Analysis and Implementation of Optoelectronic

Mongkol Raksapatcharawong

CENG 98-20

Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, California 90089-2562
(213-740-4482)
December 1998

Analysis and Implementation of Optoelectronic Network Routers

By

Mongkol Raksapatcharawong

A Dissertation Presented to the

FACULTY OF THE GRADUATE SCHOOL

UNIVERSITY OF SOUTHERN CALIFORNIA

In Partial Fulfillment of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY

(Electrical Engineering—Systems)

December 1998

Acknowledgements

I am truly grateful to my dissertation advisor, Dr. Timothy Mark Pinkston, for his invaluable support, guidance and encouragement throughout the course of my work. He has been working much harder than I have to bring me to this day. With his broad knowledge and insight, working with him is like a cultivating process that transforms me to a well academically educated person.

I thank my qualification and defense committee members: Prof. Alexander A. Sawchuk, Prof. B. Keith Jenkins, Prof. Monte Ung, Prof. Daniel P. Dapkus, and Prof. Clifford Neuman for their valuable time and constructive suggestions.

A fellowship from the Royal Thai Government was also important to my success today. The continuing support from 1992 to 1997 had covered most part of my education and is forever appreciated. The services from the staffs at the Office of Educational Affairs, Royal Thai Embassy, in Washington DC are always acknowledged.

Collaboration among the SMART group members is also another important factor to my success. I have received uncountable constructive suggestions and comments from them (in alphabetical order): Yungho Chooi, Joon-Ho Ha, Wei Hong Ho, and Sugath Warnakulasuriya, and our group's past members—Anjan K. V. and Seelan.

I also thank Joe Anadian and his group at MIT for the OPTOCHIP project (WARRP core), and Ashok Krishnamoorthy and his group at Lucent Technology lab for the CMOS/SEED project (WARRP II). The help from Dr. Charlie Kuznia, Chi-Hoa Chen, Bogdan Hoanca, and Jen-Ming Wu during the course of design and testing both chips are greatly thankful. The equipment donated by Altera (Joe Hanson) and Xilinx (Jason Fiensmith), and the EPOCH/EGGO CAD tool software donated by CADCADE (Ray Farbarik) are very helpful and I would like to thank for their support here. In addition, I

am deeply appreciated Regina Morton and Mary Zittercob for their tremendous administrative help.

Last but not least are the endless love and encouragement from my mom, Rada Raksapatcharawong, and my sisters, Sukhumal Kasemsook, Khuntalee Raksapatcharawong and Piyapat Wiboonsrisajja, and my nephew and nieces. They are always there for me. I also thank to all my friends in Thailand and around the U.S. for their constant support.

Abstract

Network routers based on optoelectronic technology have the potential to solve the network bandwidth problem which is becoming more and more critical in multiprocessor systems. By combining high-bandwidth optoelectronic I/O technology and high-performance CMOS logic technology, optoelectronic network routers promise both sophisticated switching functions as well as ample bandwidth that scales well with the performance of current and next-generation processors. Performance analysis and implementation of optoelectronic routers or other optoelectronic chips with this level of complexity, however, have not been pursued to a great extent before. This dissertation uses analytical and semi-empirical models to quantify and estimate the performance of optoelectronic routers at the chip and system levels, and it studies the feasibility of implementing such routers using GaAs MESFET/LED/OPFET and CMOS/SEED integrated technologies. The results show that optoelectronic routers may not only be technologically viable but also can provide certain architectural advantages in multiprocessor systems. Nevertheless, as shown in this dissertation, three major requirements must be met to effectively utilize this new technology. First, small and robust packaging at the chip and system levels that ensure high-bandwidth operation at useful interconnection distances and topologies are needed. Second, optoelectronic compatible CAD tools that effectively integrate a large array of optoelectronic devices with complex circuitry while retaining the potential performance of optoelectronic chips are needed. Third, optoelectronic devices must have uniform characteristics and reliability. In addition, advanced architectural techniques that efficiently exploit high-bandwidth optical interconnects are also required.

Contents

List of Tables

Chapter 1

Introduction

1.1 Motivation and Objectives

Microprocessor performance has improved continuously since the introduction of the microprocessor a few decades ago. The main reason for this is that constant progress in semiconductor technology has enabled more sophisticated architectures to be fabricated on a chip. Architectural techniques such as speculative execution, superscaling, superpipelining, multithreading, prefetching, etc., have allowed the processor to operate at faster clock rates and simultaneously execute several instructions at once (i.e., increase hardware parallelism). Current processors can theoretically perform up to several billions of operations per second. However, to achieve that level of performance, data must be delivered to/from the processor at the same rate as that by which it is being processed. This demands a great amount of bandwidth from the processor-memory bus, which is currently in the range of gigabytes per second and increasing. Unfortunately, the conventional electrical-based bus is not keeping pace with that bandwidth figure and is becoming a bottleneck. The consequence is obvious: the processor spends more time waiting for data, which significantly degrades its potential performance. A solution could be as simple as integrating the processor core and memory on the same chip in a so-called Intelligent RAM (IRAM) [1] configuration. Mitsubishi has commercialized this approach in its M32R/D processor [2] which integrates a 32-bit RISC core into a 16-Mbit DRAM chip running at 100MHz.

In a multiprocessor system, processing nodes which consist of (but are not limited to) processor(s), memory, and a network interface are distributed throughout the system and are connected via the underlying interconnection network to provide system communication capability. The bandwidth problem is therefore propagated on all levels of interconnects. This configuration does not benefit much from IRAM because data references are not limited to only local memory; some have to go through the network to remote memory located on other nodes. With IRAM, the problem at the interconnection network is exacerbated because the communication within a processing node is now much more efficient than that among the processing nodes. Therefore, from a system perspective, having a low-latency high-bandwidth interconnection network is more important in harnessing the potential performance of a multiprocessor system because remote access latency can be several orders of magnitude larger than that of local access.

The network router is an essential component that routes and manages the traffic in the interconnection network. Current semiconductor technology allows the network router to incorporate multiple functional units as well as advanced architectural techniques such as true-fully adaptive, deadlock recovery routing [3] or an enhanced crossbar structure [4]. As with the microprocessor, those schemes result in faster operation and higher bandwidth utilization as long as the network bandwidth (i.e., link bandwidth) is sufficiently large. State-of-the-art network routers that are commercially available now operate at a humble 375MHz or less clock rates and 20-bit-wide or less datapath [5, 6, 7], which yields less than 1GB/s of raw bandwidth per direction per port. This is essentially limited by the bandwidth provided by electrical interconnects. This bandwidth figure is being outpaced by the bandwidth demanded by current and next-generation processors.

Like semiconductor technology, optoelectronic technology has been successfully developed to the point where large arrays of optoelectronic devices can be effectively

integrated on a high-performance VLSI circuit. This novel technology has paved the way to the development of optoelectronic network routers that can potentially solve the network bandwidth problem. Optoelectronic network routers feature high-bandwidth optical interconnects by means of a large number of I/O pin-outs, each capable of operating at very high speeds. The optoelectronic network router can be further optimized to achieve performance beyond that of an electronic-based network router (Section 6.2). A performance study of optical networks (Chapter 3) shows that increased bandwidth allows flexible choices of network topology while keeping the network latency virtually constant. Despite the great promise of optoelectronic technology, implementation of optoelectronic chips as complex as network routers has just recently begun and, thus, there are many unknowns to consider.

The goals of this thesis are to quantify the performance advantages of an optoelectronic network router at the system network level and to identify the cost and performance issues in designing optoelectronic network routers. This work argues in favor of the development of optical interconnects. It also provides evidence that optoelectronic technology is ready for complex designs such as network routers. In the course of this investigation, two optoelectronic network router chips were successfully designed and implemented using GaAs MESFET/LED/OPFET and CMOS/SEED integrated technologies (Appendix C). In summary, the results of this study answer fundamental questions pertinent to the development of optoelectronic network routers and optical-based multiprocessor networks.

1.2 Research Approach

I believe that performance analysis and implementation of optoelectronic network routers at the levels useful to computer architects and chip designers will provide sufficient information to validate the viability of an optical network based on optoelectronic routers. This is summarized in the following thesis question and hypothesis:

Thesis Question:

Can an optoelectronic network router achieve a significant performance advantage compared to an electronic network router in a multiprocessor network environment? Can it be effectively and efficiently implemented with present or near-term technologies?

Hypothesis:

Performance analysis based on analytical and semi-empirical models at the network and chip levels will show that an optoelectronic router has the potential to outperform its electronic counterpart, given that some design and packaging issues are effectively addressed.

Performance evaluation using an analytical model is conducted here to determine the performance advantages of optical networks in the context of a multiprocessor system. Assuming the well-known k-ary n-cube class of networks and Diffractive Reflective Optical Interconnects (DROI) [8], I establish a relationship between network-level parameters and device-level parameters that is capable of identifying the cost and performance of an optical network. Due to the model's generality, it is also used to determine the performance of electrical networks with very little modification.

Implementation of optoelectronic network router chips is the approach used to evaluate the technological feasibility at the chip level. Implementations using monolithic GaAs MESFET/LED/OPFET and hybrid CMOS/SEED integrated technologies are explored. This gives an insight into how technology affects the implementation of optoelectronic chips, e.g., number of optoelectronic devices available, possible device switching speed, etc.

The design of an optoelectronic router also imposes some requirements. The optoelectronic network router possesses a circuitry far more complex than what has previously been implemented. Therefore, it requires more sophisticated CAD tools that

can handle the wiring between a structured array of optoelectronic devices and randomly distributed I/O ports. This leads to performance tradeoffs that do not exist for electronic chips. Experience with optoelectronic chip implementation has led to the conclusion that wiring can be a major problem, especially when dealing with hybrid CMOS/SEED technology. I evaluate the effects of the wiring problem on reduced transistor density and increased critical path lengths. A semi-empirical model, in which some parameters are obtained by experiments, is established. Based on available technology roadmaps, this model shows the performance trends of optoelectronic chips that can be used to compare with that of electronic chips. The results of this analysis will be used to determine the conditions for which optoelectronic chips are superior to conventional chips.

1.3 Related Work

The work presented here is comprehensive and encompasses several major research areas including interconnection technology, performance modeling, and optoelectronic chip design and implementation. While there has been related work in these areas, this work distinguishes itself by offering a complete research framework for optoelectronic network routers which aims to tackle the bandwidth problem in multiprocessor networks.

Although optical interconnects have been regarded as an alternative high-bandwidth technology for short-haul communications, some researchers are still searching for ways to revitalize the performance of electrical interconnects. Simply put, the problem with electrical interconnects are a limited number of I/O pin-outs and a modest signaling speed. The bidirectional signaling scheme [9] addresses the problem of a limited number of I/O pin-outs by allowing signals to simultaneously flow in both directions (employing a multilevel signaling scheme) on the same point-to-point interconnect. Although this technique can reduce the number of pin-outs required by half, it complicates the design of the transceiver circuits. Signal detection can become an issue in a multilevel signaling environment and may limit the operational speed. In addition, for an application like a

network router, doubling the available pin-outs may not significantly improve performance and flexibility.

On the other hand, the equalized serial line scheme [10] uses an equalizer circuit to compensate the frequency-dependent attenuation in an electrical wire such that the total frequency response is flat throughout the operating frequency. The requirement for an equalizer circuit and the additional serializer/deserializer circuits (required to interface with the internal datapath) results in an extremely large transceiver circuit and which limits the number of available pin-outs. To its credit, this technique allows longer interconnection distances at high speed. However, compared to various parallel optical links [11, 12, 13, 14, 15, 16] which require simpler and smaller transceivers and yet yield higher bandwidth, the future of the equalized serial line technique may never flourish. Those parallel optical link projects strongly encourage the development of an optoelectronic network router that features parallel links and operates comparably to an on-chip clock rate while keeping the transceiver circuit as simple as possible such that large arrays of optoelectronic I/O pin-outs can be incorporated.

Others have performed extensive work in optical interconnects modeling. Fundamental limits on the communication capacity of optical interconnects, for a given communication volume, are shown in [17, 18]. I introduce a similar concept called "connection capacity" which represents the number of optical links that can be realized in a given volume. Performance modeling of a hierarchical optical network called Optical Multi-Mesh Hypercube (OMMH) which uses diffractive analysis is performed in [19]. The work presented here follows a similar procedure but assumes a widely employed k-ary n-cube class of network of various sizes using DROI. In addition, a performance metric of interest to network designers is also developed based on the model proposed for electrical interconnects [20] with an extension to incorporate the connection capacity concept in optical interconnects.

The field of optoelectronic chip implementation has been very active. Early efforts were put towards the development of small circuits and, therefore, did not gain much attention from computer architects. Such "pixel-based" design paradigms incorporate a small amount of transistors and optoelectronic devices to form a small circuit with optical I/O pin-outs called a "smart pixel." To fully utilize the chip area and I/O bandwidth, this smart pixel is replicated throughout the chip forming a 2-D array of smart pixels. Hence, pixel-based designs are very useful for massively parallel applications which require simple functions such as signal processing [21], bit-slice arithmetic logic unit (ALU) [22], and simple switch [23]. In order to gain momentum, the optoelectronic chip must incorporate large and complex circuitry and a large number of optoelectronic devices. This approach has recently been conducted and is having success for implementing more complex optoelectronic chips. Due to the circuit size and complexity, this design paradigm is called "core-based" which is considered to be in the same class as "genius pixel" [24] (but the design is not necessarily pixel-like). Examples of core-based design are the WARRP core [25], the WARRP II router chip [26], the AMOEBA switch chip [16], and a 64-bit microprocessor core [27].

Wiring between an array of optoelectronic devices and the randomly distributed circuit I/O ports can be a problem in core-based designs. This has never been a problem in pixel-based designs because they are self-contained; most connections are local within the pixel. However, core-based designs can be as large as the entire chip area and can have a significant number of global connections. The requirements of an imaging system and interconnection patterns further complicate the wiring problem. For instance, chip input-output pairs must be placed in a structured pattern, and there can be a lot of global crisscrossing connections. To completely wire the connections, there must be sufficient wiring resources available (e.g., metal layers and wiring channels—the space between groups of standard cells). Consequently, core-based chips have less transistor density and longer wires compared to pixel-based or electronic chips. These performance tradeoffs

must be evaluated in order to validate the expected performance gain of optoelectronic chips.

Previous core-based designs simplified the wiring problem by proposing several layout schemes [28, 29]. Such schemes require manual wiring of optoelectronic devices, which does not deliver the full potential of complex optoelectronic chips. Recent development of optoelectronic compatible CAD tools has enabled the wiring to be fully automatic and, thus, yield better chip performance and chip area utilization. It has also made performance evaluation possible. The first core-based design to employ the optoelectronic compatible EPOCH/EGGO tools is a DSP core [30]. However, the only performance-related parameter reported is the average wire length between the transceiver circuits and the optoelectronic devices, which is shown to be only 25% of the manually wired core-based chip (100μm compared to 400μm). The research presented here goes further in that it models the effects of the wiring problem on the chip performance using these CAD tools. The semi-empirical model predicts the chip performance in terms of reduced transistor density and reduced achievable off-chip clock rate, which is essential to the success of core-based designs. This work also makes performance comparisons between core-based and electronic chips more meaningful.

1.4 Thesis Contribution

The optoelectronic network router is a very interesting concept with the potential, many believe, to be able to solve the network bandwidth problem. However, the development of an optoelectronic chip of this complexity is still in the early stage and there is not much literature to prove its feasibility and performance. This dissertation attempts to fill in this information gap by giving a detailed performance analysis as well as an implementation of an optoelectronic network router. It is hoped that this investigation will stimulate accelerated research in this field. The main contributions of this dissertation are as follows:

- Explanation of the network bandwidth problem in detail and establishment of convincing arguments for optoelectronic network routers.
- Performance modeling of an optical network (and an electrical network) based on k-ary n-cube network topologies. This model yields performance parameters that are useful to network designers. Insight into how network and device parameters affect the network performance is also presented.
- Introduction of the "connection capacity" cost model for 3-D free-space optical interconnect.
- Implementations of optoelectronic network routers based on monolithic and hybrid optoelectronic/VLSI integration techniques.
- Identification of the wiring problem in designing complex optoelectronic chips.
- Introduction of a semi-empirical model to analyze the performance of complex optoelectronic chips affected by the wiring problem. This provides significant information to validate the expected performance of optoelectronic chips.
- Suggestions for advanced architectures to further improve the performance of optoelectronic network routers. Such architectures include efficient channel configurations interfacing between the internal and external datapath, asynchronous token-based virtual channel arbitration, flit-bundling external flow control, and efficient buffer management.

1.5 Thesis Organization

This thesis is divided into six chapters as follow:

1. **Introduction**: This chapter gives the motivation and objectives for this research. It also presents the thesis question and research approach, and it discusses related work and the contribution of this dissertation.

2. **Background**: This chapter explains in detail why network bandwidth is becoming a problem in multiprocessor networks and why optoelectronic technology is a potential solution. The problem is described in terms of limited I/O pin-outs and the widening performance gap between on-chip and off-chip clock rates.

3. **Performance Modeling of Optical k-ary n-cube Wormhole Networks**: This chapter describes a performance model that incorporates parameters of both optoelectronic devices and network systems. This model is used to analyze the network performance of optical and electrical interconnects. Other considerations that affect the network performance are also discussed.

4. **Design Issues in Core-based Optoelectronic Chips**: This chapter describes the wiring problem associated with complex optoelectronic chip designs including the WARRP router. The problem and its effects on chip performance are evaluated using a semi-empirical model. The methodology for obtaining the model parameters is explained. Performance estimation of both CMOS and optoelectronic chips using the model is presented

5. **Implementation of an Optoelectronic WARRP Router**: This chapter evaluates the technological feasibility of optoelectronic network routers. Implementations of the WARRP router based on monolithic and hybrid optoelectronic/VLSI integration technologies are described and related issues are discussed.

6. **Conclusions and Future Work**: This chapter summarizes the results obtained in this dissertation. Advanced architectural techniques that can be incorporated in an optoelectronic network router to further enhance network performance are suggested.

Chapter 2

Background

2.1 Bandwidth Trending

Progress in semiconductor technology seems to be without limit (at least for the near future). For every new technology generation, die size grows by 30% and transistor feature size decreases by 30%, which almost triples the number of transistors on a chip. This increased number of transistors enables more advanced architectures and/or multiple functional units to be incorporated in high-performance chips such as microprocessors or complex application specific ICs (ASICs). Not only do the additional circuits require a larger number of metal layers to connect but they also require more I/O pin-outs because more data is generated/required to/from outside the chip. This relationship can be represented by Rent's rule [31] as:

$$gate = \left(\frac{I/O}{k} \right)^{c}, \tag{1}$$

where *gate* is the average number of gates supported by I/O terminals, *I/O* is the number of signal terminals or chip pin-outs, k is a constant value of which depends on terminal sharing (i.e., $0.15 < k < 0.2$ for memory chips, $0.4 < k < 0.6$ for switching and computer chips, $k > 1$ for high performance logic chips), and c is a constant in the 1.5 to 3.0 range.

Based on technology trends, shown in Table 1, predicted by the Semiconductor Industry Association (SIA) [32], Eq.[1] was used to approximate the number of required I/O pin-outs for next generation processors assuming $k = 1.2$ and $c = 1.8$. Surprisingly,

the trend line resulting from those data points are in excellent agreement with those of past and current processors. The results are compared with the number of pin-outs offered by the high-performance Ball Grid Array (BGA) packaging as depicted in Figure 1. Clearly, BGA packaging would be unlikely to satisfy the pin-outs required by the processors by the new millenium! This situation explicitly calls for dense I/O pin-outs packaging which is not likely to happen with electrical interconnect technology.

Table 1. Semiconductor and optoelectronic SEED technology roadmaps.

Year of first shipment	1999	2001	2003	2006	2009
Technology (μm)	0.18	0.15	0.13	0.10	0.07
# Transistors (millions)	6.2	10	18	39	84
On-chip/Off-chip Clocks (MHz)	1250/480	1500/785	2100/885	3500/1035	6000/1285
# Pin-outs Required (pins)	1570	2000	2400	3270	4400
# BGA Package Pin-outs (pins)	1500	1800	2200	3000	4100
# SEEDs (per chip)	8000	12000	20000	35000	47000
Bonding Pad size (μm)	9	8	7	5	4

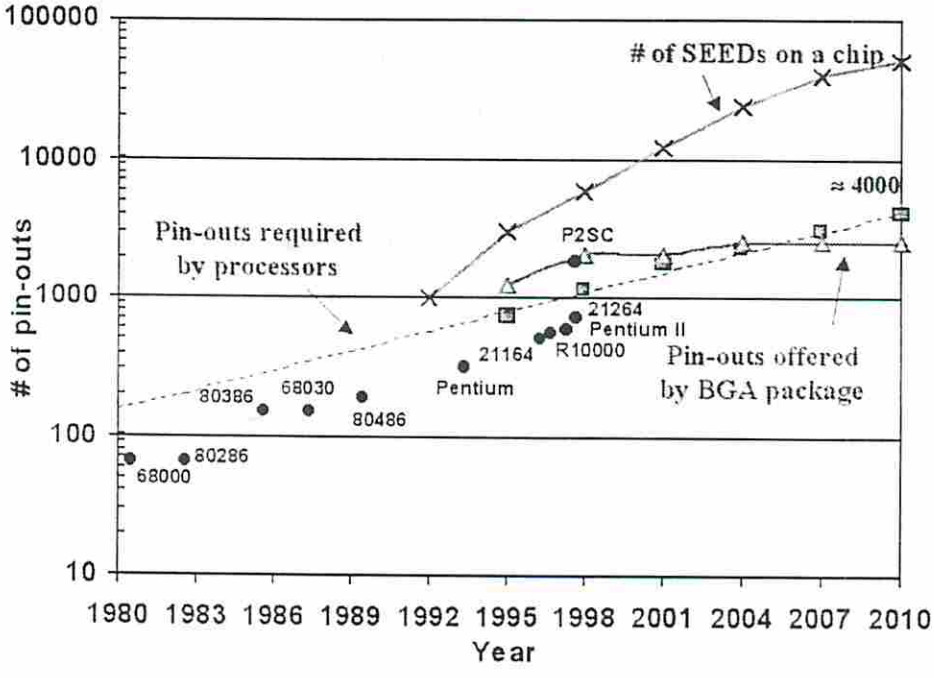


Figure 1. Pin-outs demand and supply trends of past, current, and next-generation processors. Optoelectronic SEED trend is also plotted showing the number of devices available on a chip that is far greater than the pin-outs required by the processors.

Alternatively, Self-Electrooptic-Effect-Devices (SEEDs), one of the most promising optoelectronic technologies to date, has been successfully integrated on top of CMOS-VLSI circuitry. Currently, the integration of up to 32x64 SEEDs has been reported [33]. In keeping with what has previously been achieved, Krishnamoorthy has predicted a similar technology trend for SEEDs [34], as shown in Table 1. Seemingly, optoelectronic SEED technology exhibits a promising number of pin-outs that can easily sustain pin-outs required by next generation processors.

Limited performance of electrical interconnects leads to an increasing gap between on-chip and off-chip clock rates. Table 1 also shows that an on-chip processor clock could achieve as high as 1.9GHz while an off-chip clock will slowly reach 475MHz in the next decade. Together with limited number of I/O pin-outs, which may cause the external datapath to be much narrower than internal datapath, off-chip bandwidth available could be an order of magnitude less than that required by the processor core. Fortunately, optoelectronic SEED technology has been proven to be efficient in this regard as well; each pair of SEEDs can currently transfer data up to 2.48Gb/s in dual-rail mode [35]. Hence, this technology is capable of providing tens to thousands of terabits per second of aggregate off-chip bandwidth which is far beyond the required bandwidth of the processors.

