

where, P = pitch of the devices, d<sub>det</sub>, d<sub>source</sub> = diameter of the active area of the detector and source, D = diameter of POF, NA = numerical aperture of POF, θ<sub>POF</sub> = acceptance angle of the POF, T = substrate thickness, η<sub>GaAs</sub>, η<sub>InP</sub> = index of refraction, θ<sub>FWHM</sub> = FWHM angle of the emitter and

\[
\theta = \theta_{\text{POF}} \quad \text{if} \quad \theta_{\text{POF}} > \theta_{\text{FWHM}}
\]

\[
\theta = \theta_{\text{FWHM}} \quad \text{if} \quad \theta_{\text{POF}} < \theta_{\text{FWHM}}
\]

To study the transmission efficiency of the POF-based interconnect as a function of the working distance L we simulated both the emitter and detector module via ray tracing and radiometric calculations. When simulating the emitter side we modeled the VCSEL with a user-defined source featuring a circular geometry with a uniform emittance distribution. We also assumed the intensity to have a revolution angular distribution. We assumed a Gaussian angular intensity distribution. We then calculated the coupling efficiencies of the sources for the different POFs. In a next step we simulated the receiver side to calculate how much of the light emerging from the POF impinges on the detector area. Here again, a user-defined source models the light that is coupled out of the POF. Multiplying the values of the coupling efficiencies of both the emitter and receiver side for an identical working distance L then gives us the transmission efficiency of the optical interconnection system for this working distance.

4.2.2 Simulation results and discussion

We show the results of the transmission efficiency for two combinations of the diameter and the NA of the POF as a function of the distance between the POF and the emitter or receiver in Figure 10.

We observe that the NA of the fiber used does not affect the coupling efficiency at the emitter side because of the small divergence angle of the laser. This means that a fiber with a smaller NA and diameter can be used with the result being an improved coupling efficiency at the detector side and a more relaxed cross-talk condition.

Figure 10. Transmission of the POF based guided-wave interconnect using a VCSEL.

5. CONCLUSION

We have shown in this paper that our design is suitable and feasible for very high performance computing. The system is characterized by immense bisection bandwidth, scalability, and low interconnects complexity. Our design
meets all the performance objectives we set out to achieve. The design is able to control combinatorial explosion of complexity by encapsulating complexity within the basic building blocks or nodes. We have showed that optical interconnection will be an inevitable solution to the bandwidth needs anticipated in the quest for petaFLOP performance. Analyses of the optical interconnection network were employed to further support our claim that our design achieves outstanding performance.

REFERENCES

BIOGRAPHY

Ekpe Okorafor received the B. Eng in ECE from the University of Nigeria in 1996, the MSc in ECE from Texas A&M University in 2001, and is currently rounding up his PhD at Texas A&M University. He has been a Graduate Research Assistant from 2000 to date. He has also worked in different research labs including HP labs and IBM. His research interests are in the areas of optical interconnection networks, massively parallel and distributed computing, high speed computer network systems, grid computing, mobile computing and computer architectures.

Dr. Mi Lu received her MSc and PhD from Rice University in 1984 and 1987 respectively. She is a Professor of ECE at Texas A&M University. She has co-authored over 90 technical publications and is a senior member of IEEE. She has served as associate editor for several journals. She has advised over 40 PhD/MS students and received several national awards. Her research interests are in the areas of parallel computing, distributed processing, computer architectures and networks, computational geometry, parallelizing compiler and VLSI algorithms.