Memory access time is another performance bottleneck. The processor-memory performance gap, in terms of memory access time and processor clock cycle, has been increasing. For instance, the access time of a high-performance PC100-SDRAM utilizing a 100MHz memory-bus is around 20ns to 30ns whereas the clock cycle of the Pentium II-400 MHz processor is about 2.5ns and is decreasing. As a result, the processor spends more time waiting for data to arrive. The situation is exacerbated in distributed memory multiprocessor systems where this latency can be as high as several hundreds of processor cycles. To sustain the potential performance of the processor core, latency-

tolerating techniques are required. Prefetching [36] is a well-known latency tolerating technique that issues a memory request before it is actually accessed, moving the data closer to the processor. Hopefully, when the data is referenced it would be a cache hit. In effect, prefetching pipelines multiple memory accesses and overlaps them with program execution. Multithreading [37] is another latency tolerating technique that allows the existence of multiple execution contexts called "threads" running concurrently in a processor. When the processor encounters a cache miss, it suspends the executing thread and selects another eligible thread to execute, thereby overlapping the memory access with thread execution. Both techniques, nevertheless, increase the off-chip bandwidth because of multiple outstanding memory accesses. To illustrate this effect on a multithreading processor, SPEC92 benchmark data [38] was used to calculate the required off-chip bandwidth for single- and multithreaded processors, assuming the DEC Alpha processor architecture with performance extrapolated to the year 2007. The results are plotted against the available processor bus bandwidth in Figure 2.

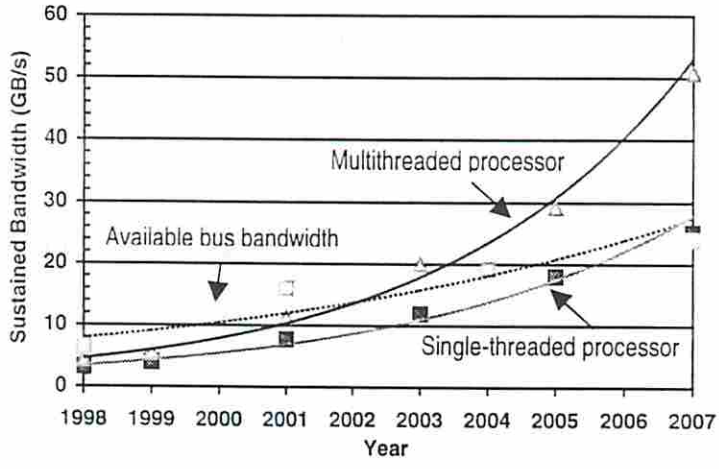


Figure 2. Off-chip bandwidth required by single– and multithreaded processors versus off-chip bandwidth supplied by the processor bus (calculated using SIA data).

Figure 2 shows that the off-chip bandwidth of multithreaded processors is about twice the bandwidth required by single-threaded processors and could be in excess of 50GB/s in a decade. This number rapidly outpaces the bandwidth provided by the processor bus.

In this case, high-bandwidth interconnects are *necessary* to enable latency tolerating techniques which are required to achieve higher performance. A simple solution to tackle the processor-memory performance gap is to integrate a processor core onto a memory chip called IRAM [1, 2]. This architecture significantly improves performance in uniprocessor systems by *increasing* memory bandwidth and *decreasing* memory access time. Alternatively, high-bandwidth optical interconnects may also be used with the inclusion of latency-tolerating techniques. Seemingly, IRAM may currently have an edge over optical interconnects but its electrical interconnects will soon limit the achievable bandwidth. Therefore, optical interconnects are considered a longer-term solution due to their higher bandwidth scalability.

2.2 Multiprocessor Network and Network Router

Bandwidth trending at the network level in multiprocessor systems resembles that of a processor-memory bus in uniprocessor systems but the solution to the problem is not as simple. To make the discussion more understandable, the components of a distributed multiprocessor system are depicted in Figure 3.

In a distributed multiprocessor system, processing nodes which consist of (but are not limited to) processor(s), memory, I/O controller, and network controller are connected together via the interconnection network. Memory references are satisfied either locally on the same processing node or globally on the other nodes. The remote (global) references are handled by the interconnection network and, hence, affected by the network performance. According to processor bandwidth trends, it is reasonable to assume that the required network bandwidth to handle such remote references will increase rapidly. The main contributions to the required network bandwidth are faster processors, latency tolerating techniques, larger system size, and the application's communication behavior.

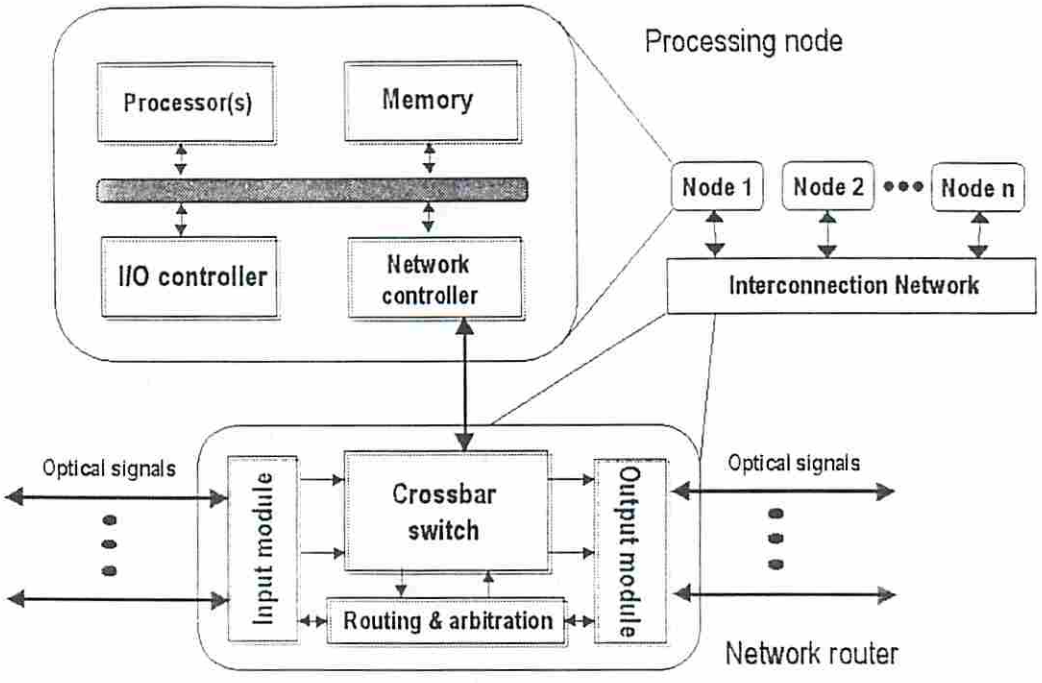


Figure 3. A distributed multiprocessor system with an optical interconnection network (a regular electrical interconnection network would be the same except all signals are electrical).

There are two reasons that IRAM is not sufficient to achieve the potential system performance (and in some cases can even degrade the performance). First, some parallel applications do not exhibit good locality and generate mostly remote references. Hence, improving local accesses as accomplished in IRAM does not drastically improve the performance. Second, each memory level in the system can be represented as a pipelining stage. Reducing the latency of one stage without improving the latency and/or throughput of the consequence stages would result in prohibitively long queuing delay as suggested by queuing theory. In this case, the network controller and the interconnection network can be overwhelmed by the references that are misses in local memory.

In contrast, high-bandwidth low-latency interconnection network is always beneficial to the overall performance in multiprocessor systems because network latency is much higher than memory access latency. Since the interconnection network is simply a collection of "network routers" and "physical links," we can tackle the bandwidth

problem by improving the performance of the network router, or increasing the link bandwidth, or both. Each option has different pros and cons, and will be discussed in greater detail in the next section.

2.3 Current State-of-the-Art Network Routers

The network router is the brain of the interconnection network. It performs almost every function including routing, switching, and managing the network traffic, leaving only the transporting function to the physical link. All advanced network routers currently available (including the experimental ones) incorporate several architectural techniques to improve the utilization of network bandwidth.

Table 2. Current state-of-the-art network routers.

Router	On-chip/Off-chip clock rates (MHz)	Internal/External channel width (bits)
SGI Spider [5]	100/200 (double-edge)	80/20
Intel Teraflop [6]	200/200	16/16
Cray T3E [7]	75/375	70/14

Table 2 summarizes three state-of-the-art network routers that were built for commercial large-scale distributed multiprocessor systems and their corresponding on-chip and off-chip bandwidth (shown as clock rates and datapath/channel width). The SGI Spider features an adaptive routing algorithm via on-chip table lookup. Adaptive routing allows a packet to take one of several paths to move towards its destination, preferably the less congested one. In contrast, non-adaptive or deterministic routing allows a packet to use only a single path throughout its course. Hence, adaptive routing can evenly distribute the traffic throughout the network, which improves the network bandwidth utilization and, in most cases, reduces the network latency. To reduce the on-chip latency, for each packet, the Spider chip looks ahead at the routing information for all the possible output ports for the next Spider chip on the selected path. This pipelining strategy overlaps the table lookup latency, which is 10ns, with the crossbar arbitration and CRC check. In addition, the Spider chip sends data to the links on both clock edges. This

effectively doubles the link bandwidth. Each link has 4 virtual channels associated with it. The virtual channel is simply a buffer with associated control circuit that is used to improve bandwidth utilization by temporarily suspending the blocked packet and allocating the link bandwidth to others.

The Intel Teraflop router, called Cavalino, was designed to support thousands of processing nodes. To achieve the highest possible on-chip throughput, the Cavalino chip distributes the routing and flow-control to each link. Link bandwidth utilization is optimized by the use of 4 virtual channels. The crossbar employs 2-level arbiter to pipeline the arbitration latency. Unlike the Spider chip, Cavalino features deterministic routing. To reduce the number of required I/O pin-outs, simultaneous bidirectional signaling is used.

The Cray T3E router features both adaptive and non-adaptive routing schemes. Each physical link has 6 virtual channels where only one virtual channel allows adaptive routing. This router supports adaptive routing as proposed by Duato [39] and prevents "deadlock" in the network. While adaptive routing can improve network performance, it can also lead to deadlocks. Deadlock can happen by some packets holding on to network resources while requesting resources occupied by others in a cyclic fashion [40]. When this occurs, all packets involved in deadlock cannot make any progress and eventually the whole network becomes stagnate. Duato observed that by routing non-adaptively in some channels, deadlocks can be prevented altogether and, hence, the technique is called deadlock avoidance routing. In this scheme, however, the bandwidth of non-adaptive channels cannot be optimally utilized, which sacrifices some of the network performance.

Table 2 shows a rule of thumb in designing network routers, which is to make the on-chip bandwidth at most equal to the available off-chip bandwidth. This rule comes about because network routers cannot use caching techniques to reduce the amount of off-chip communication; all incoming traffic must be routed to the output ports. Therefore, the

design of network routers is governed by the available off-chip bandwidth which is currently very limited. The situation is getting more critical as progress in semiconductor technology continues. As a matter of fact, a network router's operation is much simpler than a processor's and can be optimized to operate at a very high clock speed. Without high-bandwidth interconnects, the network router can barely take advantage of the progress in semiconductor technology.

The essence of off-chip bandwidth to network routers can be further explained in terms of their I/O pin-outs requirement. A network router is essentially an intelligent switch which always benefits from having wide communication channels and more communication ports. Both features reduce the network latency and give design flexibility to the network. It is interesting to estimate the I/O pin-outs requirement of advanced network routers while keeping in mind which configurations are supported by current technology and which are not (thus, requiring alternative technologies).

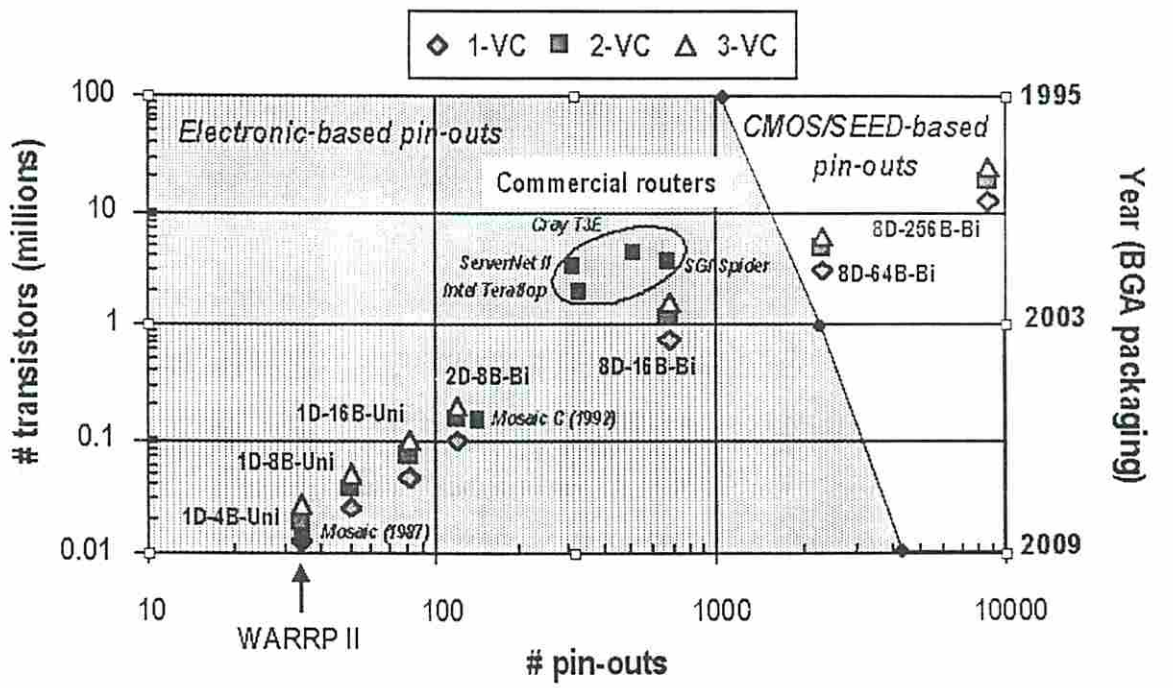


Figure 4. Network router complexity and its pin-outs requirement.

In Figure 4, the Wormhole Adaptive Recovery-based Routing via Preemption (WARRP) router (developed by the SMART Interconnects group [41]) complexity and its pin-outs requirement is juxtaposed with those of the past and current network routers. The leftmost data points represent a torus-connected 4-bit-wide unidirectional channel (1D-4B-Uni) WARRP router with 1, 2, and 3 virtual channels. The WARRP II implementation (discussed in Chapter 5) has only one virtual channel. Despite a comparable complexity, the WARRP router features fully adaptive deadlock recovery routing whereas the MOSAIC router features a simple dimension order routing. Each generation, which is about 5 years apart, network routers (e.g., MOSAIC—1987, MOSAIC C—1992, and SGI SPIDER—1997) require almost an order of magnitude more in terms of pin-outs and transistors. These routers were implemented with electrical packaging as shown in the dark-shaded area. However, if this trend continues, next generation network routers could require the number of pin-outs that is beyond the capability of high-performance BGA packaging. This is the reason why optoelectronic pin-outs (shown in white area) should be investigated as an alternative technology to facilitate the development of future high-performance network routers (including the more complex configurations of the WARRP router). Current commercial network routers are usually constrained by limited pin-outs to have a wider internal datapath compared to external channel width. Consequently, on-chip clock rate is forced to be slower than the off-chip clock rate (which cannot be very fast either!) to balance between on-chip and off-chip bandwidth as are the cases for the SGI Spider and Cray T3E routers. In conclusion, the design of high-performance network routers is impeded by the limited off-chip bandwidth.

2.4 High-performance Electrical Interconnect Technology

Improving the performance of electrical wires is also a direct solution to increase the network bandwidth. Prior attempts to enhance the bandwidth of electrical wires include

the equalized serial line [10] and simultaneous bidirectional signaling [9] schemes. The former tries to increase signaling speed so the off-chip clock can be faster, and thus increases the interconnection distance. The latter utilizes multilevel signaling on a wire to double the bandwidth (i.e., reducing the I/O pin-out requirement for data by half). The equalized serial line scheme uses the equalizer circuitry to compensate the frequency-dependent attenuation in the transmission line such that total attenuation is flat throughout the operating frequency range. Recent experiments have shown that this technique can yield 4Gb/s bandwidth at up to a 6-meter interconnection distance. The simultaneous bidirectional signaling scheme reduces the number of pin-outs by simultaneously transmitting signals in both directions on a point-to-point interconnect. This technique requires that the transceiver must be able to detect multilevel signals, as opposed to regular two-level signals. Bandwidth of 2.5Gb/s/wire at several centimeters on a printed circuit board has been demonstrated. Nevertheless, both schemes complicate the design of the transceiver circuits. In addition to the equalizer, the equalized serial line's transceiver requires multipexer/demultiplexer to interface between internal datapath and external serial line. Thus, each transceiver requires a very large die area which means that only a few I/O pin-outs can be realized on a chip. Simultaneous bidirectional signaling requires a very sensitive transceiver, which is susceptible to signal noise and ground bounce. A comparison of transceiver sizes and operating speeds of a regular I/O pad, SEED transceivers, and an equalized serial line, based on a 0.5μm CMOS HP14B process, are shown in Figure 5.

In Figure 5, numbers in caption's parenthesis indicate the relative size of each circuit compared to the regular I/O pad. The optoelectronic transceiver is ~50 times smaller yet operates 12 times faster. Also shown (not to the same scale) is the equalized serial line transmitter. Its operation speed is comparable to the optoelectronic transceiver but its size (excluding I/O pad driver) is more than 2200 times larger.

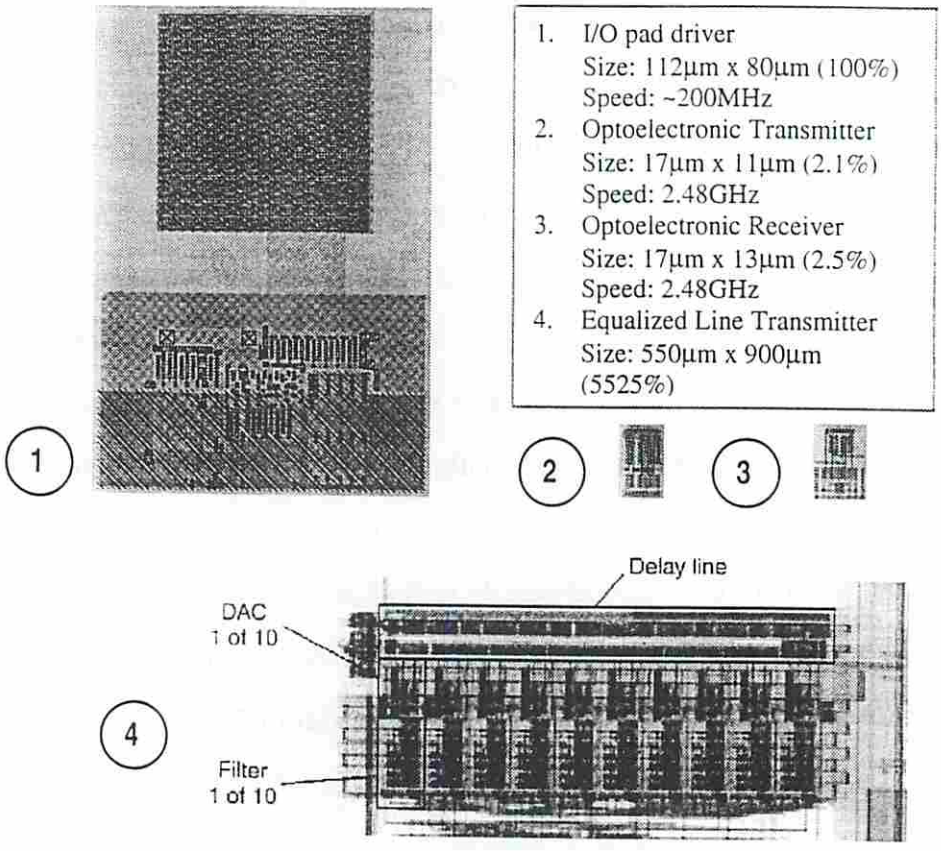


Figure 5. Transceiver size and speed comparison among various interconnection technologies.

It is evident that electrical interconnect technology can still advance, albeit at the expense of increasing complexity. The question then becomes to what extent can electrical interconnect technology be improved? Miller [42] has shown that the fundamental limitation of electrical wires is the aspect ratio of the interconnection length (l) to the total cross-sectional dimension of the interconnect wiring ($\sqrt{A}$). For a broad range of electrical cables this fundamental limit is shown to be approximately $10^9 A^2/l$ Gb/s. For example, the best MCM-D technology can achieve only 750Gb/s bandwidth for a 25cm^2 module with 40μm line width. Moreover, signal skew and jitter are critical to performance at high data rates, causing electrical wires to operate only in serial mode at a longer distance. Although such factors exist in the optical domain, they are less severe. Table 3 shows current parallel optical links that feature 10-bit wide or more links

and yield a data transfer rate of 1Gb/s/fiber or more, at few hundreds of meters interconnection distance. Clearly, such configurations cannot currently be realized with electrical interconnects.

Table 3. Current bit-parallel optical link interfaces comparison.

Interface	Year	Transmission rate (GHz)	Channel width (bit)
OETC [11] (the Optoelectronic Technology Consortium)	GE/AT&T/Honeywell/IBM, 92	0.625	32
Jitney [12]	IBM/3M/Lexmark, 93	0.400	24
POLO [13] (Parallel Optical Link Organization)	USC/HP, 96	1.000	10
Optobus II [14]	Motorola, 95	0.800	10
ChEEtah [15] (Cost Effective Embedding of High Performance Interconnects)	USC/Honeywell et al., 97	> 1.000	12
AMOEBA [16]	Lucent Technology/UNCC, 96	0.050	11

Up to this point, the ability of optoelectronic technology to provide high-bandwidth interconnects via dense I/O devices and high-speed signaling is shown to be essential for developing high-performance network routers required by next-generation processors. Rapid progress in both fabrication and packaging techniques is making this alternative interconnect technology a nearer-term solution. For example, a hybrid integration of CMOS VLSI circuitry with GaAs-based multiple-quantum well (MQW) modulators (i.e., SEEDs) via flip-chip bonding with more than 16,000 devices on a single die at 99.878% yield [33] has been reported. Recent experiments have shown that each device can operate up to 2.48Gb/s with less than 300μW optical power in dual-rail mode [35]. Monolithic integration of GaAs MESFET circuitry with VCSELs/MSM photodetectors that operate at more than 9GHz is also possible [43]. Some other combinations are also actively being investigated such as CMOS circuitry with VCSELs/MSM photodetectors [44] and GaAs MESFET circuitry with LEDs/OPFET photodetectors [45].

2.5 Proposed Solution: An Optoelectronic WARRP Router

Having identified the pros and cons of previous attempts to solve the bandwidth problem, this dissertation proposes an optoelectronic implementation of the WARRP router [41, 46]. It addresses the bandwidth problem by increasing the network bandwidth via optoelectronic I/O technology.

The first question regarding the proposed solution is, "How does it benefit the multiprocessor networks?" Conceptually, we should expect optoelectronic I/O technology to improve the network performance in two ways. Firstly, the design flexibility of multiprocessor networks can be enhanced through a large number of available I/O pin-outs, i.e., a wide-range of topologies is efficiently supported. Secondly, the design of high-performance network routers is possible through high-speed signaling. For example, a fully pipelined network router can be designed to operate at full-speed without being limited by off-chip clock rates. Both are major advantages over the ubiquitous electrical I/O packaging technology and are discussed in Chapter 3.

The next question is, "What are the major issues pertinent to the success of optoelectronic network routers?" A design issue that can be a major problem is the wiring between SEEDs and CMOS circuitry. This affects the expected performance of optoelectronic network routers. Nevertheless, we should expect an optoelectronic network router to outperform its electronic counterpart in terms of available bandwidth and number of pin-outs. Performance evaluation is explained and discussed in Chapter 4.

The last question to be answered in this dissertation is, "Can optoelectronic network routers be implemented?" This question is answered through various implementations of the WARRP router based on different optoelectronic technologies which are elaborated in Chapter 5.

To further improve network bandwidth availability and utilization, some advanced network router architectures are also suggested in Chapter 6. Despite being suggested for optical interconnect technology, such architectures are applicable to any high-bandwidth low-latency interconnect technologies.

Although the questions and concomitant issues raised are addressed using the WARRP router as an example, the knowledge gained about optoelectronic technology can be applied to network routers in general. At this moment, the proposed solution appears advantageous in the development of next-generation high-performance network routers which incorporate features to efficiently deal with the network bandwidth problem.

Chapter 3

Performance Modeling of Optical k-ary n-cube Wormhole Networks

This chapter presents a cost and performance modeling of optical-based interconnection networks based on DROI. It is shown here that optoelectronic routers can significantly improve the network performance and efficiently support a wide range of network configurations.

3.1 Free-Space Optical k-ary n-cube Wormhole Networks

Table 4. k-ary n-cube network characteristics (for unidirectional links).

• node and edge symmetric
• regular topology
• connectivity = 2n
• maximum degree = n
• maximum diameter = n(k-1)
• average distance D_{avg} = n(k-1)/2 (for uniform message distribution)
• channels = $nN = nk^n$
• bisection width (channels) = $2N/k = 2nk^{n-1}$
• nodes in system = $N = k^n$

The k-ary n-cube class of networks has a number of favorable characteristics as given by Table 4 and, hence, is among the more popular. They employ static, direct point-to-point connections between nodes and support locality of communication to reduce delay of messages in the network. Topologies for this class of networks have channels which span n dimensions and have k nodes connected in each dimension (radix). The switching

technique, however, can have a greater influence on the delay of messages in the network than topology.

The network analysis here assumes wormhole switching [20] which pipelines the transfer of flits[1] along the path from source to destination. Once a node receives the header flit of a message (which contains all the relevant routing information), the header flit is routed to an appropriate output channel. If that channel is free, the header is transferred to the next node; all other flits follow sequentially. If the required channel is busy, all flits are blocked behind the header and wait until the channel becomes available. Therefore, the latency resulting from wormhole switching can be expressed simply as

$$T_{lat} = T_C \cdot \left(D + L_F \cdot \frac{F}{W} \right) + T_{contention}, \tag{2}$$

where T_C is the channel cycle time for transceiving and routing flits, D is the number of network hops required from source node to destination node, L_F is the data message length in filts, F is the flit size in bits/flit, and W is the physical channel width in bits (also referred to as the phit size). The congestion along the path from source to destination due to messages contending for the same channel is parameterized by the $T_{contention}$ variable. Note that the contention delay is not modeled in this work which assumes low-load networks. This is sufficiently accurate because such delay is very small compared to other latency components in low-load operating regions. The channel cycle time, T_C, is the maximum between external and internal router delays assuming both input and output are buffered [47]. Note that by pipelining logic functions in the network router, the external propagation delay of signals (i.e., signal propagation time in the interconnection medium and signal conversion/re-generation, if applicable) can become the critical path which determines the channel cycle time. The internal router delay

---

[1] A flit or flow control unit is the unit of message transfer on which flow control is performed.

includes the decision time to route the header flit (t_r) and the switching time to switch a flit from input to output buffers (t_w). A latency diagram of low-load wormhole switched networks assuming the channel cycle time is determined by the internal router delay and is depicted in Figure 6.

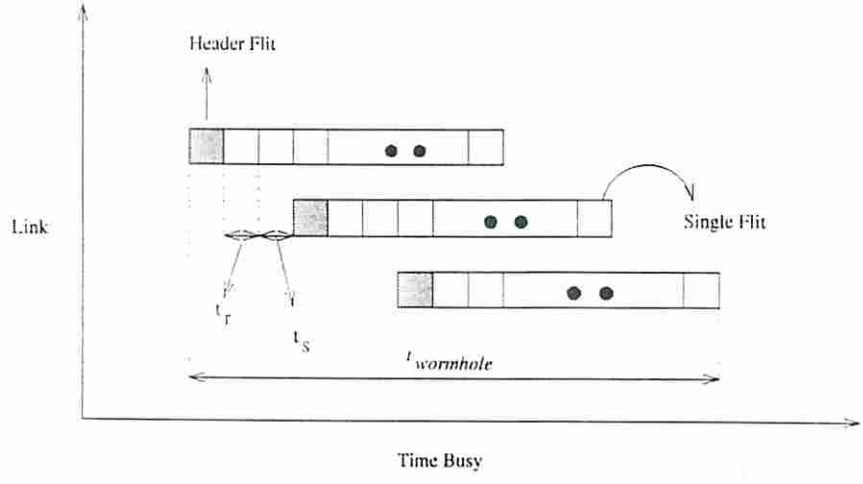


Figure 6. Wormhole switched network latency diagram.

Conceptually, free-space optical interconnects can comprise a transmitter plane, a receiver plane, and an optical imaging system in between. Optical beams are transmitted by transmitters (e.g., light sources or modulators) in the transmitter plane, deflected and/or split by the optical imaging system (e.g., holograms, lenses, etc.), and detected by sensors (e.g., photodiodes, modulators, etc.) at the receiver plane. Free-space optical interconnects can freely make use of the third dimension to route signals. In contrast, wiring freedom in electrical interconnects is limited to only on the same plane or vertically through multiple stacked planes. This is a major difference between optical and conventional electrical interconnects.

Wiring in two-dimensional (planar) electrical VLSI implementation can be made as long as there is enough wiring space between the source and destination points. Thus, wiring can be made locally or globally according to a given topology. Local connections represent all wiring to near neighbors whereas global connections are the connections that

cross from one side to another side of the wiring plane. Dally observed that these global connections are critical to the wireability of k-ary n-cube networks and are densest at the middle of the system wiring area [20]. Hence, the notion of *bisection width* constraint was introduced, which is the number of wires crossing an imaginary plane that divides the system into two equal halves. This notion can be extended to develop expressions for network latency in optical interconnects.

Unlike electronics, connections are established over a volume in optics where each connection shares the same implementation cost. Therefore, the notion of bisection width is extended to better evaluate optical interconnects by introducing the notion of the connection capacity constraint, which is the number of connections that can be established for a given imaging system. The number of beam steering elements limits the number of connections that can be established in the system (which is defined as "connection capacity" in this study). Consequently, optically interconnected systems evenly distribute the implementation cost of all connections.

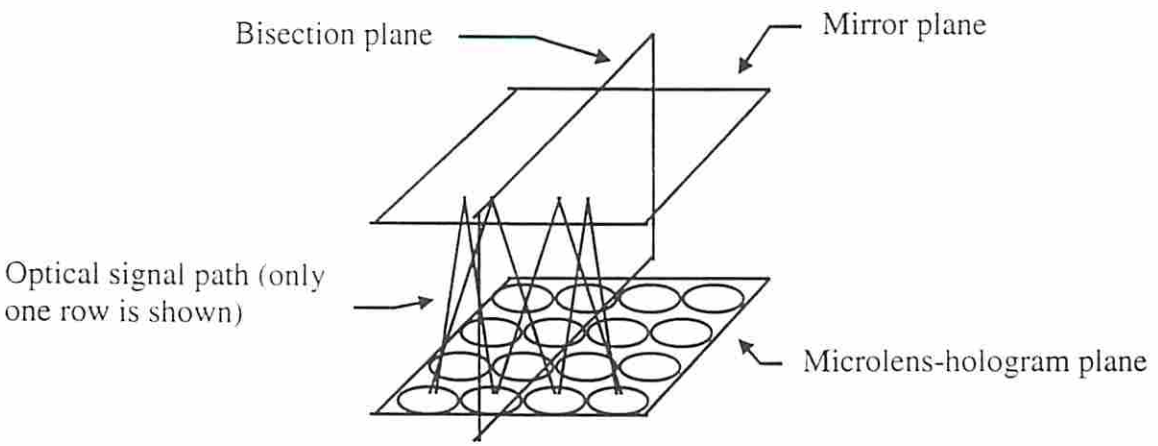


Figure 7. DROI of 4-ary 2-cube (torus) network.

For instance, Figure 7 shows a 4-ary 2-cube torus network using DROI. The bisection plane shown in the figure is an extension of Dally's bisection width applied to 3-D optically interconnected systems. It is shown here that bisection width fails to accurately represent implementation cost constraints in optical interconnection systems. To

demonstrate this, two systems with the following parameters are assumed: System A is a 16-node torus with 32 connections and a bisection width of 8, and System B is an 8-node hypercube with 24 connections and a bisection width of 8 as well.

In a wire-limited VLSI system with a bisection width of 8, both systems can be implemented. However, an optical system that has a connection capacity of 24 can only implement the second system even though both systems have the same bisection width. Therefore, the connection capacity constraint takes all connections into account while the bisection width constraint takes only global connections into account. *Connection capacity is a more accurate and appropriate implementation cost metric for 3-D optical interconnection networks.* This leads to the difference in performance sensitivity of optical- and electrical-based networks to the topologies, as shown in the next section. However, the bisection width of an optical interconnection network can always be found given that its connection capacity is known. This relationship is useful for comparing the performance of electrical- and optical-based networks using the conventional bisection width parameter.

3.2 The Model

An analytical model for k-ary n-cube optical networks with wormhole switching is developed in this section. This model is an extension of Dally's analysis [20] applied to optical interconnects. The analysis is primarily based on the notion of connection capacity as opposed to bisection width. Any free-space optical interconnect system can be described in terms of its connection capacity. Below, the connection capacity is shown to have a significant impact on the types of topologies that can be efficiently supported.

3.2.1 *Connection-Efficient Topologies*

Let C be the connection capacity of an optical imaging system (C is constant). The number and width of channels supported for various k-ary n-cube topological configurations can be described in terms of connection capacity:

$$C = lW = nNW = nk^nW, \tag{3}$$

where l, the number of unidirectional channels or links required by a k-ary n-cube network, is given in Table 4. Hence, a relationship between bisection width (B) and connection capacity can be written as

$$B = \frac{C}{nN} \cdot \frac{2N}{k} = \frac{2C}{nk} = \frac{2C}{n\sqrt[n]{N}}. \tag{4}$$

Eq.[4] shows that bisection width of an optical network with constant connection capacity changes with topology, and becomes larger for high-dimensional networks $\left(B \propto 1/n\sqrt[n]{N}\right)$. This is because the number of channel crossing between two equal sub-networks increases faster than the decreasing channel width (Eq.[4]) for higher dimensions. Therefore, there is a trade-off between network latency and bisection width in implementing an optical-based network.

Comparisons between various topological configurations are more comprehensible when we normalize connection capacity to that of the hypercube (binary n-cube) topology with unity channel width. Normalized connection capacity results in $C = N\log N$. The channel width $W(k,n)$ of a k-ary n-cube with normalized connection capacity is therefore given by[2]

$$W(k,n) = \frac{N\log N}{nN} = \frac{\log N}{n} = \log k\,. \tag{5}$$

---

[2] Throughout this dissertation, log x stands for $\log_2$ x.

This expression for channel width with a normalized connection capacity is different from that for the normalized bisection width derived by Dally [20], where a *constant bisection width* was assumed. Under that assumption, channel width was shown to grow linearly with increasing k (i.e., $W(k,n) = k/2$). Intuitively, when k increases the dimension, n, decreases accordingly, increasing the number of links and, thus, decreasing the channel width, W. Clearly, the model shows that optically implemented topologies are less sensitive to the radix, k, because the logarithmic function ($log\ k$) changes less rapidly than the linear function ($k/2$). Hence, the advantages expected from lower dimension networks (namely, wider channels) should not be as pronounced for optical (connection capacity limited) networks as they are for electrical (bisection bandwidth limited) networks. Insight into how this conclusion impacts the expected latency of optical interconnects is shown below.

Latency is the time required to deliver a message from source to destination. The average latency for a k-ary n-cube network can be found as follows. If the source and destination nodes P_s and P_d are randomly selected with equal probability, the average number of hops between them is given by

$$D = \left(\frac{k-1}{2}\right) \cdot n, \tag{6}$$

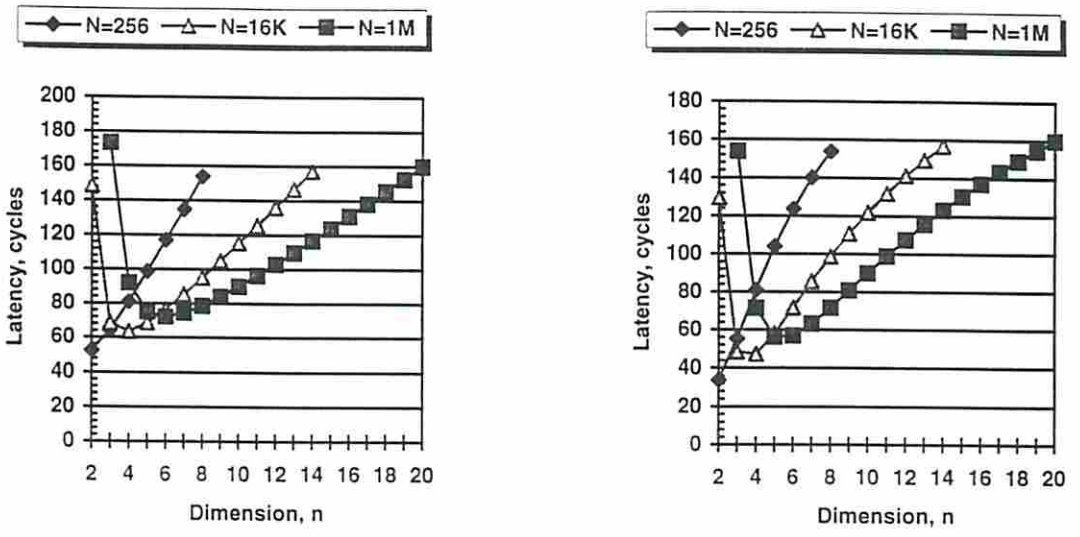
where $(k-1)/2$ is the average number of hops the message travels in each dimension given that links are unidirectional, and n is the number of dimensions.

Substituting this expression and the channel width from Eq.[5] into Eq.[2], the average latency of an optical k-ary n-cube network is:

$$T_{lat} = T_C \cdot \left(\left(\frac{k-1}{2}\right) \cdot n + L_F \cdot \frac{F}{\log k}\right) + T_{contention}. \tag{7}$$

The second term in Eq.[7] is the only difference between the optical model and Dally's electrical model; it is less sensitive to the radix of the network than the electrical model.

Therefore, latency should increase more slowly with dimension as compared to Dally's model. Latency characteristics for both models are illustrated in Figure 8.



(a) Optical model.

(b) Electrical model (Dally, 1990).

Figure 8. Latency versus dimension with unit channel cycle time.

Figure 8 depicts the average network latency as a function of dimension for k-ary n-cube networks with N=256, 16K and 1M nodes. A unit channel cycle time is assumed, and the message length $L=F \cdot L_F$ (in bits) is assumed to be 150 bits (flit size and channel width are assumed to be equal henceforth). Thus, the above figure represents the latency for *constant delay* of both optical and electrical signals regardless of the physical distance between source and destination nodes. *It should be noted that Figure 8 is not intended to compare the latencies of optics and electronics*; rather, it shows how dimension affects latency for each design space (optics or electronics).

For each curve, the rightmost data point corresponds to a hypercube and the leftmost data point corresponds to a 2-D torus. In low-dimensional networks, messages travel a greater number of hops. Latency is dominated by this hop distance even with wormhole routing (network congestion would further degrade performance). In contrast, messages

suffer increased transfer time (in flits) between nodes for high-dimensional networks due to the smaller channel width offered by the topology. Here, latency is dominated by message length. Hence, the results agree with [20] in that low-dimensional networks outperform high-dimensional networks in terms of latency for both design spaces. However, as we will see later, higher dimensional networks are not as disadvantageous for optics, especially given that it is less difficult to implement higher dimensional networks with free-space optics than with wire-limited electronics. This is because the model takes into account all connections. Therefore, under constant connection capacity assumption, low- and high-dimensional networks are just as costly (as noted in Section 3.1), and performance in terms of channel width is less topology-dependent.

3.2.2 The Channel Cycle Time (T_C)

The previous analysis assumed constant channel cycle time. In what follows, the model is developed in more detail to include the effects of optical signal delay assuming that the path external to the router defines channel cycle time (i.e., T_C = max[external router delay ($T_{C\text{-}ext}$) , internal router delay ($T_{C\text{-}int}$)]). The time to convert and propagate an optical signal between a pair of nodes is given by the following:

$$T_{C\text{-}ext} = T_{e/o} + T_{prop} + T_{o/e}. \tag{8}$$

The first term, $T_{e/o}$, is the electro-optical conversion time for the optical source (or modulator) circuit. The last term, $T_{o/e}$, is the opto-electronic conversion time of the receiver circuit. The second term, T_{prop}, is the light propagation time which is approximately 1ns per foot in a vacuum. The assumed optical signal delay model is depicted in Figure 9.

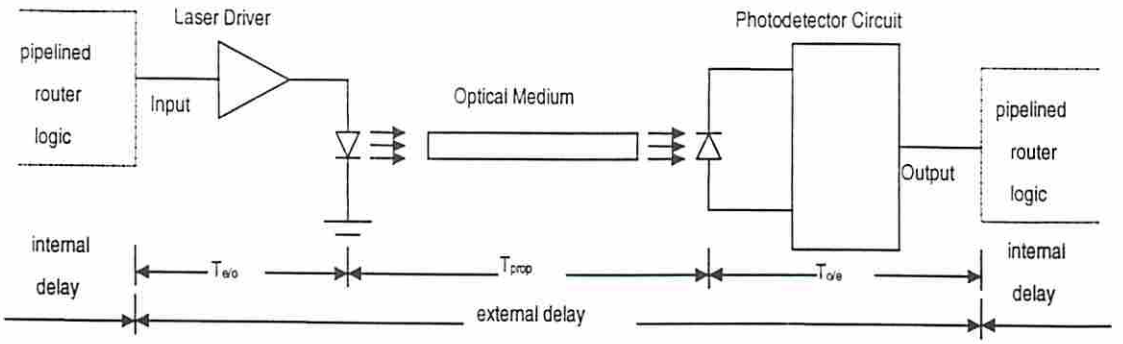


Figure 9. The optical signal equivalent propagation path.

This work assumes a two-phase clock with equal length phases: the first phase for signal propagation and synchronization, and the second phase for routing decision. This scheme allows each router to operate asynchronously with a reasonable clock skew margin [47]. For example, given the internal delay of 6ns and the external delay of 5ns (T_C = 6ns). The network clock would be 2*6 = 12ns. Thus, the system would allow 1ns clock skew with a throughput of one phit per 6ns when both phases are pipelined. This margin can be increased to 7ns at the expense of a lower throughput of one phit per 12ns. In contrast, if external delay is greater than internal delay, some margin to the channel cycle time to accommodate the clock skew may be added. Since signal skew is less in optical systems [48], this problem should not affect the performance of optical interconnects as much as electrical interconnects.

The transmitters are assumed to be the differential output amplifier [20] as shown in Figure 10. The output of a logic gate (V_{in}) is compared with a reference voltage (V_{ref}). This comparison results in changing the current flowing through the Vertical-Cavity Surface-Emitting Laser (VCSEL). V_{source} in the figure is used to provide a constant current source for the differential amplifier. To enhance switching speed, the VCSEL is biased at approximately its threshold by V_{bias}. When the amount of current exceeds that threshold, the VCSEL starts emitting light, resulting in an electrical to optical modulation. The electrical to optical conversion delay is described by

$$T_{e/o} = \left(\frac{1}{\beta_n \left(V - V_{tn} \right)} + \frac{1}{\beta_p \left(V + V_{tp} \right)} \right) \cdot \left(C_o + C_{in} \right) + T_{diff} + T_{laser} \ , \qquad (9)$$

where C_o is the output capacitance of the output gate (neglecting wiring capacitances), C_{in} is the input capacitance of the differential amplifier, V is the supply voltage, β_n, β_p, V_{tn}, and V_{tp} are the n– and p–MOS transistor gains and threshold voltages of the output gate, respectively [49], T_{diff} is the delay of the differential amplifier, and T_{laser} is the laser response time.

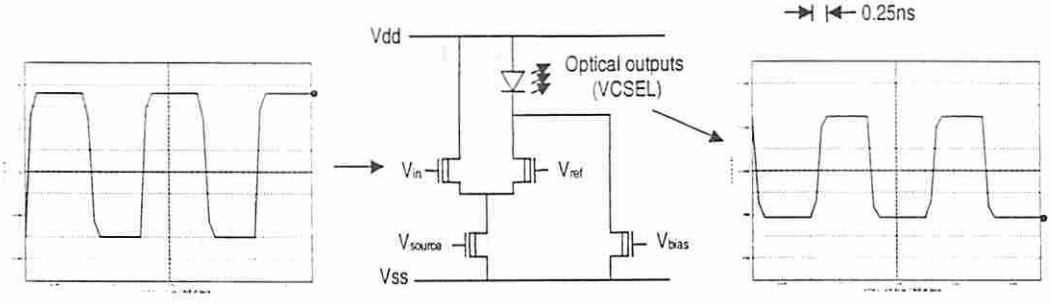


Figure 10. Schematic of a transmitter circuit and its SPICE waveforms.

Current VCSEL technology available from Sandia National Laboratories [50] provides 1.5mW optical output power at 2mA input current. Assuming a 0.5μm CMOS process with 3.3V supply voltage (HP14B process from MOSIS), a symmetric delay, unit-sized inverter gate yields C_o = 7.38fF. The input capacitance, C_{in}, of the differential amplifier is 8.82fF. The n– and p–MOS transistor gains and threshold voltages are 169μA/V^2, 211.3μA/V^2, 0.65V, and –0.93V. The laser delay, T_{laser}, is typically less than 0.1ns. The differential amplifier circuit was simulated assuming a VCSEL threshold current of 2.5mA with 2mA driving current. SPICE simulations in Figure 4 shows that the delay of the differential amplifier, T_{diff}, is approximately 0.25ns. These values yield $T_{e/o} \approx$ 0.36ns for electro-optical conversion delay according to Eq.[9].

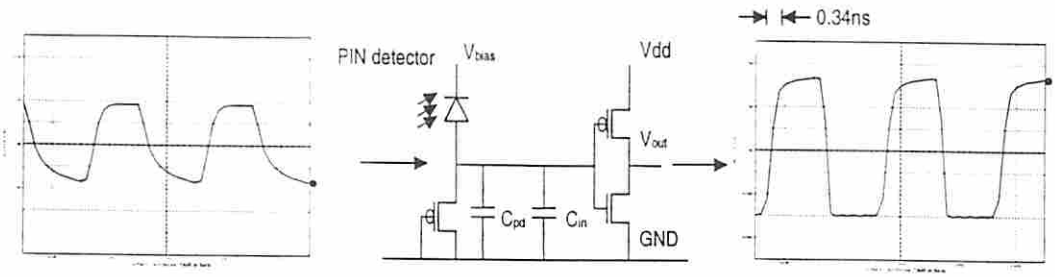


Figure 11. Schematic diagram of a P-I-N photodetector circuit and its SPICE waveforms.

Likewise, the receivers are assumed to be the receiver circuit [51] as shown in Figure 11. The receiver consists of P–I–N diode and an output driver. An expression for the delay in charging up voltage at the receiver is shown to be [51]

$$T_{o/e} = \frac{V}{S\eta P_{laser}}\left(C_{PD} + C_{in}\right), \tag{10}$$

where S is the P–I–N detector sensivity, P_{laser} is the optical power emitted by the VCSEL, V is the supply voltage, C_{PD} is the P–I–N detector capacitance, C_{in} is the input capacitance of the output driver, and η is the optical link efficiency (which includes that of the hologram and the microlens).

Assuming the same CMOS process and an input capacitance of a symmetric delay, unit sized inverter of 3.53fF, we get C_{PD} = 53fF for a detector area of $15\mu m$x$15\mu m$. Its sensivity is 0.5 A/W at 15V reverse-bias. In a DROI design with a 1.5mW VCSEL and an optical link efficiency of 63% (81% hologram efficiency for 4-level diffractive optical element (DOE) [52] and 99.5% microlens efficiency), we get an optical to electrical conversion delay of $T_{o/e} \approx 0.4$ns. SPICE simulation shows that the optoelectronic delay is 0.34ns. This number is not an under-estimate as detectors that operate beyond 2Gb/s with less than 300μW optical power have been reported [35].

The last major component of external channel cycle time is the propagation delay. This delay is dependent on the medium and its length. The most efficient way to

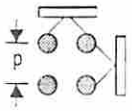
implement a network topology in a volume (where nodes are to reside in a plane) is to map the connections as symmetrically as possible so as to minimize connection length. Previous attempts to map optical k-ary n-cube topologies in a volume did not consider wrap around connections [53]. Here, the longest connection between two nodes in the system is defined as the "maximum connection path." Figure 12 shows a suggested layout of nodes and the mapping of connections in a volume (3-D) for various 4-ary n-cube topologies. With this layout, the maximum connection path, $R_{\max}$, is given by

$$R_{\max} = \begin{cases} \dfrac{p \cdot 2^{\left\lceil \frac{n}{2} \right\rceil - 1}}{\sin\theta} & k = 2, \\ \dfrac{p \cdot 2^{n-1}}{\sin\theta} & k = 4, \\ \dfrac{2 \cdot p \cdot k^{\left\lceil \frac{n}{2} \right\rceil - 1}}{\sin\theta} & \text{any other } k, \end{cases} \tag{11}$$

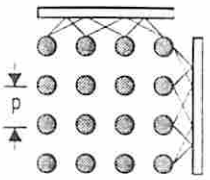
where $p = \sqrt{A/N}$ is the minimum connection length (lateral distance) between adjacent nodes, N is the number of nodes, A is the square area of the node plane, θ is the maximum hologram deflection angle, n and k are the dimension and the radix of the network. It should be noted that Eq.[11] applies to any configuration of n and k (which are integers) that fit perfectly in a square area. Therefore, the light propagation time is

$$T_{prop} = \frac{R_{\max} n_x}{c}, \tag{12}$$

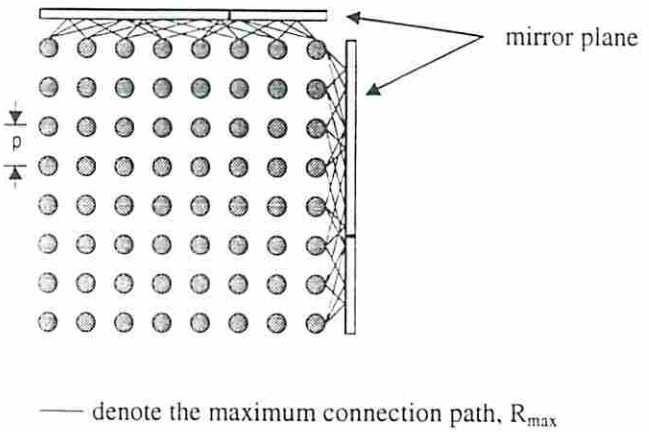
where c is the speed of light and n_x is the refraction index of the material. We can now estimate the external channel cycle time, $T_{C\text{-}ext}$. According to the values calculated previously, $T_{C\text{-}ext} = 1.76$ns for an optical link distance of $R_{\max} = 1$ foot in vacuum.



(a) Layout and mapping of 1-D network.



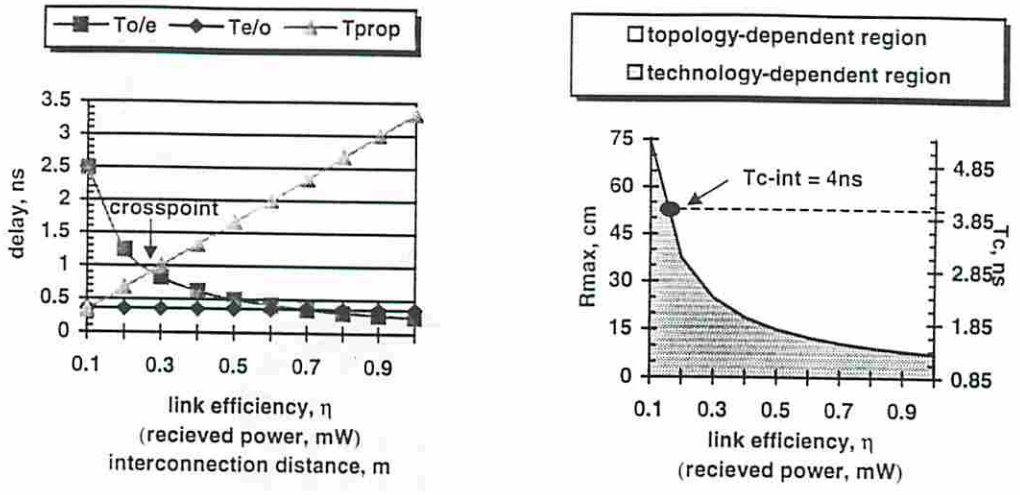
(b) Layout and mapping of 2-D network.



(c) Layout and mapping of 3-D network.

Figure 12. Embedding of 4-ary n-cubes in a volume (nodes in 2-D plane) for n=1,2,3. Only the connections of nodes along edges are shown for clarity. Moreover, the mirror plane shown here would be above the transmitter-receiver plane in a real system.

As previously described, the external channel cycle time can be decomposed into signal conversion delay and propagation delay. The former is *technology-dependent* (Eq.[10]) whereas the latter is *topology-dependent* (Eq.[11, 12]), making external channel cycle time dependent upon both technology and topology. Figure 13 shows the relationship between optical link efficiency (η) and conversion delays ($T_{e/o}$ and $T_{o/e}$), T_{prop} versus R_{max}, and the crosspoint curve where $T_{o/e} = T_{prop}$.



(a) Conversion delay and propagation delay. (b) Crosspoint curve which separates $T_{C\text{-}ext}$

Figure 13. (a) Plots of $T_{e/o}$ and $T_{o/e}$ vs link efficiency and T_{prop} vs $R_{\max}$ assuming a 1.5mW VCSEL and parameters in Section 3.2.2 and propagating in vacuum. (b) Regions defined by the set of crosspoints, which are the values of link efficiency where $T_{o/e} = T_{prop}$ (one crosspoint is shown in (a)), and the corresponding overall $T_{C\text{-}ext}$ value.

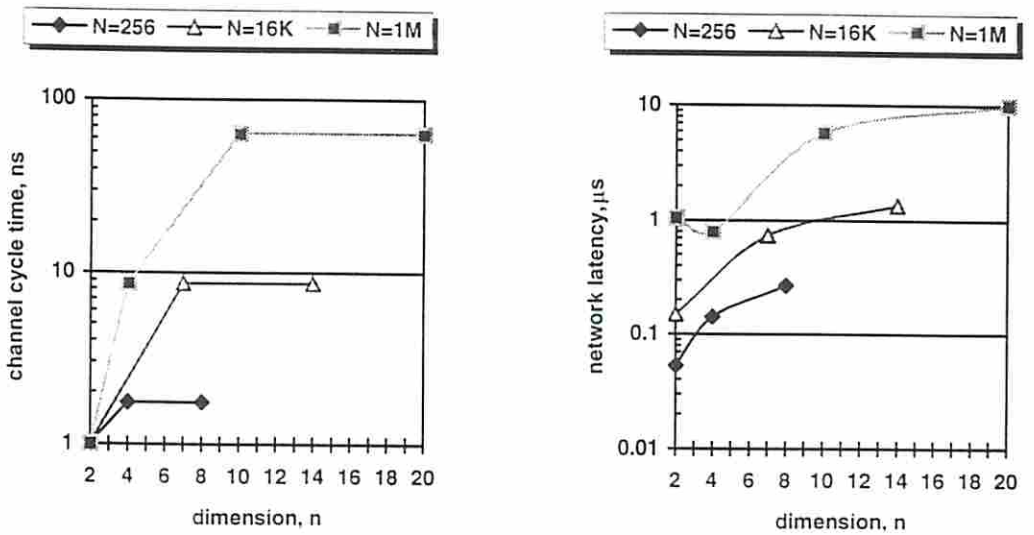
As depicted in Figure 13(a), low efficiency imaging systems can result in significant conversion delay ($T_{o/e}$) which can dominate external channel cycle time. In these cases, channel cycle time is technology-dependent and virtually constant for all topologies (Figure 13(b) shows this technology-dependent region). Small systems are likely to operate in this region because of shorter interconnect distance, including my hypothetical 64-node hypercube network which has $R_{\max} \approx 15cm$ (Section 3.3). It is interesting to note that the technology-dependent T_C is always larger than the topology-dependent T_C for a given R_{max}. Also, transmitters take less time to generate 1.5mW than detectors take to detect this power, as shown in Figure 13(a). This is because the area of the VCSEL is smaller than that of the P–I–N detector ($25\mu m^2$ compared to $225\mu m^2$ in our study).

From Figure 13(b), for a given internal router delay (e.g., $T_{C\text{-}int}$ = 4ns), overall channel cycle time (T_C) is bounded by *external delay* for points lying on the left side of the curve, and is bounded by *internal delay* for points lying on the curve to the right of the intersection. Therefore, for very fast routers with small internal delay (e.g., $T_{C\text{-}int}$ =

2ns), the imaging system should be as efficient as possible to minimize overall channel cycle time.

3.2.3 *Network Latency with Linear Optical Signal Delay*

The latency figures shown in Section 3.2.1 do not reflect the more realistic situation where the channel cycle time in optical networks is not constant but depends on interconnect distance. Assuming that the efficiency of a free-space optical system does not depend on distance, $T_{e/o}$ and $T_{o/e}$ in Eq.[8] remain constant. In this case, *linear optical signal delay* results where $T_C \propto T_{prop} \propto R_{\max}$.



(a) Channel cycle time. (b) Network latency.

Figure 14. Channel cycle time and network latency with linear optical signal delay ($T_{contention}$ excluded) for systems with N=256, 16K, and 1M nodes when normalized connection capacity is assumed (only topologies where k is an integer are plotted). Message length is 150 bits and the minimum connection length, p, is assumed to be 1.5 cm.

As expected, channel cycle time increases with n because of a greater R_{max}. This is made clear if we rewrite Eq.[11] as $R_{\max} \propto 1/\sqrt[n]{N}$. The smaller channel width further accentuates the latency difference between low- and high-dimensional networks. Together they make hop count less significant to network latency (this describes the

difference between the latency curves of Figure 14(b) and those in Figure 8(a)). Hence, with the linear delay assumption, low-dimensional networks still outperform high-dimensional networks in terms of network latency for a broad range of system sizes.

3.2.4 *Connection Capacity (C)*

Channel width of each topology in k-ary n-cube optical networks is determined by connection capacity. In general, the connection capacity for an optical imaging system is expressed as

$$C = \frac{A_{system}}{A_{spot}}, \tag{13}$$

where A_{system} is the area over which interconnects can be established and A_{spot} is the maximum light beam area along the propagation path. Assuming diffractive-reflective optical interconnects (DROI) and Gaussian beam propagation [Appendix A], these two parameters are shown to be functions of other system parameters [Appendix B]:

$$\begin{aligned} A_{system} &= F(\theta, h, p, n, k) \\ A_{spot} &= F(f, w, \lambda, \theta, h) \end{aligned}, \tag{14}$$

θ is the hologram deflection angle, h is the separation between mirror and microlens planes, f is the microlens focal length, w is the transmitted beam radius, λ is the wavelength, p is the minimum connection length, n is the dimension and k is the radix. The hologram deflection angle itself is also a function of other system parameters [Appendix B]:

$$\theta = F(\lambda, n_x, L_b, w_f), \tag{15}$$

where n_x is the index of refraction of the material through which optical signals propagate, L_b is the number of hologram levels, and w_f is the minimum feature size of

each hologram. Figure 15 illustrates the DROI geometry. Shown is one optical signal connection path.

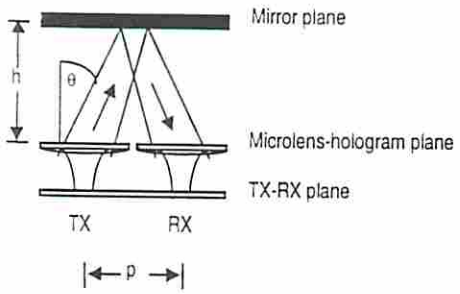


Figure 15. A DROI geometry.

In this study, A_{system} is held constant and h is allowed to vary according to the topology (no volume constraint). In addition, the spot size area, A_{spot}, is assumed to be equal to the microlens area, M_D^2. Hence, the connection capacity for the assumed DROI simplifies to

$$C = \frac{A_{system}}{2M_D^2}. \tag{16}$$

The factor of two in the denominator takes into account the fact that both transmitters and receivers are in the same plane. For example, assuming the DROI optical imaging system supports interconnection of nodes over an area of $A_{system} = 64cm^2$ and the lens diameter of each interconnection is $M_D = 125\mu m$, the connection capacity is 204,800 connections for all k-ary n-cube topological configurations.

3.3 Application of the Model: Optical vs Electrical Interconnects

It is useful to determine whether optics performs better than electronics and if so, by how much. Therefore, this section compares the latency given by the optical model with that given by [20] for k-ary n-cube networks. The optical model has as its constraint the connection capacity whereas Dally's electrical model has bisection width as its

constraint. These constraints, although different, are actually related since they are both used to determine channel width of the various topologies.

3.3.1 *Electrical Interconnect Delay Model*

The latency model of an electrical interconnect is based on distributed RC effects of a transmission line using a microstrip conductor with no transmission line effect [54]. The channel cycle time in electrical interconnects is given by

$$T_{C\text{-}elec} = T_{prop\text{-}elec} + T_{RC} . \tag{17}$$

Here, $T_{prop\text{-}elec}$ is the propagation delay in an electrical medium which is $\approx$ 0.148ns/in [54]. The RC delay, T_{RC}, takes into account the distributed RC effect of the transmission line [49] and delays associated with driver and receiver circuits as shown in Figure 16.

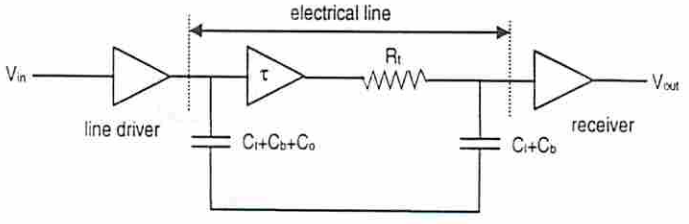


Figure 16. Simple model for electrical interconnect delay.

Here, C_i is the input capacitance of a receiver, C_o is the output inverter capacitance, C_b is the bonding pad capacitance, τ is the signal delay on a transmission line, C_t and R_t are the total lumped capacitance and resistance of the transmission line. The signal delay on the transmission line is given by [49]

$$\tau = \frac{rcl^2}{2}, \tag{18}$$

where r and c is the unit length resistance and capacitance which are 45.4mΩ/in and 1.0pF/in for a 5-mil-wide, 5-mil-apart, and 2.7-mil-thick conductor [54], and l is the transmission line length.

The total RC delay, T_{RC}, is expressed as follows:

$$T_{RC} = \left(\frac{1}{\beta_n (V - V_{tn})} + \frac{1}{\beta_p (V + V_{tp})} \right) (C_t + C_b + C_o) + \tau + R_t (C_t + C_b). \qquad (19)$$

The driver is assumed to provide an output current of 10mA. Thus, the output inverter capacitance, C_o, based on the same CMOS HP14B process is 91.1fF. The input capacitance of the unit-sized receiver, C_t, is 5.88fF. The n– and p–MOS driving transistor gains and threshold voltages are 2848μA/V^2, 3560.7μA/V^2, 0.65V, and –0.93V, respectively. The bonding pad capacitance is 0.4pF for the $100\mu m^2$ pad area. Given a maximum connection path, $R_{max\text{-}elec}$, the channel cycle time of an electrical interconnect can be found by using Eqs.[17-19]. For a 1-foot interconnection length, the channel cycle time of the above interconnects is 4.92ns (T_{RC} = 3.14ns and $T_{prop\text{-}elec}$ = 1.78ns).

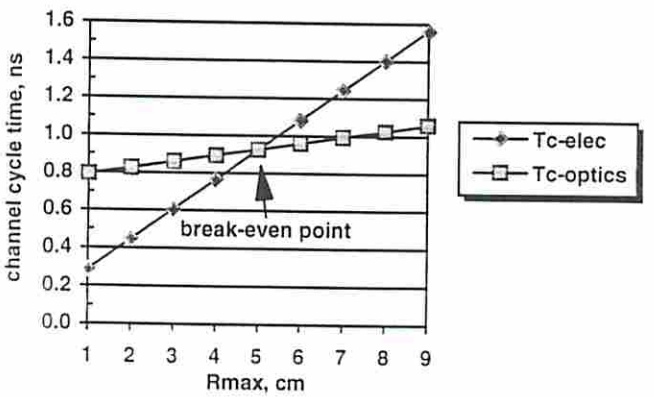


Figure 17. T_C and break-even point.

The channel cycle time for both optical and electrical interconnects varies with the maximum connection path R_{max}. Figure 17 shows the relation between the two according to the assumed parameters ($T_{e/o}$ = 0.36ns and $T_{o/e}$ = 0.4ns) using Eqs.[8-12] and Eqs.[17-19]. Channel cycle time in optics is less sensitive to R_{max} because only its propagation delay is topology-dependent whereas both propagation and RC delays are topology-dependent in electronics. Figure 17 shows the *break-even point* indicating the value of

R_{max} where both technologies yield the same external channel cycle time. Optics is superior for R_{max} greater than this point. The optic curve shifts downward for systems with a higher link efficiency, which results in a lower break-even point. Moreover, free-space optical systems are inherently more compact than electronic systems, so they generally operates at a smaller R_{max}. For instance, given a system size of $144in^2$ for electronics and $144cm^2$ for optics (assuming a deflection angle of 24°), we get $R_{max\text{-}elec}$ = $18.53cm$ and $R_{max\text{-}optics}$ = $7.41cm$ for a 2-D torus network.

3.3.2 *Channel Width*

Substituting Eq.[16] into Eq.[3], we express channel width for optical networks as a function of topological and implementation parameters:

$$W_{optics}(k,n) = \left(\frac{A}{2M_D^2 \cdot N \log N} \right) \cdot \log k \,. \tag{20}$$

Applying the bisection width notion to electrical networks, we express electrical channel width as a function of topological and implementation parameters:

$$W_{elec}(k,n) = \left(\frac{L\sqrt{A}}{NT_w} \right) \cdot \frac{k}{2}, \tag{21}$$

where A is the printed circuit board (PCB) area (or area of the microlens plane in optical networks), N is the system size, T_w is the electrical wire pitch, L is the number of PCB layers that can be routed in same direction, and M_D is the microlens diameter.

3.3.3 *Latency Comparison*

A comparison between electrical interconnects with aggressive PCB technology and optical interconnects with available optoelectronic and micro-optic technologies is performed in this section. Parameters assumed are listed in Table 5 and Table 6 for electrical and optical interconnects, respectively.

Table 5. Parameters for assumed electrical system.

Chip area	$1in^2$
PCB size	12in x 12in
# of layers	20
min. connection length (p)	1.5in

Table 6. Parameters for assumed optical system.

Laser wavelength (λ)	850nm
VCSEL beam radius	5μm
VCSEL output power	1.5mW
VCSEL efficiency (η_{VCSEL})	75%
P-I-N detector size	15μm x 15μm
Microlens diameter	125μm
chip area	$1\ cm^2$
Interconnection area	12cm x 12cm
Usable microlens area (A)	$64cm^2$
min. connection path (p)	1.5cm
Index of refraction (n_x)	1.5
max. deflection angle (θ_{max})	~ 24 °

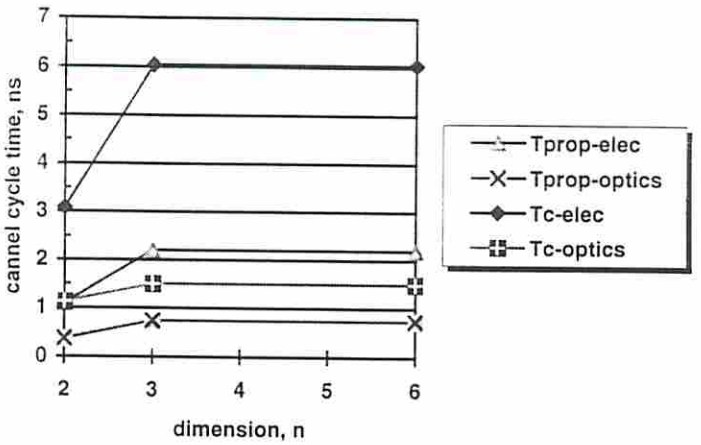
The electrical interconnects are implemented using a 20–layer PCB ($12in$x$12in$) in which 10 layers can be used to route signals in the same direction. The router die size is assumed to be the same as the Chaos Router chip which is $10mm$x$10mm$ [55]. Due to die packaging, each node occupies a square area of $1in^2$. All nodes are placed $0.5in$ apart, thus, the minimum connection length is $1.5in$. The conductor and spacing are $10mils$. This number is reported and implemented by Hewlett Packard [56]. Substituting these values yields a bisection width for the electrical interconnects of 12,000 connections. Other parameters for latency calculations are assumed to be the same as in Section 3.3.1.

The optical interconnects are assumed to be implemented with a $12cm$x$12cm$ transmitter-receiver plane. Each node occupies a square area of $1cm^2$ and is separated by $0.5cm$ from its neighbors for a minimum connection path of $1.5cm$ (die only). Recent studies confirm that packaging of multiprocessor free-space optical interconnects at this level of compaction is feasible [57, 58]. The VCSEL and P–I–N detector arrays are integrated on top of the CMOS circuits via flip chip bonding [59]. Therefore, only $64cm^2$

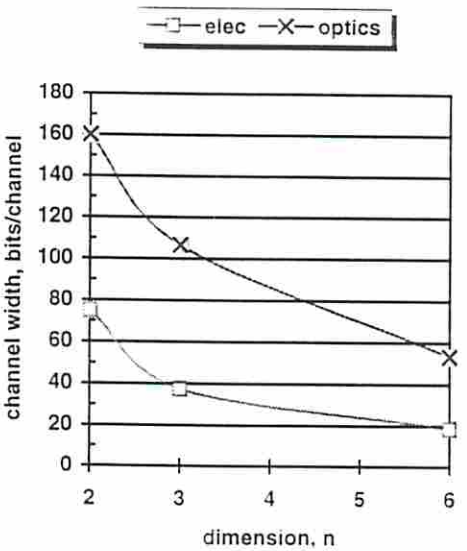
is available for microlens and hologram fabrication. Minimum feature sizes of $0.35\mu m$ for holograms and $0.5\mu m$ for CMOS circuits are assumed. Every plane is packaged together within glass with a refraction index of 1.5. Under the Gaussian-beam propagation assumption, microlens with $125\mu m$ diameter is sufficient to collect light with 99.5% efficiency and, hence, a connection capacity of 204,800 connections results. All other parameters for $T_{e/o}$ and $T_{o/e}$ are consistent with Section 3.2.2.

In comparing both types of interconnects, we can immediately see the great disparity in connectivity; optics provides ~17 times more connectivity. The maximum volume needed to sustain this connectivity is a modest ~$980cm^3$ ($12cm$x$12cm$x$6.78cm$). (a 64-node system was chosen due to the limited PCB area and transmitter-receiver plane area for the nodes.) The channel cycle time and network latency of the 64-node system with a message length L=1024 bits (or 128 byte packets) are plotted in Figure 18, assuming 10% of connection capacity is used for data lines due to practical considerations such as power, ground, and control lines. The dimension n is varied to observe the channel cycle time and network latency as given by both models (only topologies with integer values of k are observed).

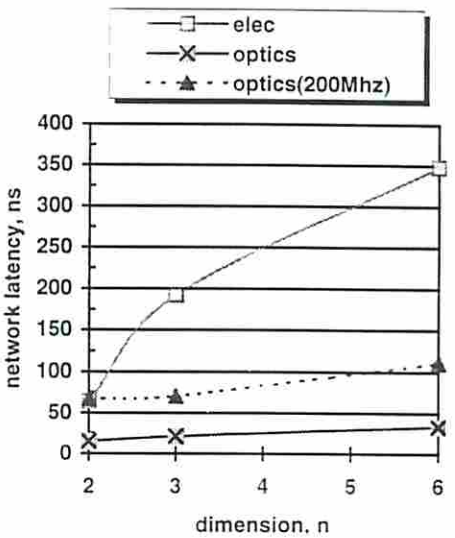
Channel cycle time and its components for both optical and electrical interconnects are shown in Figure 18(a). Every component in $T_{C\text{-}elec}$ grows with dimension whereas only the propagation delay does so in $T_{C\text{-}optics}$. This makes the channel cycle time in optics grow much more slowly. Although routing in the third dimension increases the propagation distance, the effect is negligible.



(a) Channel cycle time and its components (electronics and optics).



(b) Channel width.



(c) Network latency.

Figure 18. Latency and channel width of the 64-node system.

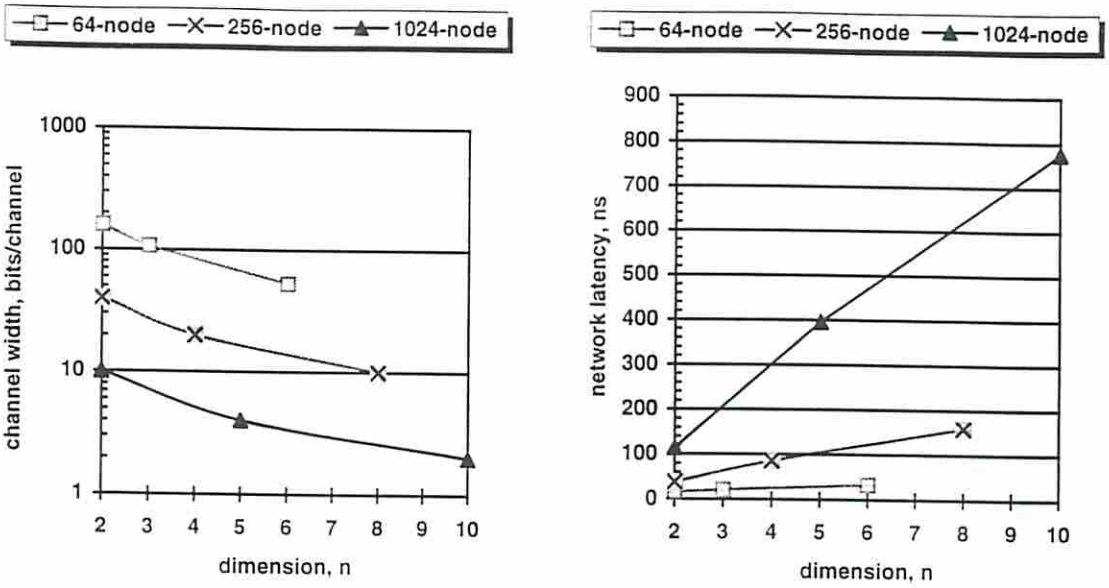
The channel width for both schemes is depicted in Figure 18(b). As the figures show, optical interconnects can be implemented more compactly with lower latency for all topologies shown here. Optics' wider channel width makes network latency less dependent on message length even for higher dimensions. Together with wormhole switching, which makes hop distance have even less of an impact on network latency,

optical interconnects are closer to achieving constant minimal network latency for various k-ary n-cube configurations as is shown in Figure 18(c).

When the channel cycle time is determined by internal router delay, an optical system benefits only from wider communication channels. This effect is shown by *optics(200MHz)* curve in Figure 18(c) assuming channel cycle time is fixed by internal router delay (T_C = 5ns, e.g., Intel Teraflop [6]). In this case, optics is still more than two times faster than electronics for the binary 6-cube. This trend becomes more pronounced for lower clock rates (i.e., slower routers).

The performance of larger networks is also studied as shown in Figure 19, again assuming 10% of connection capacity are used for data lines. Due to the problem of free-space packaging, larger networks are assumed to be implemented on a plane that is the same size as the TX-RX plane ($12cm$x$12cm$) with smaller die for each node. This implies that each node has fewer connections (microlens diameter does not change). Channel cycle time and volume requirements are smaller in low-dimensional networks because nodes are located closer to each other. High-dimensional networks do not benefit from this configuration because the maximum connection path remains unchanged.

Network latency becomes unmanageably large for massively large networks, especially in high-dimensional networks, because of the rapid decrease in channel width. However, low-dimensional networks, as shown in Figure 19(b), still achieve tolerable network latency. This suggests that optical networks are moderately scalable and the 3-D connection ability in optics becomes less beneficial in networks that are larger than 256 nodes.



(a) Channel width

(b) Network latency

Figure 19. Channel width and network latency for 64-, 256-, and 1024-node systems.

3.4 Other Considerations

Two other important considerations in implementing free-space optical interconnects are power dissipation and packaging tolerance. From a performance perspective, an insight into discuss how they affect the network latency of optical interconnects is discussed. In this DROI study, power dissipation and cooling capability for current technology put limits on the channel width. Misalignment in system packaging leads to larger transmitter and/or receiver microlenses. Given that the power dissipation, cooling capability, and interconnection area are known, the channel width, network latency, and packaging tolerance of a system can be determined.

3.4.1 *Power Dissipation*

Power dissipated as heat generated by VCSELs is given by the following:

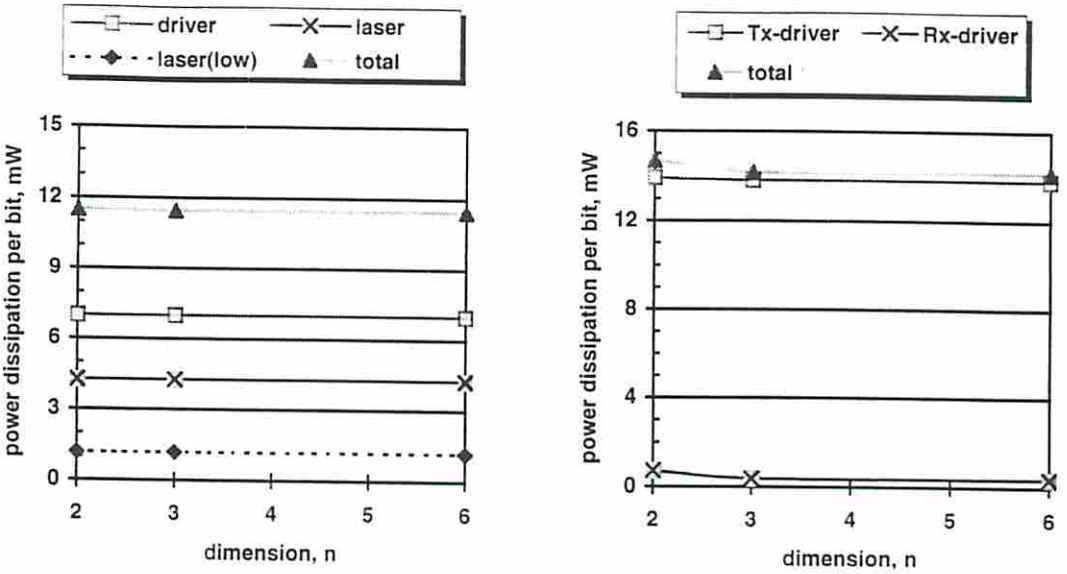
$$P_{laser} = P_{th} + P_o\left(\frac{1-\eta_{VCSEL}}{\eta_{VCSEL}}\right), \tag{22}$$

where P_{th} is the threshold power, P_o is the optical output power, and η_{VCSEL} is the slope efficiency (W/A). Similarly, the power dissipation of an electronic circuit is given by

$$P_{elec} = \frac{1}{2T_c}CV^2, \tag{23}$$

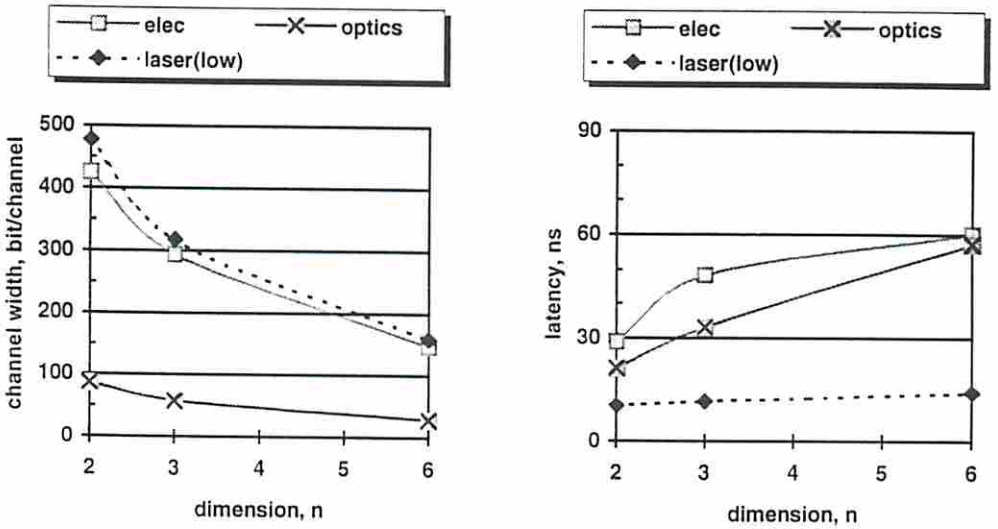
where C is the total load capacitance, V is the supply voltage, and T_C is the channel cycle time. Therefore, the overall power dissipation per optical channel takes into account the heat generated by the VCSEL, laser driver, and receiver circuit. Each electrical channel includes only the heat generated by the transmitter (line driver) and receiver circuits.

Electrical interconnects dissipate most of the power to drive the transmission line while optical interconnects dissipate most of the power on the VCSELs. This dissipation is so large that it dominates the other components. Figure 20 shows that power dissipation per bit for both technologies is virtually independent of topology. It is obvious why optical interconnects maintain constant power dissipation by examining Eq.[22]. In contrast, Eq.[23] shows that power dissipation in electrical interconnects depend on capacitance (mostly line capacitance) and channel cycle time. Because these quantities change with topology at almost the same rate and have an inverse relationship, power dissipation remains virtually constant. Although both electrical and optical interconnects share the same power dissipation trend, electrical interconnects have about 20% more power dissipation. This gap is likely to become larger in the near future due to the emergence of low-threshold, high-efficiency VCSELs which have significantly lower power dissipation as given by the *laser(low)* curve in Figure 20 [60].



(a) Optical interconnects. (b) Electrical interconnects.

Figure 20. Power dissipation of the 64-node optical and electrical interconnects.



(a) Channel width. (b) Network latency.

Figure 21. Latency and channel width of the 64-node system with 2W/cm^2 cooling capability.

To see the effect of power dissipation on network topology, a cooling capability of 2 W/cm^2 interconnection area was assumed. The number of I/O channels is constrained by cooling capability as well as by connection capacity or bisection width. Therefore, a

smaller channel width and larger network latency, as shown in Figure 21, are expected. Only chip areas are used for cooling which are $64cm^2$ for optics and $400cm^2$ for electronics, and 100% of connections available are used for data lines.

Electronics provides a wider channel width due to its larger chip area to remove heat. Although this is the case, network latency in optics still surpasses that of electronics because of optics' lower channel cycle time. Thus, when power dissipation and cooling capability are considered, optical interconnects still have a performance advantage but not as much as when connection capacity is the only consideration. However, it must be emphasized again that progress in optoelectronic and packaging technologies will enhance the performance advantages of optics as given by the *laser(low)* curve.

The *laser(low)* curves in Figure 20 and 21 represent the power dissipation, channel width, and network latency of low threshold VCSELs [60]. These VCSELs have the following characteristics: threshold current of 212 μA at 1.755 operating voltage, 1.9mW maximum output, and 55% slope efficiency. For a 1mW output, the power dissipation is only 1.18mW/VCSEL, about a quarter of the commercially available VCSELs'. This technological breakthrough allows not only denser optical channels but also faster conversion time. Therefore, wider channel width and faster clock rates are expected. In addition, better cooling capability will similarly enhance system performance.

3.4.2 Packaging Tolerance

Section 3.4.1 shows that the power dissipation for commercially available technology places more stringent limitations on channel width than the connection capacity of an optical interconnect. This section uses the channel width given in Section 3.4.1 to further evaluate the allowable packaging tolerances corresponding to each type of misalignments in a DROI k-ary n-cube system.

Three common misalignments exhibited in free-space optical systems are shown in Figure 22. The first is lateral misalignment, which is a horizontal misalignment between the transmitter-receiver plane and the microlens-hologram plane. The second is longitudinal misalignment, which is a vertical misalignment between the microlens-hologram plane and the mirror plane. The last is angular misalignment of the mirror plane to the horizontal plane. The effect of each type is somewhat similar to each other.

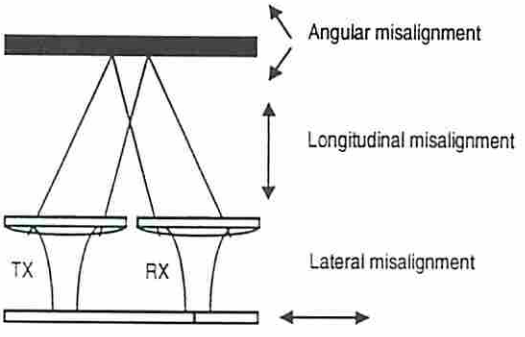


Figure 22. Three types of misalignment in DROI systems.

1) Lateral Misalignment: To maintain a microlens efficiency of 99.5%, the microlens diameter must be four times larger than the spot radius. Since this misalignment shifts both transmitters and receivers laterally, it is necessary to fabricate larger transmitter and receiver microlenses to satisfy this condition. The lateral misalignment can be expressed as

$$\Delta x = \Delta_{Lat}, \tag{24}$$

where is Δx the lateral shift of the optical beam respect to the microlens-hologram plane and is Δ_{Lat} the lateral misalignment. For the assumptions given in Section 3.4.1, the system supports a maximum of 1,936 microlenses/cm^2 each with a diameter of 227μm. Compared to the correctly aligned microlens with 125μm diameter, it is possible to tolerate Δ_{Lat} = 102μm.

2) *Longitudinal Misalignment:* Misalignment either upward or downward with respect to the microlens-hologram plane results in a lateral shift of optical beams at the receiver microlenses given by

$$\Delta x = 2\Delta_{Long} \tan\theta, \tag{25}$$

where Δx is the lateral shift of the optical beams, Δ_{Long} is the longitudinal misalignment, and θ is the maximum hologram deflection angle. The effect of this misalignment is somewhat different from the previous one, as it requires only larger receiver microlenses. By keeping transmitter microlenses at $125\mu m$ diameter, the receiver microlenses can be enlarged to $330\mu m$. From Eq.[25], it is possible to tolerate $\Delta_{Long} = 230\mu m$.

3) *Angular Misalignment:* Similarly, the angular misalignment of the mirror plane results in a lateral shift of optical beams at the receiver microlenses given by (for small angle approximation),

$$\Delta x = 2h\Delta_{\theta}, \tag{26}$$

where Δx is the lateral shift of the optical beams, h is the mirror to microlense-hologram plane separation, and Δ_{θ} is the angular misalignment. For a height of h = $6.77cm$ (the 64-node 2-ary n-cube network) and receiver microlenses with $230\mu m$ diameter, the system can tolerate $\Delta_{\theta} = 7.75\text{x}10^{-4}$ radian or only 0.044°!

3.4.3 Wavelength Variation

Uniformity of optoelectronic devices is not easily achieved, particularly VCSEL arrays. Although other VCSEL parameters such as threshold current and threshold voltage are also non-uniform, the wavelength variation causes the most severe performance degradation. This issue is analyzed below.

Wavelength variation affects two things in DROI system: the spot radius at both transmitter and receiver microlenses due to Gaussian beam propagation (see Appendix A) and the lateral shift of optical beams with respect to the receiver microlenses. To accommodate both changes, larger microlenses are needed to maintain efficiency. Since the variation has only minor effects on spot radius, only the lateral shift of optical beams is addressed here.

The hologram deflection angle changes proportionally to the wavelength (see Appendix B) and can be expressed as (for small angle approximation)

$$\Delta\theta = \frac{\Delta\lambda}{n_x T}, \tag{27}$$

where $\Delta\theta$ is the deflection angle sensitivity, $\Delta\lambda$ is the wavelength variation, n_x is the refractive index of material, and T is the hologram period. This deflection angle sensitivity resembles the angular misalignment previously discussed and causes a lateral shift between the optical beams and the receiver microlenses:

$$\Delta x = \frac{2h\Delta\lambda}{n_x T}. \tag{28}$$

Assuming a receiving microlens with $230\mu m$ diameter and a 64-node 2-ary n-cube network, a wavelength variation of $0.8nm$ is allowed.

These results clearly show that progressing optical interconnects has a few important, but not insurmountable obstacles to overcome before they can be widely employed. Free-space optics' small volume makes both heat removal and packaging much harder than with electronics. Together with device nonuniformity such as wavelength variation, these obstacles can considerably reduce the performance advantage of optical interconnects. This is why free-space optics are not likely to be implemented in a large single volume in the near term. Once these issues have been substantially improved, free-space optics will certainly become practical for multiprocessor interconnects. Recent

research [61] shows a significant improvement in VCSEL technology that has a yield of 99.8% on a 3-in wafer and uniformity of better than ±9% in threshold current, ±1% in threshold voltage, and ±1.5% in maximum optical output power.

3.5 System-level Integration: Is It Feasible?

The results from the analytical model for optical interconnects introduced in this chapter clearly show that optical networks can support much wider communication channels compared to a similarly configured electrical network. A misconception about signal conversion time (electrical to optical and vice versa) is also elucidated here; the conversion times are not a bottleneck in optical interconnects as long as the imaging system is efficient. This is usually the case using current technologies. Altogether the network latency in an optical network is dramatically less and, due to the connection capacity cost model of free-space optical interconnects, that latency is less sensitive to the network topology. In other words, optical networks provide more design flexibility while retaining the desired network performance.

There are several issues relating to system-level integration to be solved. Identified and evaluated here are power dissipation, packaging tolerance, and device uniformity. Note that none of them are (or at least have not reached) the fundamental limits for the construction of optical networks. Since optoelectronic and related technologies are progressing at an impressive rate, those issues are becoming less of a concern. For instance, ultra low current VCSELs [60] with very high device uniformity are being developed [61]. Alternatively, passive devices such as SEEDs can be used as both transmitters and receivers to reduce on-chip heat dissipation. Therefore, more devices can be integrated on a chip without exceeding the system cooling capability.

The success of Micro-Opto-Electro-Mechanical Systems (MOEMS) [62] has significantly simplified the packaging tolerance issue. This technique makes the fabrication and alignment of micro-optic apparatus on a chip more feasible and accurate.

The scalability disadvantage of the optical network assumed in this study is not a fundamental limit either. This only applies to a tightly integrated system like the configuration assumed here. For a geographically separated system, such as board-to-board interconnections or beyond, each node can be ruggedly and compactly packaged and connected through free-space or fiber ribbons. This packaging paradigm provides both better scalability and less susceptibility to heat dissipation and, hence, features a very dense array of optoelectronic devices. Examples are the optical module designed at McGill University [63] supporting 1024 (32x32 array) SEEDs via free-space optical interconnects, and the AMEOBA switch [16] module supporting 1024 (32x32 array) SEEDs via Dense Wavelength Division Multiplexing (DWDM) at hundreds of meters interconnection range.

In conclusion, the optical network has potential advantages over the conventional electrical network. System-level integration appears to be feasible using current technologies and should be further pursued. To achieve the potential performance provided by an optical network, network components such as switches and routers must be sufficiently sophisticated to utilize the high-bandwidth and low-latency of optical interconnects. Hence, the next question then becomes—can complex optoelectronic chips be built? This issue is addressed in the following chapters.

Chapter 4

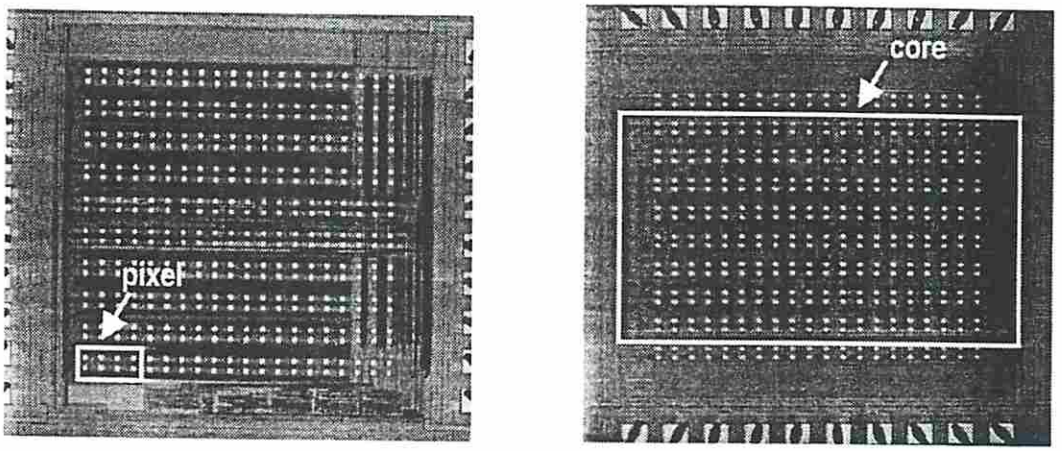
Design Issues for Optoelectronic Chips

Optical interconnects have been successfully employed in long-haul communication systems. Recent progress in optoelectronic technology and increasing demand in bandwidth together have brought optical interconnects down to a shorter-distance interconnection range. From a computer architect perspective, optics provides ample bandwidth that can be exploited at the chip-to-chip interconnect level up to the local- or campus-wide network level. In order to realize the concept, the integration of CMOS circuits and optoelectronic devices on a single chip must be developed with clearly defined functions; the former handles computations and controls whereas the latter handles off-chip communications. This chapter particularly focuses on CMOS/SEED integrated technology (Appendix C.2) in which CMOS-VLSI circuits are flip-chip bonded to an array of GaAs-based SEEDs.

4.1 Pixel-based versus Core-based Designs

Early attempts at the integration of CMOS circuits and optoelectronic devices consisted of a small circuit connecting to SEEDs forming *pixel-like* circuit called "smart-pixels." A pixel is usually replicated throughout the entire chip to form a 2-D array of smart-pixels. These "pixel-based" designs are very common for non-sophisticated high-parallelism applications such as signal processing [21], bit-slice arithmetic logic unit (ALU) [22], or simple switch [23]. Generally, this type of design can be efficiently optimized because it

is typically small and self-contained with most wires being local and usually fits within the pitches of the SEED array. Figure 23(a) shows a pixel-based chip where a pixel is depicted in a rectangular box.



(a) A pixel-based design
(the TRANSPAR chip—courtesy A. Sawchuk et al. [64]).

(b) A core-based design
(shown is the WARRP II chip).

Figure 23. Comparison of CMOS/SEED chip designs.

Pixel-based designs, however, are neither sophisticated nor powerful enough to implement complex chips such as network routers. With the continuous progress in CMOS technology and the introduction of more powerful processors, off-chip bandwidth is becoming a critical factor to achieve higher system performance. Therefore, optoelectronic integration of such bandwidth-sensitive chips is being actively investigated. In light of this significance, recent CMOS/SEED designs tend to integrate large and complex CMOS circuitry with the SEED array, referred to as Level-5 genius pixels [24]. Such designs are not necessarily pixel-like, in fact, they tend to be a tightly integrated core and, hence, are called "core-based" designs and are shown in Figure 23(b).

4.2 Issues in Core-based Designs and Their Effects on Chip Performance

Although core-based designs have potential usefulness for high-bandwidth high-performance applications, they also raise critical design issues that must be addressed. First, the large number of SEED transceivers must be integrated with the CMOS core. Second, the CMOS core I/Os are not perfectly aligned with the SEED array. This makes it difficult to connect those I/Os to a regularly distributed SEED array. Third, at least the top metal layer must be reserved exclusively for SEED wiring. Fourth, the I/Os should be laid on the dies in a structured pattern in order for the chips to be connected by a space-invariant optical system because the locations of transmitter-receiver pair are interdependent. Fifth, to achieve production-level yield, the SEED array is limited to a $3.7 \times 3.7 mm^2$ area [66], which is typically smaller than that of a complex CMOS-VLSI circuit. Thus, optoelectronic I/O pin-outs could only be located over a specific area of the chip. Note that these issues are exclusive to core-based designs only because pixel-based designs are much easier to optimize and fit in a structured array under the SEEDs.

The consequences of this are considerably long wires between SEED transceivers and CMOS I/O ports and/or SEED bonding pads. These wires take away some of the wiring resources (e.g., metal layers) from the CMOS core and effectively decrease the number of wireable transistors on the chip. Longer connections between the bonding pads and the receivers decrease the signal-to-noise ratio at the receivers (which are analog circuits) thereby requiring more optical power and/or achieving slower detection rates. Moreover, longer wires can decrease the achievable on-chip clock rates because they may lie in the signal critical paths. In conclusion, core-based CMOS/SEED chips tend to decrease the

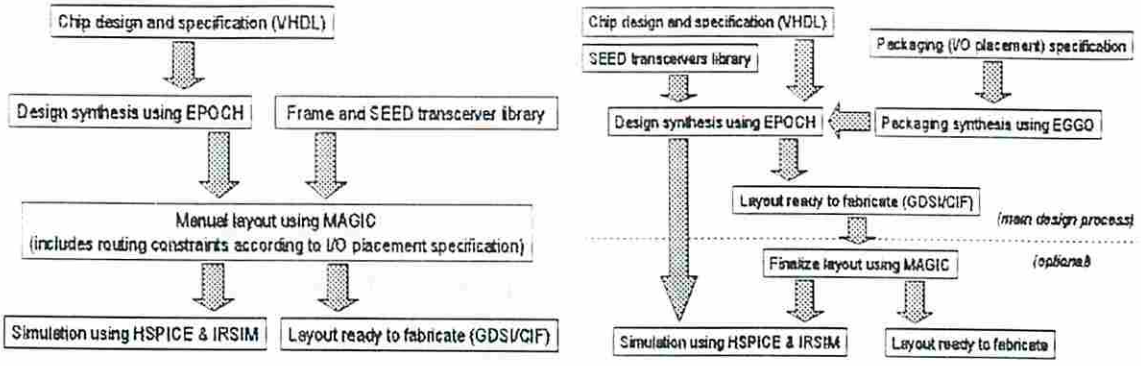
transistor density and the on-chip clock rates in exchange for more I/O pin-outs. These issues are collectively referred to as the wiring problem[3] of core-based designs.

High-performance BGA packaging, on the other hand, also features area-distributed I/Os (called "balls") for the CMOS chips. It may seem that some issues related to core-based designs should apply to this packaging technique as well. In fact, none of them do because of the following: First, it requires a transceiver that is already compatible with CMOS which can be seamlessly integrated with the core circuitry. Second, each transceiver can be wired to any of the nearest balls. Third, metal layers are not exclusively used for ball wiring. Fourth, the ball array can be distributed throughout the entire chip area. Nevertheless, this packaging technique still provides fewer I/O pin-outs compared to CMOS/SEED integration technique (see Table 1).

It is obvious that to effectively design the core-based chips, optoelectronic-compatible CAD tools are required which automatically integrate the SEED array, the transceivers, and the CMOS core. Since there were no such tools available for this investigation, prior core-based designs were carried out by manual integration which exploited only part of the SEED array as in WARRP II (Figure 23(b)). In designing the WARRP II chip, only 20% of the SEEDs was used as optical I/Os whereas the remaining 80% was occupied by the CMOS circuitry (details will be discussed in Section 5.4). Others have simplified the wiring problem by placing the SEED array on top of the CMOS core and the transceivers on the periphery [28]. Alternatively, the SEED array was placed on top of the transceivers and the CMOS core was placed on the periphery [29]. Both techniques result in very long wires connecting between the CMOS I/O ports, the SEED transceivers, and the SEED array, and very low chip area utilization.

---

[3] Throughout this dissertation, the term "wiring" is used in lieu of "routing" commonly used in circuit designs. This is to distinguish from the routing function in network routers.



(a) Manual integration. (b) Automatic integration using EPOCH/EGGO tools.

Figure 24. Suggested design process of complex CMOS/SEED chip.

Recent development of EGGO CAD tool, a supplemental area-distributed I/O packaging tool to EPOCH, has allowed the efficient and automatic integration of the SEED array and complex CMOS circuitry. Figure 24(a) shows the manual integration process, which CMOS circuit is synthesized by EPOCH and later on is manually placed and wired to the SEED array and transceivers using MAGIC. In contrast, the automatic integration process simultaneously places the standard cells and the transceivers according to the SEED I/O assignments and later on routes the entire design automatically, as shown in Figure 24(b). This makes the MAGIC tool optional. In other words, the automatic integration process fully utilizes both the SEED array and the area underneath, making CMOS/SEED integration more appealing to complex designs. An example of complex optoelectronic chips that utilizes automatic integration is a DSP core [30].

To the best of my knowledge, the area of core-based CMOS/SEED design is relatively new and very few aspects have been investigated, especially performance modeling and estimation. Performance estimation is an important part of the design process because the design can be reiterated to meet cost/performance requirements without actually synthesizing the layout, which is computationally intensive and time-

consuming. This dissertation, therefore, aims to establish a performance model for core-based optoelectronic chips based on EPOCH/EGGO CAD tools.

4.3 Wiring Capacity and Wiring Cost Models

The model is established by considering the wiring resource available to and taken by SEED integration. Wiring resource considered here are the number of metal layers available. In core-based design, CMOS I/O ports are wired to the transceivers that are subsequently connected to the bonding pads. These pads are simply square pieces of the top metal layer required to attach to the SEED array via flip-chip bonding. Once the standard cells and transceivers are synthesized by EPOCH, the bonding pads are placed and wired by EGGO. Since EGGO does not support rip-up of any previous wires, at least the top metal layer must be reserved exclusively for the SEED wiring[4]. Additional metal layers may be required to complete the wiring of large SEED arrays. For small arrays (e.g., 20x10 devices as in the WARRP II chip), just using the top metal layer is sufficient.

4.3.1 The Models

X- and *y*- wiring styles as used in EPOCH are assumed, each metal wiring layer is associated with either the *x*- or *y*- direction. In addition, the SEED array is large enough to utilize *at least* two metal wiring layers and there are sufficient metal layers to wire both the CMOS core and the SEED array. These assumptions are reasonable considering the potential core-based designs as will be shown in Section 4.4. The notion of "wiring capacity" is used to represent the number of wires that can be placed per unit, where each unit is the area surrounding each SEED as shown in Figure 25. The wiring capacity available in *x*- and *y*-directions can be written as:

---

[4] The SEED wiring is the wiring from the transceivers to the bonding pads.

$$X_C = K_i \cdot \left(\frac{Y_{pitch}}{m_{X-pitch}} \right) \cdot D \tag{29}$$

$$Y_C = K_j \cdot \left(\frac{X_{pitch} - P}{m_{Y-pitch}} \right) \cdot D \tag{30}$$

Where D is the total number of SEEDs, P is the bonding pad size, K_i and K_j are the wiring utilization for metal layer i and j, X_{pitch}, Y_{pitch}, $m_{X\text{-}pitch}$, and $m_{Y\text{-}pitch}$ are the pitch of SEED and the pitch of metal layer in x- and y-directions, respectively. Also m_X is the top metal layer to be used as the bonding pads and m_Y is the subsequent metal layer under m_X and thus it can use all the SEED area to place wires. Wiring utilization is defined as the ratio of the area used to wire signals over the entire design core area, for a given metal layer. This parameter adjusts the model to accurately represent the wiring performance of the real designs. The methodology to find wiring utilization is explained in the next section.

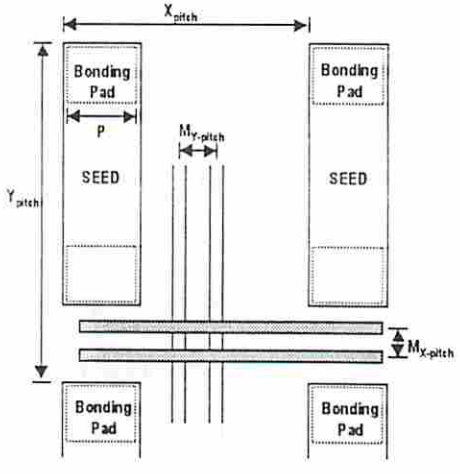


Figure 25. SEED placement and wiring assumptions.

To calculate the wiring cost, all signals are assumed to be dual-rails and the SEED array is split equally into two groups of transmitters and receivers that are placed symmetrically to the optical axis (parallel to x-axis is assumed). In the worst case where

CMOS I/O ports are randomly placed with respect to the bonding pads, the wiring cost for the entire SEED array in x- and y-directions can be written as:

$$X_R = \frac{D}{2} \cdot \frac{D_x}{2}, \tag{31}$$

$$Y_R = \frac{D}{2} \cdot \frac{D_y}{4}. \tag{32}$$

Where D_x, D_y is the number of diodes in x- and y-directions, respectively. The ratios of Eq.[29]/Eq.[31], and Eq.[30]/Eq.[31] simply determine if additional metal layers are required in order to wire the SEED array. If so, the procedure is repeated until both x and y wiring costs are covered by the wiring capacity with P in Eq.[30] disappearing since bonding pads are no longer a wiring constraint.

4.3.2 *Methodology to find Wiring Utilization*

Since several interdependent factors affect wiring performance of the CAD tools, it is impractical to establish an analytical model for wiring utilization. Here, this parameter is estimated by exhaustively synthesizing the WARRP router and varying its features, ranging from about 10,000 to 50,000 transistors in complexity. In addition, different versions of EPOCH with different technology files (or ruleset in EPOCH terminology), HP14B and AMI0.6μm, were used to capture the effects of the synthesis tool and fabrication technology on wiring utilization. Note that both rulesets are actually 0.6μm drawn feature size with the three metal layers. In most cases, the achieved layouts are comparable in size. Experimental configurations are summarized in Table 7.

Table 7. Experimental configurations of the WARRP router.

Architecture	Configuration	#transistors
WARRP II	Ring topology, unidirectional 4-bit-wide channels, 1VC, 2-flit-deep buffers	~11,000
WARRP III	2-D torus topology, unidirectional 4-bit-wide channels, 3VCs, 2-flit-deep buffers	~50,000
WARRP IV	WARRP II with unidirectional 16-bit-wide channels	~27,000
WARRP V	WARRP II with unidirectional 32-bit-wide channels	~48,500

The syntheses without EGGO were performed *to find the effects of SEED wiring in manual integration*. To do so, the designs were synthesized with 2 and 3 metal layers. Three-metal-layer layouts signify the pure-CMOS designs whereas two-metal-layer layouts presumably represent the core-based designs. This assumption holds because all designs employ less than 100 optical I/Os and can be wired using only Metal–3.

Table 8. 2-metal-layer layout characteristics synthesized by EPOCH.

Designs	Die size (mm^2)	TX density (TXs/mm^2)	TX area (%)	WU-m1 (%)	WU-m2 (%)	Critical path (ns)
0.6um (AMI) WARRP II	0.91x1.13 = 1.03	10,505	33	80	43	11.182
0.5um (HP14B) WARRP II	0.83x1.24 = 1.03	10,505	26	78	50	11.536
0.6um (AMI) WARRP III	2.04x3.01 = 6.13	8,111	26	84	84	15.052
0.5um (HP14B) WARRP III	2.19x2.91 = 6.37	7,806	26	84	68	12.256
0.6um (AMI) WARRP IV	1.67x1.64 = 2.74	9,774	33	85	51	14.325
0.5um (HP14B) WARRP IV	1.50x1.78 = 2.67	10,064	28	85	60	12.086
0.6um (AMI) WARRP V	2.37x2.03 = 4.79	10,107	34	89	58	16.574
0.5um (HP14B) WARRP V	2.13x2.47 = 5.26	9,204	28	84	67	14.550

Table 8 shows the result from synthesizing the designs with 2 metal layers. Critical paths were measured by using a timing simulator called TACTIC in EPOCH tool. Different technology does not have a significant impact on the layout characteristics of each configuration, i.e., less than 20% difference for all cases. Except for WARRP III, which features more complex architectures. The rest of the designs all feature a similar core with a wider datapath. Wiring utilization of Metal–2 (WU-m2) shows that the wider the datapath, the higher the wiring utilization. This suggests that WARRP III has the largest datapath circuits compared to others while WARRP II has the smallest. The average WU for Metal–1 and –2 are 83.63% and 60.13%, respectively. This is not surprising because EPOCH employs "cell-based" wiring style in which wiring in horizontal direction (Metal–1) is preferred over vertical direction (Metal–2). The most complex control circuitry incorporated in WARRP III results in the largest die area and

the lowest transistor density because this type of circuit cannot be as well optimized as the datapath.

Table 9. 3-metal-layer layout characteristics synthesized by EPOCH.

Designs	Die size (mm^2)	TX density (TXs/mm^2)	TX area (%)	WU-m1 (%)	WU-m2 (%)	WU-m3 (%)	Critical path (ns)
0.6um (AMI) WARRP II	0.86x0.82 = 0.70	15,714	44	66	57	44	11.042
0.5um (HP14B) WARRP II	0.84x0.82 = 0.69	15,751	34	70	39	51	9.325
0.6um (AMI) WARRP III	2.01x1.95 = 3.91	12,717	40	64	83	66	14.210
0.5um (HP14B) WARRP III	2.18x1.98 = 4.31	11,536	41	74	57	68	11.416
0.6um (AMI) WARRP IV	1.38x1.32 = 1.82	14,765	47	68	70	58	14.482
0.5um (HP14B) WARRP IV	1.50x1.20 = 1.80	14,929	38	73	47	56	11.394
0.6um (AMI) WARRP V	1.92x1.83 = 3.51	13,793	46	69	83	68	16.511
0.5um (HP14B) WARRP V	2.10x1.67 = 3.51	13,793	39	76	57	65	13.682

The results of pure-CMOS syntheses using 3 metal layers are shown in Table 9. As expected, the availability of Metal-3 significantly reduces the die size, increases the transistor density, and reduces the critical path. It also reduces WU-m1 because EPOCH now has two horizontal wiring layers and tries to distribute the wires more evenly. The effects of excluding Metal-3 required by SEED wiring are summarized in Table 10.

Table 10. Effects of SEED wiring on the layouts.

Die size (%)	TX density (%)	WU-m1 (%)	WU-m2 (%)	WU-m3 (%)	Critical path (%)
+48.97	-32.87	+16.29	+3.74	-100.00	+4.25

From Table 10, the layout is enlarged by almost 50% and the transistor density is reduced by about 33%. One might think that these results should not hold because the transceivers are not included in the layout. However, such circuits typically consist of a simple inverter for the transmitter and a small analog amplifier for the receiver and, therefore, they account for less than 5% of the total transistors in the experimental configurations.

The effect on critical path requires additional investigation to be fully understood. Critical path is increased by two factors: larger die (as a result of less metal layers) and additional wires required to connect between the CMOS I/Os to the SEED transceivers and to the bonding pads. The results obtained in this experiment reflect the first factor (larger die) that is 4.25% longer than the pure-CMOS layouts.

The second experiment was performed based on EPOCH/EGGO tools *to quantify the effect of additional wires between the transceivers and the bonding pads on critical path, and to compare the wiring utilization obtained by the manual integration.* The ability to automatically integrate the SEED array on the CMOS core significantly increased the synthesis time and memory requirements. Only the WARRP II and WARRP III designs were synthesized successfully.

Table 11. Layout comparison between core-based designs w/ and w/o SEED integration.

Designs	Die size (mm^2)	TX density (%)	WU-m1 (%)	WU-m2 (%)	Critical path (ns)	Increased critical path (ns)
WARRP II (w/ SEEDs)	1.22x1.18 = 1.44	31	78	54	13.430	1.894
WARRP II (w/o SEEDs)	0.83x1.24 = 1.03	26	78	50	11.536	–
WARRP III (w/ SEEDs)	2.77x1.85 = 5.11	30	88	64	14.995	2.739
WARRP III (w/o SEEDs)	2.04x3.01 = 6.13	26	84	68	12.256	–

Surprisingly, all layout characteristics were very close to that of 2-metal-layer layouts in the previous experiment except the critical path, as shown in Table 11. The increased critical path corresponds to the second factor mentioned above—additional wires from the CMOS I/Os to the bonding pads—and is around 19% on the average. These wires were specified to be within 400μm wiring window in the experiment (this number should scale proportionally with optoelectronic and CMOS technologies). Of course, a smaller wiring window can be specified to reduce the critical path length but it would increase the synthesis time and may not allow the wiring of all the SEEDs. Other values including 100μm, 200μm, and 300μm were attempted but all failed to complete the SEED wiring.

The total synthesis times for the WARRP II and WARRP III designs were 15 minutes and 135 minutes, respectively.

The results from both experiments are not sufficient to predict the performance of large core-based designs where more than one metal layer is required for SEED wiring. However, the results do show some trends that could be used to extrapolate the parameters and effects of new rulesets that would support more than three metal layers. Additional metal layers are assumed to yield the wiring utilization between 60%–75% range. Each metal layer is also estimated to account for circuit wiring proportionally to its wiring utilization and inversely to its minimum pitch. For example Metal–2 would affect the transistor wiring more than Metal–3 as it has finer pitch and higher wiring utilization. The effect of larger die on the critical path is assumed to be 5% per metal layer and is accumulative whereas the effect from additional wires is 20%. Note that the effect from additional wires is quite accurate. The performance results of the WARRP II chip (see Section 5.4) show that the longest additional wire is approximately 1500μm long (about 4 times the length assumed here) and decreases the achievable clock rate by half of the original design with three metal layers. According to the model, the critical path would increase by 85% (5% from increased die size and ~80% from additional wire to the bonding pad). The 15% error should account for the connections that were wired using the Polysilicon layer.

4.4 Core-based Optoelectronic Chip Performance Estimation

This section uses the model to predict the performance of core-based optoelectronic chips in comparison with pure-CMOS chips based on published semiconductor technology trends from SIA [32] and CMOS/SEED integrated technology trends from Krishnamoorthy [34]. The performance prediction provides the missing piece of information between the two technological trends. This information justifies core-based

designs and validates the expected performance advantage promised by this technology. All data fields in Table 12 are excerpted from [32, 34] except the SEED pitches. This information was not available so the densest 64x32 SEED array on a 2.3x2.3mm^2 area available in 1997 was taken to be the base value and a 20% pitch reduction per generation was assumed. This number reflects a similar 20% size reduction of the bonding pad corresponding to the total number of SEEDs available on-chip predicted by Krishnamoorthy.

Table 12. Semiconductor and optoelectronic SEED technology roadmaps.

Year of first shipment	1999	2001	2003	2006	2009
Technology (μm)	0.18	0.15	0.13	0.10	0.07
Transistor Density (per mm^2)	140,000	160,000	240,000	400,000	640,000
On-chip Local Clock (MHz)	1250	1500	2100	3500	6000
Off-chip Clock (MHz)	480	785	885	1035	1285
# Package Pin-outs (pin)	1500	1800	2200	3000	4100
Aggregate Bandwidth (GB/s)	225	315	440	750	1281.3
Maximum Wiring Layers	6-7	7	7	7-8	8-9
Minimum Contacted Pitch (μm)	0.46	0.40	0.34	0.26	0.19
# SEEDs (per chip)	8000	12000	20000	35000	47000
Bonding Pad size (μm)	9	8	7	5	4
SEED x- and y-pitches (μm)	29,58	23,46	18.5,37	15,30	12,24

The wiring utilization for horizontal and vertical metal layers is assumed to be 75% and 60% (in accordance to the results obtained in previous section), respectively, and each additional metal layer has a 20% pitch increase. For example, Metal–2 yields 60% wiring utilization and 0.54μm contacted pitch; Metal–3 yields 75% wiring utilization and 0.66μm contacted pitch, and so on, with 0.18μm technology in 1999. Also, part of metal layers left from SEED wiring cannot be used for CMOS core wiring. Using Eq.[29-32], the wiring capacity, wiring cost, and number of metal layers required for SEED wiring, in x- and y-directions can be estimated.

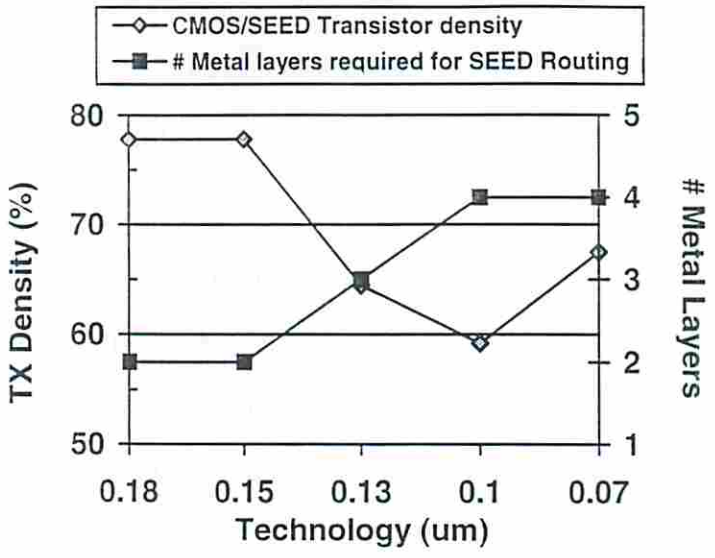
With some number of metal layers excluded, the wireability of the transistor decreases and, thus, effectively reduces the transistor density per unit area. As mentioned earlier, each metal layer is assumed to contribute to circuit wiring proportionally to its

wiring utilization and inversely proportional to its minimum pitch, excluding Metal–1 that is used exclusively for power and local wiring. For instance, the wiring contribution of Metal–3 is 40% for HP14B and AMI0.6μm technologies.

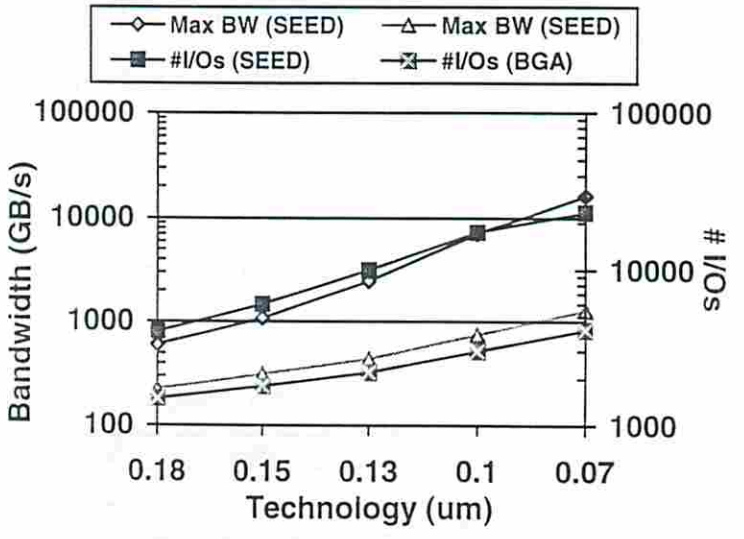
In summary, the model predicts that core-based designs require between 2 and 4 metal layers to complete the SEED wiring and reduce the transistor density by as much as 41%. The performance predicted by the model is summarized in Table 13 and plotted in Figure 26. In general, the number of metal layers required for SEED wiring increases, going hand-in-hand with the increasing number of SEEDs. In contrast, the transistor density keeps falling but rises again at 0.07μm technology because an additional metal layer is available. Although the on-chip clock rates of the CMOS/SEED chip can be reduced by almost 30% due to their increased critical paths, the more important off-chip clock rates can be as high as the core circuit. This assumption is not overestimated, as many believe that SEEDs can operate at much higher rates than the core circuit. In addition, at higher number of available I/O pin-outs, CMOS/SEED chip is capable of 2 to 10 times higher aggregate off-chip bandwidth. From the bandwidth perspective, CMOS/SEED chip is more appealing as an intermediate-term solution to the bandwidth problem.

Table 13. Performance comparison of complex CMOS/SEED and CMOS/BGA chips.

Year of first shipment	1999	2001	2003	2006	2009
Technology (μm)	0.18	0.15	0.13	0.10	0.07
# of Metal Layers Required (x, y)	1,1	1,1	2,1	2,2	2,2
Normalized Transistor Density	0.778	0.778	0.645	0.592	0.675
Normalized On-chip Clock	0.768	0.768	0.737	0.706	0.706
Normalized Aggregate Bandwidth	2.131	2.740	4.392	7.210	9.716



(a) Effective transistor density and number of metal layers required for SEED wiring in CMOS/SEED chip.



(b) Available I/O pin-outs and aggregate off-chip bandwidth.

Figure 26. Predicted performance of complex CMOS/SEED and CMOS/BGA chips.

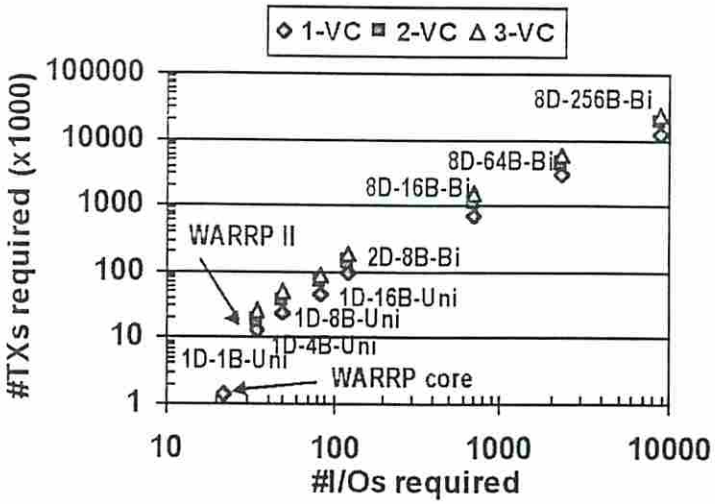


Figure 27. WARRP router complexity plotted in terms of the number of transistors and I/Os required (excluding power and ground pins), ranging from a small 4-bit-wide unidirectional-link torus with 1 virtual channel (1D-4B-Uni-1VC) to a large 256-bit-wide bidirectional-link 8-dimensional torus with 3 virtual channels (8D-256B-Bi-3VC). Most data points (up to 8D-16B-Bi-3VC) were extrapolated the results obtained by EPOCH. With current processor trends, 64-bit-wide or 256-bit-wide channels should be soon common.

The performance predicted by the model can be applied to a specific CMOS/SEED design to evaluate its performance given that necessary design information is available. This work uses the design information of the WARRP router as depicted in Figure 27. The configurations range from a simple deadlock recovery router core—the WARRP core (described in Chapter 5)—with 1200 transistors and 22 I/Os, up to a complex fully-functional WARRP router with ~20 million transistors and ~9000 I/Os.

Based upon available information, the following questions can be answered. Can this configuration be implemented? If so, what is the cost? As an example, the 64-bit-wide bidirectional-link 8-dimensional torus with 3 virtual channels configuration (8D-64B-Bi-3VC) which requires approximately 6 million transistors and 2200 I/Os is chosen. It can immediately be seen that, from Table 12 and 13, CMOS/SEED technology allows this configuration to be implemented as soon as the year 1999 with 4000 dual-rail I/O pin-outs in a 6.8x6.8mm^2 (core circuit only) chip area. In contrast, this configuration can be implemented with CMOS/BGA technology by the year 2003 or after (due to insufficient

I/O pin-outs available) with 2200 I/O pin-outs in a 5x5mm^2 chip area. Although implemented with 1999 technology, the CMOS/SEED chip still provides 36% more aggregate off-chip bandwidth than the CMOS chip implemented with 2003 technology. In other words, CMOS/SEED technology accelerates the development of high-performance designs that would otherwise be limited by the number of available I/O pin-outs.

The implementation cost often takes precedence in chip designs. The question then becomes what configurations can be implemented for a given implementation cost. For example, the design has to fit in a 6x6mm^2 chip area which supports ~4.62 million transistors (from Table 12 and Table 13), with 4000 dual-rail I/O pin-outs as available in year 1999. It is obvious that the 8D-64B-Bi-3VC configuration is not supported. In this case, we may choose to implement 3 copies of 8D-16B-Bi-3VC to realize the 8D-48B-Bi-3VC configuration instead. This option requires about 4.5 million transistors and 1800 I/O pin-outs. Since the design uses up all the available transistors, it is called "transistor-bound." On the other hand, the design that uses all the available I/O pin-outs is called "I/O-bound." High-performance designs are typically I/O-bound because the number of available I/O pin-outs increases at a much slower rates than the number of transistors on a chip.

4.5 Core-based Optoelectronic Chip Design: Is It Effective?

This chapter presents a methodology to capture the negative effects of core-based design on the chip performance. To do so, a semi-empirical model is established based on the optoelectronic compatible CAD tools—EPOCH/EGGO—to evaluate the performance of the complex CMOS/SEED chips. This model serves two main purposes. First, it provides the relevant information bridging between the CMOS and CMOS/SEED technologies. Second, it validates the expected benefits of the CMOS/SEED chips. Although some parameters of the model were obtained though the syntheses of the

WARRP router, they should generally represent other core-based designs as well. This is because the experimental configurations were chosen to represent various combinations of control and datapath circuits inherent in most complex circuits.

The results from the model show that complex CMOS/SEED chip *can* be effectively designed with somewhat of a performance penalty, which is at most a 30% decrease in on-chip clock rate. Although the transistor density is reduced by as mush as 41% during the integration, it is considered an abundant resource that is worth paying for in exchange for an order of magnitude more I/O pin-outs. In terms of off-chip bandwidth, CMOS/SEED chips are an intermediate-term solution to the bandwidth problem but, in terms of available I/O pin-outs, they are an immediate solution to the I/O limited designs. CMOS/SEED integration accelerates the emergence of I/O-bound chips (e.g., the 8D-64-B-Bi-3VC configuration can be implemented with the CMOS/SEED technology four years ahead of the CMOS chip). In some cases, it is *the only* choice of implementation in the foreseeable future (e.g., the 8D-256B-Bi-3VC configuration).

Chapter 5

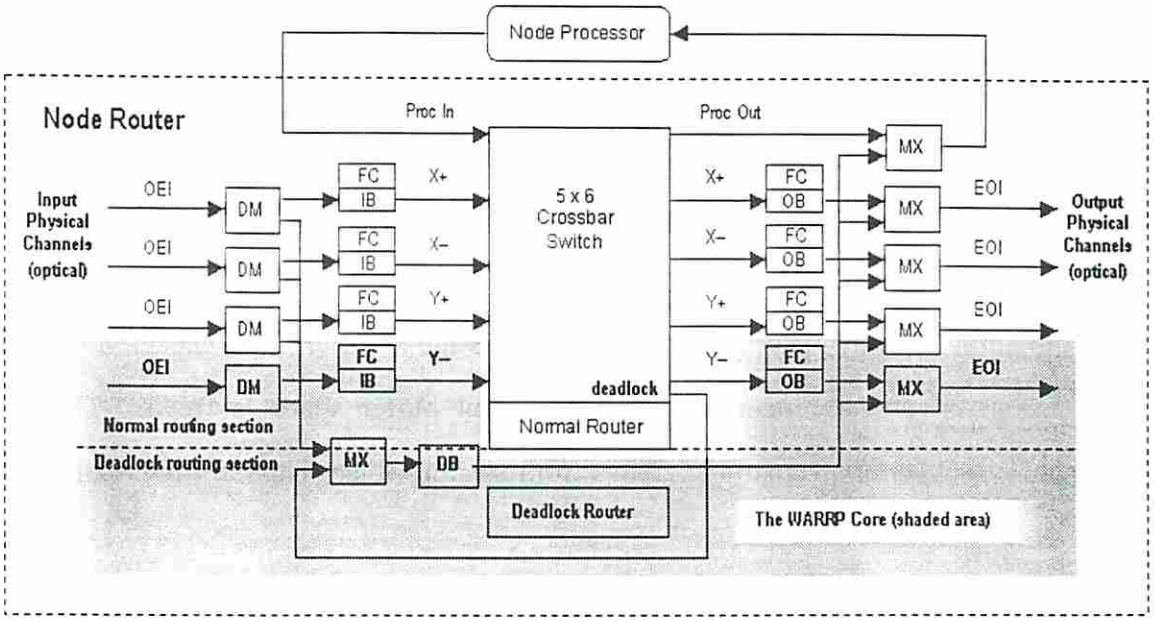
Implementation of an Optoelectronic WARRP Router

Having identified the wiring problem associated to complex optoelectronic chip designs, this chapter tries to evaluate the technological feasibility of such optoelectronic chips which is considered the last piece of information needed to determine the conditions where optoelectronic network routers are feasible. The experiments on chip-level integration of an optoelectronic router through various implementations of the WARRP router were performed. Not only the technological feasibility is evaluated but also the issues related to each technology are identified and used to speculate on which technology is more suitable for optoelectronic network routers.

5.1 WARRP Core Design and Operation

WARRP core is the first optoelectronic implementation of the WARRP router [41]. It implements key elements of progressive deadlock recovery-based adaptive wormhole routing sufficient to demonstrate the interaction between normal packet transmission and deadlock packet recovery in a bit-serial, torus-connected network. Figure 28 shows the block diagram of a fully functional WARRP router and the components comprising the WARRP core. Circuits included on the chip are input buffers, output buffers, deadlock buffers, external flow control logic, channel preemption logic, and deadlock arbitration logic. The objective of this chip design effort was to verify that network bandwidth can

be made globally accessible to packets in an unrestricted manner while remaining deadlock freedom and to show that optoelectronic smart-pixel implementation of network router deadlock handling mechanisms can achieve this. Although the WARRP core itself was designed to handle concurrent deadlock recovery [65], it can also handle sequential deadlock recovery [3] as well by adding a simple external token circuit.



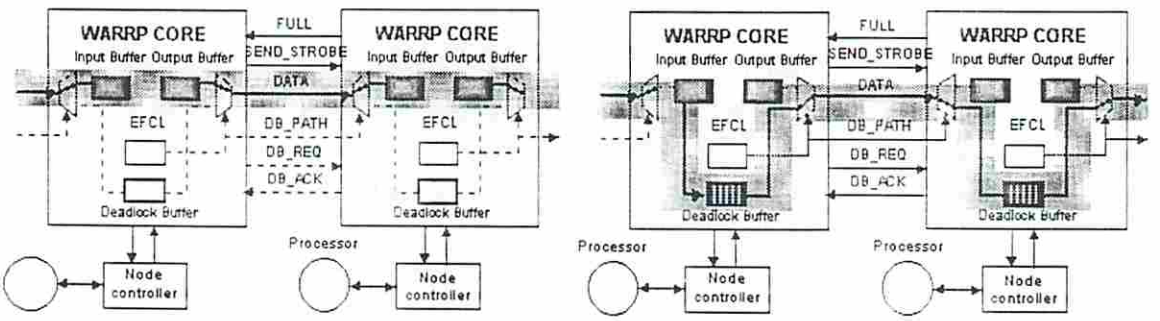
Legend: DM: Demultiplexer MX: Multiplexer FC: Flow Controller
IB: Input VC Buffers OB: Output VC Buffers DB: Deadlock buffer
OEI: Opto-Electronic Interface EOI: Electro-Optic Interface

Figure 28. Block diagram of the WARRP router and the WARRP Core components.

The WARRP core uses two external flow control signals (*SEND_STROBE* and *FULL*) to control flit flow between routers source-synchronously. To make full use of the optical channel bandwidth, the router must have sufficient *input buffer* storage for each channel so that round-trip propagation delay of the flow control handshaking signals can be mad transparent to operation (i.e., multiple flits may be in flight before the *FULL* signal is detected by a router). A router initiates external flow of normal packet flits if its *output buffer* contains data to transmit (i.e., internal router flow transfers flits from input to output buffers). Three external handshaking signals are used to control structured

channel preemption: *DB_REQ*, *DB_ACK* and *DB_PATH*. The first two are used to request and acknowledge structured access to the neighboring router's deadlock buffer; the last is used to indicate whether the packet using the channel is a normal packet or a deadlock packet.

External flow of normal packets operates as follows. If the external *FULL* signal on the channel is not asserted (indicating that the corresponding input buffer of the next router is not full) and the internal *DB_PATH* signal is also not asserted (indicating that this router's deadlock buffer is not preempting the channel), the output buffer transmits its flits on the channel while asserting the *SEND_STROBE* source-synchronous signal. The receiving router uses this signal as an indicator to latch in flits on the channel into its input buffer. Once the receiving router's input buffer nears capacity, the *FULL* signal is asserted and remains so until sufficient buffer space is freed. The sending router observes this signal and ceases from sending flits. This process is illustrated in Figure 29(a) and Figure 30.



(a) Packet transmission over normal buffers.

(b) Deadlock packet initiation and propagation over deadlock buffers.

Figure 29. Concurrent deadlock recovery operation using the WARRP Core chip.

The operation of concurrent progressive deadlock recovery by structured channel preemption in the *WARRP* core is described below. Two scenarios of deadlock buffer acquisition arise: (1) the input buffer of a router requests its own deadlock buffer (i.e., initiation of a deadlock packet); (2) the output buffer of a neighboring router(s) requests a

router's deadlock buffer (i.e., propagation of a deadlock packet). A router initiating a deadlock packet asserts an internal *DB_REQ* signal whereas a router propagating a deadlock packet asserts an external *DB_REQ* signal to its neighboring router along the recovery path. Since multiple requests for a router's deadlock buffer can be received simultaneously (neighboring routers and itself), the router's deadlock buffer arbitration logic grants mutual exclusive access to only one requestor by asserting the *DB_ACK* signal to that requestor, signifying that its request was successful. If a deadlock packet being initiated at the current router is granted access, the header along with all other flits in wormhole succession is switched from the router's input buffer to the deadlock buffer. Otherwise, the granted router asserts the external *DB_PATH* signal to preempt the channel from its output buffer in order to propagate the deadlock packet on to the next router's deadlock buffer. Preemption prediction logic performs this switching of channel usage (output buffer yielding to deadlock buffer) in a single cycle for faster recovery. At this point, the recovery path between two neighboring router deadlock buffers is established, and deadlock packet flit transmission can occur. This is illustrated in Figure 29(b) and Figure 30. In the same manner at subsequent routers along the recovery path, the deadlock packet is routed entirely on the deadlock buffer resources until it reaches its destination. Once the tail flit passes, the deadlock buffer arbitration logic releases the channel, allowing normal packet operation of input and output buffer channel usage to resume.

The operation of sequential progressive deadlock recovery under the control of an external token circuit (either optical or electrical) is now described. Here, an optical asynchronous token circuit implemented in a second chip, the hybrid CMOS/SEED OMNI chip [24], is used to control access to WARRP's deadlock recovery lane. The operation is described using two simple scenarios: normal packet transmission and deadlock packet recovery transmission. An asynchronous optical token is initiated by the Async_Token circuit in the OMNI chip upon receiving the *Input_Token* signal from the

node controller. The node controller is also responsible for packet injection/removal to/from the network through the WARRP core.

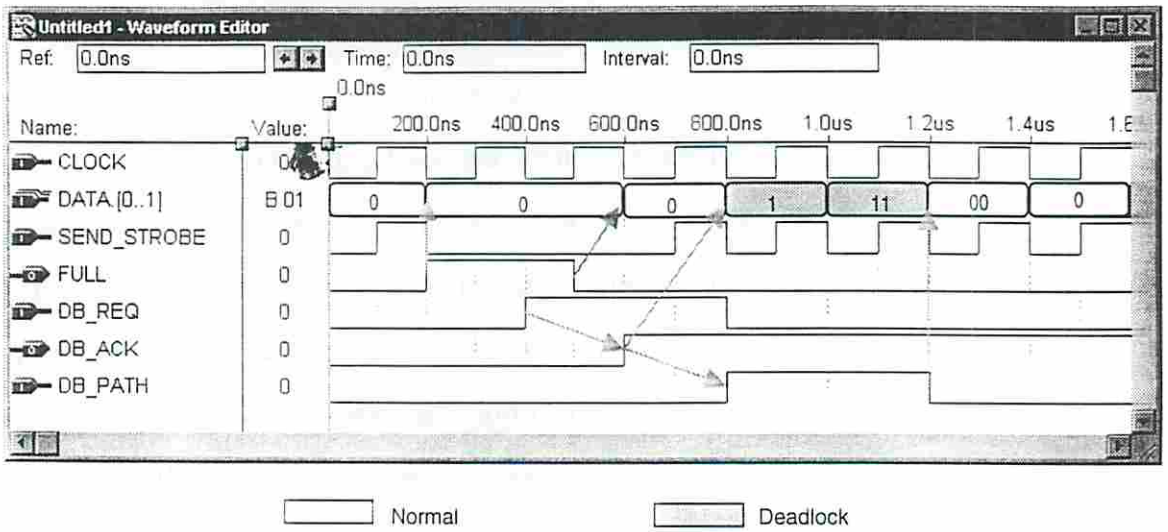
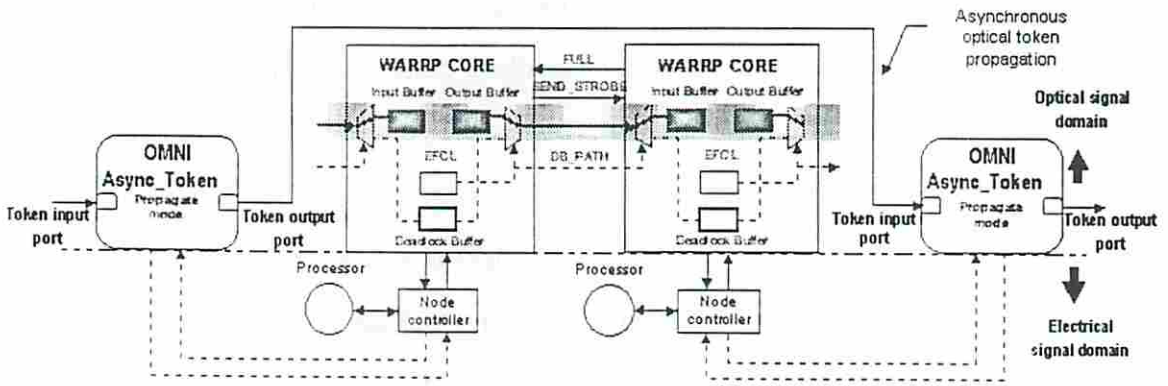
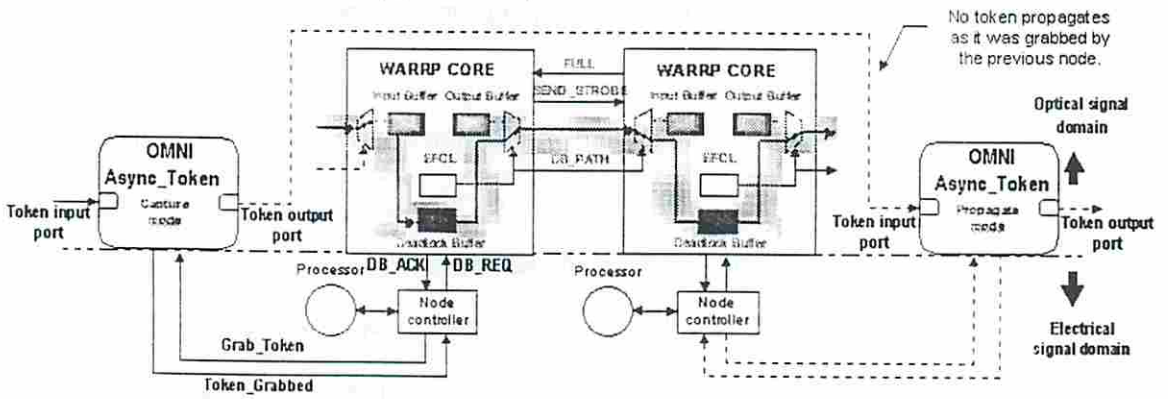


Figure 30. Timing diagram for the WARRP Core (circuit simulation).



(a) Normal packet transmission.



(b) Deadlock packet initiation and propagation over deadlock buffers.

Figure 31. Sequential deadlock recovery operation using WARRP Core and the OMNI chip.

In the first scenario, the token is optically propagated through OMNI's Async_Token smart pixel port which is programmed in *propagate mode*. The token that arrives at the Async_Token circuit is simply regenerated and sent out to the next chip without any intervention from the controller. In the second chip, the WARRP core, all packets are routed independently according to their destinations. The *FULL* and *SEND_STROBE* optical signals are used to implement external flow control between neighboring nodes. Once a packet reaches its destination, it is removed from the network by the node controller. The node controller does not interact with the Async_Token circuit during normal packet transmission. Figure 31(a) depicts the normal packet transmission scenario. Note that only signals involved in the operation are shown, and inactive signals are represented by dotted lines.

The second scenario occurs when deadlock is detected by the deadlock detection logic. The node controller asserts the *Grab_Token* signal to program the Async_Token circuit into *capture mode* to start the deadlock recovery process. As soon as the token is captured, the Async_Token circuit asserts the *Token_Grabbed* signal, notifying the controller that it has exclusive access to the deadlock buffer. The deadlock buffer is allocated to the deadlocked packet after the *DB_REQ* and *DB_ACK* handshaking signals are exchanged. The *DB_PATH* optical control signal is used for constructing a recovery lane by preempting the channels from normal packets. After the deadlocked packet has reached its destination, the recovery lane is torn down and normal packet transmission resumes. The controller asserts *Regen_Token* signal to program the Async_Token circuit into *regeneration mode* in which the token will be internally regenerated and optically propagated to the next node. Figure 31(b) depicts the deadlock packet recovery scenario.

5.2 WARRP Core Smart-Pixel Implementation

The WARRP core design was implemented in a $2\times1 mm^2$ area with 27 electrical I/Os. All functional circuits and photodetectors were fabricated based on the Vitesse H-GaAs III

process, a 0.6μm, four-level metal technology. The LEDs were separately fabricated later using E-O-E technology by the MIT OPTOCHIP group [Appendix C.1]. A micro-photograph of the chip is shown in Figure 32. About 40% of the chip area was used for the design which contains about 1400 transistors representing the input buffer, the output buffer, the deadlock buffer, the external flow controller logic for all buffers including the channel preemption logic, the arbitrator for mutual exclusive deadlock buffer access, and MUX/DEMUXs for channel preemption (Figure 28). All buffers are four bits deep. The circuits are necessary to demonstrate the deadlock recovery scheme described. The rest of the chip area implements the I/O pads, LEDs and drivers, photodetectors, and test circuits. Due to limited chip real estate, the internal flow control logic, the router crossbar, and the routing decision logic are implemented off-chip on a node controller board, as shown in Figure 33. Limited node degree was implemented (i.e., 1-D torus) allowing ring interconnect topology which is sufficient to demonstrate recovery from deadlock. Data, status, and control signals are optically transmitted via six LED/photodetector pairs. Each forms a single-ended optical channel to neighboring chips. The node controller is electrically connected to the chip to send/receive data to/from the chip as well as to control and monitor chip status.

The surface-normal LEDs are GaAs/InGaP double heterostructure devices. The emission spectrum peaks at a wavelength of 873nm. Each LED cell has a $50\times50\mu m^2$ active area and occupies $120\times80\mu m^2$, including driver. The photodetectors are GaAs OPFET (modeled as EFET with the optical power input converted to an equivalent gate bias) with its source input connected to a diode-connected DFET load to make an optical-in/DCFL-out inverter. The size of the photodetector cell is $50\times75\mu m^2$ including a $40\times40\mu m^2$ active gate area. Transistor-level simulation using HSPICE shows that the design operates in excess of 50 MHz.

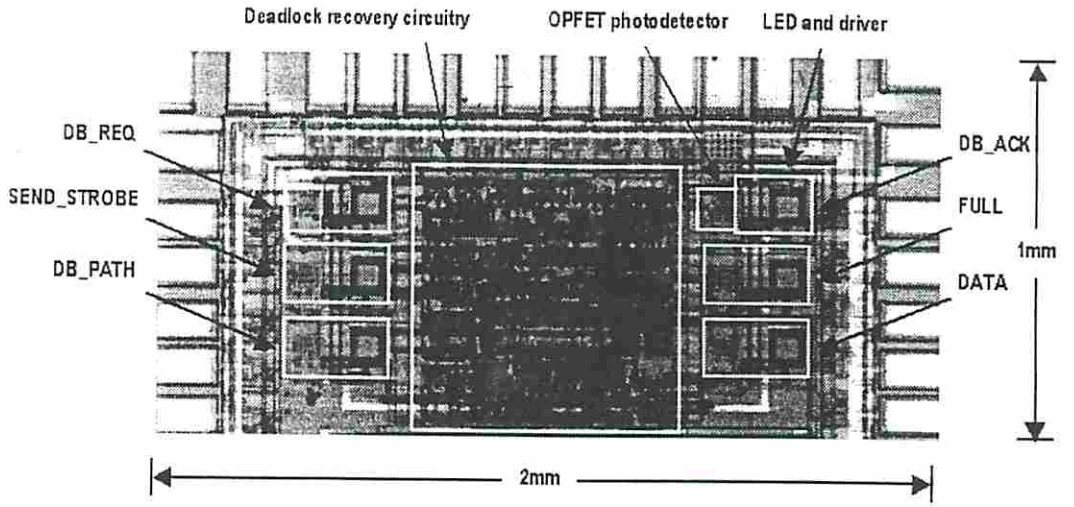


Figure 32. Microphotograph of the WARRP Core.

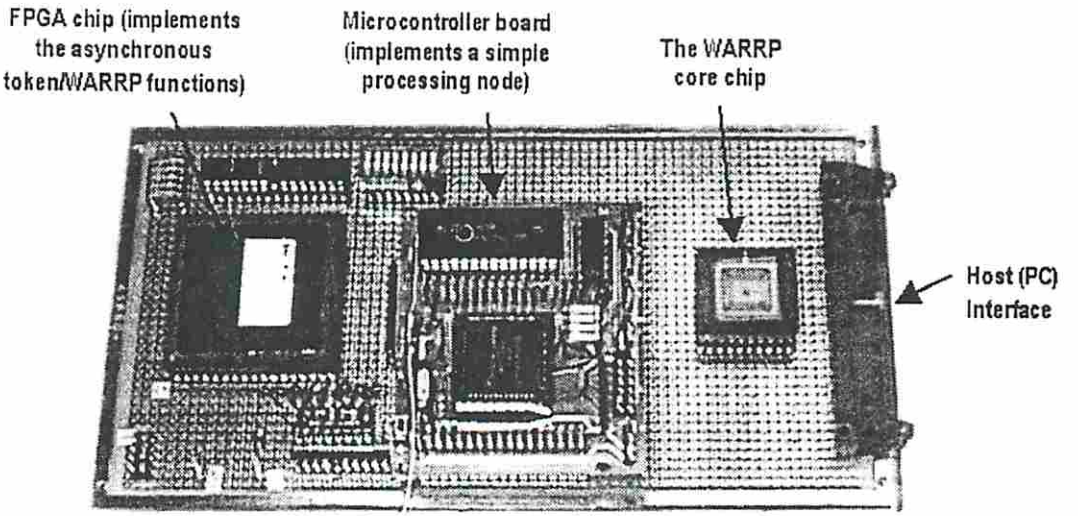


Figure 33. A fully functional microcontroller/FPGA/WARRP testing board.

The WARRP testing board is shown in Figure 33. To reduce the complexity of the optical setup and expedite the testing process, a 12000-gate FPGA (Field Programmable Gate-Array) chip is used to implement both asynchronous token circuit (in lieu of the OMNI chip) and node controller functions. This is integrated onto a 68HC11-based microcontroller board with 32-Kbyte RAM running at 2 MHz. This microcontroller allows us to develop a simple yet highly functional router node, and the flexibility of the FPGA implementation allows us to verify our WARRP core design.

An electrical version of the WARRP core chip was available and tested in 1Q97. Unfortunately the chip did not work at all. The problem might have been caused by insufficient driving capability of the high-sharing signals such as clock and reset signals. This problem was not discovered during the simulation because we mistakenly neglected the I/O pads from the input signal path. In effect, all inputs were driven from strong signals which are not realistic.

5.3 Monolithic GaAs MESFET LED/OPFET Integration Issues

- GaAs MESFET is larger than CMOS transistor and cannot be densely packed. This is to provide sufficient area for heat dissipation because GaAs MESFET dissipates static power. The electronic version of the WARRP core consists of ~4,000 transistors but requires a huge 250mA static current, equivalent to 625mW or 31.25W/cm^2 heat dissipation. For large circuits with large array of optoelectronic devices, the generated heat can be excessive and can affect the operation of the devices. In addition, such circuits may be too large to be efficiently fabricated.

- This technique integrates optoelectronic devices on the same plane of the VLSI circuitry and, hence, the integration of complex circuits with a 2-D array of devices can be very ineffective. Taking into account the requirement of an imaging system, devices are usually arranged in a regular structure, the VLSI circuitry must allocate the area that matches such structure throughout the die to accommodate the devices. The wiring of global connections that are common in complex designs can be extremely difficult. The irregularity of a complex circuit and the size of optoelectronic devices hinder the feasibility of this technology. In the WARRP core design, this problem was mitigated by placing the

optoelectronic devices on the periphery of the VLSI core circuitry. The sizes of both LEDs and OPFETs are so large that a very small area could be used for the core. It is questionable whether this technique will ever support a large array of devices.

- Experimental results show that GaAs MEFSET can easily operate beyond 1GHz but the bottlenecks lie in the optoelectronic devices, especially OPFETs which are shown to operate at only few hundreds kilohertz [45]. The OPFET was chosen as a detector device because of its high gain. However, this is not a critical issue as other devices such as VCSELs/MSMs can be used to improve the performance.
- GaAs MESFET requires voltage levels that are not CMOS compatible (e.g., −1.7 to −0.8V versus 0 to 3.3V in CMOS logic). Therefore, an additional interface is required to operate with the CMOS chips. Not only does it increase the implementation cost but it also decreases the chip performance. Note that high-performance CMOS chips are prevalent and, therefore, this issue is worth considering.

5.4 WARRP II Architecture and Implementation

WARRP II is the second attempt to implement an optoelectronic WARRP router. It features a scaled-down, fully functional version of the WARRP router integrating an array of 20x10 SEEDs on a 2x2mm^2 CMOS circuitry, via flip-chip bonding. The CMOS circuitry was fabricated by MOSIS and later on flip-chip bonded by Lucent. To simplify the integration process, Lucent requires that Metal-3 must be reserved exclusively for SEED bonding. This has a significant impact on the chip performance.

Each Self Electro-optic Effect Device (SEED) is 20x60μm^2 with a horizontal pitch of 62.5μm and a vertical pitch of 125μm, respectively, and operates at 850nm wavelength. Recent experiments have shown that this promising technology can provide more than

47,000 devices on a 3.7x3.7mm^2 area in the near future [66], and each can currently operate at up to 2.48Gb/s with only 300μW optical power input in dual-rail mode [35]. Using the HP14B CMOS process (a 0.5μm, 3-metal layer, 3.3 V supply voltage), this chip contains approximately 15,000 transistors, of which 3,500 are used for I/O pad drivers and optical transceivers. These peripheral circuits occupy almost 40% of the chip area, leaving the remaining 60% for the router circuitry.

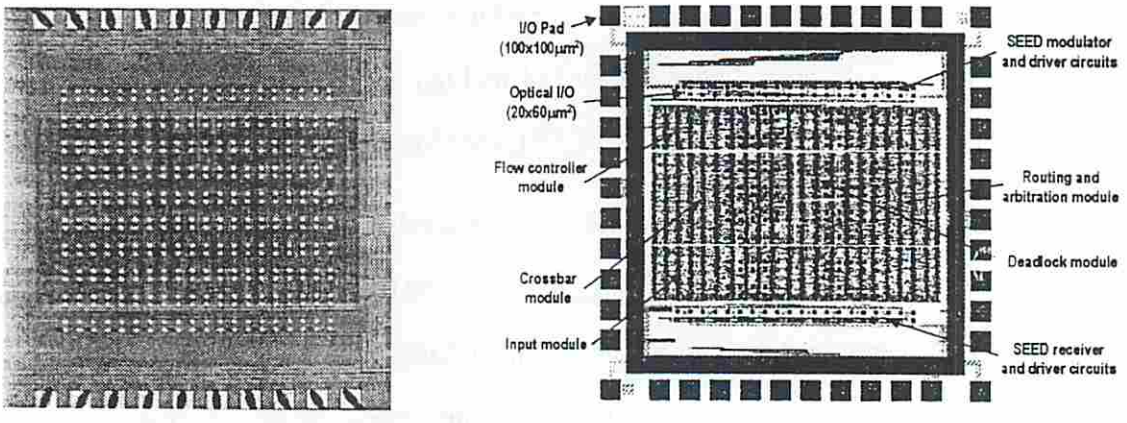


Figure 34. The WARRP II die and its floorplan.

Figure 34 shows the internal modules of the WARRP II chip which consists of 4-flit-deep input buffers, 3-flit-deep output buffers, an address decoder, a 2x3 crossbar, a crossbar arbitrator, and a deadlock core module (i.e., a deadlock buffer and its associated flow controller and channel preemption logic implemented in the WARRP core chip). This chip implements a 4-bit-wide unidirectional torus-connected topology with one virtual channel and associated deadlock recovery mechanisms using 20 optical I/O pin-outs (18 I/Os were used for router ports and 2 I/Os were used for testing purposes). Another 16 signals (for the processor port) were implemented electrically.

The design of WARRP II was split into two phases: CMOS circuitry optimization by EPOCH and manual integration of the CMOS circuitry with the SEED array using MAGIC. Because Metal-3 was reserved for SEED bonding, only two metal layers could be used by the CMOS circuitry. Another limitation was that the design had to fit in a

~1.6x1.6mm^2 area (which excluded the I/O pads). Our synthesis tool yielded a circuit density of ~6,000 transistors/mm^2 without the Metal-3 layer, or ~19,000 transistors for the entire area. Our results indicated that, on average, the layouts expanded by 35% without Metal-3. Since the transistors are moved far apart to make room for circuit wiring, the critical path length doubles. This severely affects the chip functionality and performance.

The design was extensively simulated using switch-level IRSIM (due to its complexity, exhaustive SPICE simulations were not possible given the limited design time fram and CPU resources). Maximum operation speed is estimated to be 25MHz, about half that expected in the original design. This is due to a longer critical path resulting from the exclusion of Metal-3 in circuit wiring.

An electrical version of the WARRP II chip came back and was tested in April 1998. All testings were controlled by LabView program on a host PC via a Digital I/O (DIO) board. The results were not satisfactory; only simple functions such as reset and asynchronous token circuits were working. These results were confirmed by thoroughly testing of the chip with signal generators and oscilloscope. We plan to partition the WARRP II circuitry into pieces and simulate it with HSPICE to pinpoint where the flaws could have been. Hoping to gain more experience on the SEED I/Os, we allocated two of them for transmitter and receiver testing. These two were completely isolated from the core circuitry and should be unaffected by its failure. An optoelectronic version is due in August 1998.

5.5 Hybrid CMOS/SEED Integration Issues

- The need for an external light source for SEEDs complicates the packaging. For instance, Diffractive Optics Elements (DOEs) are required to distribute (split) and focus the light source to all the devices. Experiments in chip and system

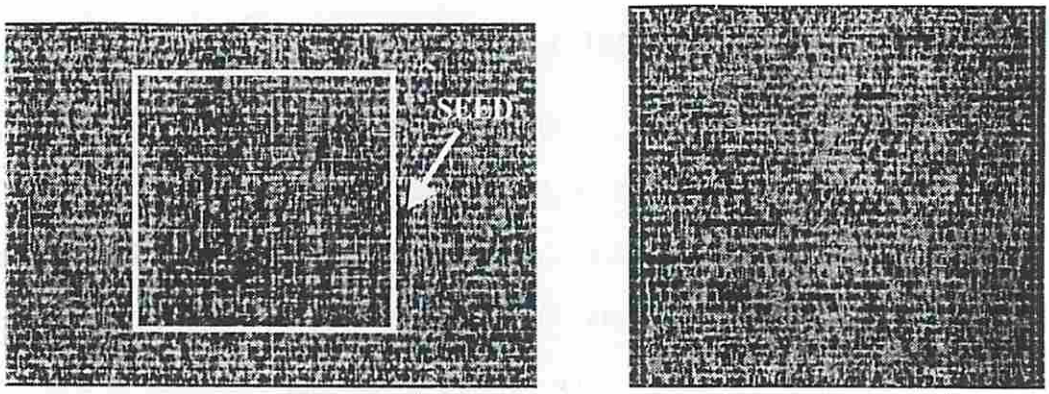
packaging for this technique have made impressive progress, e.g., free-space optical module package [63] and DWDM module for the AMEOBA switch [16]. The former uses only single wavelength (850nm) and is designed for free-space board-to-board interconnects while the latter uses multiple wavelengths, each at 0.5nm apart, and is designed for guided wave (fiber ribbons) system-to-system (i.e., few hundred meters) interconnects. In addition, as the number of devices gets larger, more powerful external light source is required. Assuming a chip with 40,000 SEEDs, 50% DOE efficiency, and 300μW optical power per SEED (to operate at 2.48Gbps [35]), a 24W external light source is required for each chip. Current high-power lasers typically generate less than 10W of continuous power. Some researchers address this issue differently by experimenting on an alternative integration technique—a hybrid CMOS/VCSEL integration [67, 68]—with modulation speed up to 800Mbit/s.

- Since this technique integrates optoelectronic devices "on top" of CMOS-VLSI circuitry via flip-chip bonding (forming a 3D structure), it may seem that this technique is free from the device placement problem found with the monolithic integration technique previously described. This is not the case. In fact, as long as the design is not pixel-based, the problem still persists. In hybrid CMOS/SEED, since the CMOS I/O ports are randomly placed throughout the circuits by the CAD tools, wiring from the I/O port to a regularly structured SEED can become a problem. Each wire can be very long which affects the performance of the chip itself. This issue has been discussed and evaluated in Chapter 4.

5.6 WARRP III Architecture and Implementation

WARRP III is a third optoelectronic design of stemming from the WARRP router design. From a functional perspective, this chip is very similar to WARRP II with 2D torus-connected topology and 2 virtual channels per port. It is the most advanced multi-processor optoelectronic network router to be designed so far, based on a hybrid CMOS/SEED integration technique, on a $3x3mm^2$ die with an array of 20x10 SEEDs. In conjunction with fully adaptive deadlock recovery routing, WARRP III incorporates the enhanced crossbar structure proposed by Choi [4]; each router port has two virtual channels and each virtual channel is connected to a sub-crossbar that system-wide forms a so called "virtual class network."

To address the wiring problem associated with complex optoelectronic chips, the WARRP III chip is being designed using the EPOCH/EGGO CAD tools, assuming the same SEED array and SEED transceivers used in WARRP II. In this software suite, EPOCH performs circuit synthesis and generates the layout, and EGGO automatically performs area-distributed I/Os (SEEDs) and transceiver placement on top of the generated layout. Our previous attempts using only EPOCH resulted in a very inefficient use of chip area and SEEDs. Since the transceivers have to be placed on the periphery of the core circuitry only, the wiring of SEEDs can be very complicated and sometimes impossible for large and complex design like the WARRP III. With EPOCH/EGGO tools, such wiring is accomplished automatically. At present, the design of CMOS/SEED chips has to reserve the topmost metal layer (Metal-3 in this case) for SEED bonding and wiring. Therefore, the layout of CMOS/SEED chip is always larger than its CMOS counterpart as shown in Figure 35.



(a) CMOS/SEED version: 2.77x1.85mm^2. (b) CMOS version: 2.01x1.95mm^2.

Figure 35. Comparison of CMOS/SEED and CMOS implementations of WARRP III (core only).

Figure 35 compares the core circuitry (without electrical I/O pads) of CMOS/SEED and CMOS versions of the WARRP III router based on AMI 0.6μm CMOS process. The boxed area in the center of Figure 35(a) represents the SEED array and the transceivers. Note that both versions could be implemented on a 3x3mm^2 die with 66 electrical pin-outs but CMOS/SEED version have the additional 100 (or 200 in single-ended mode) optical I/O pin-outs, each can operate beyond gigabit per second. (WARRP III requires 23 electrical and 48 optical I/O pin-outs, excluding power lines. So the CMOS version does not meet the I/O requirement.) However, a larger layout can result in longer critical paths thereby decreasing achievable clock rates. Because the design was too large to be simulated by HSPICE on our machine, TACTICS—a timing analysis tool in EPOCH suite—was used to measure the longest critical path of both CMOS and CMOS/SEED versions. The CMOS implementation is 18% smaller in size and ~20% shorter in critical path. It should be noted that this disparity in critical path does not always occur—it depends on the optimizations which are chosen in the synthesis tools. This chip has not yet been sent to fabrication.

5.7 Chip-level Integration: Is It Feasible?

Although the WARRP core and WARRP II chips did not function properly, there is no reason to conclude that complex optoelectronic chips are not feasible. The failure of both chips is due primarily to the lack of sufficient and accurate design simulations. There is nothing to implicate the failure of the integration techniques because other designs in the same foundry runs were successfully tested [16, 27, 69].

Considering the issues related to the integration techniques, hybrid CMOS/SEED integrated technology has been more widely employed and should be more suitable for complex optoelectronic chip design due to its high-density I/O pin-outs and CMOS logic compatibility. This is a basis for designing a CMOS/SEED version of the WARRP III chip.

In conclusion, implementing complex optoelectronic chips such as network routers is possible under current technology. However, optoelectronic compatible CAD tools are required to effectively design such chips by automatically integrating the SEED array and the CMOS circuitry. This is to alleviate the wiring problem inherent in complex optoelectronic designs. Optoelectronic technology alone is not sufficient to accomplish the design and implementation of complex optoelectronic chips.

Chapter 6

Conclusions and Future Work

Optical interconnects incorporating optoelectronic network routers are emerging as an alternative technology to tackle the network bandwidth problem. Nevertheless, the work in this area so far has not clearly demonstrated optics' niche because there are no integral frameworks that investigate the issues related to complex optoelectronic chip designs and implementations, from the system level down to the chip level. This dissertation presents such a framework, which encompasses performance evaluation at the system- and chip-level as well as feasibility assessment of some well-known optoelectronic technologies. The results here, therefore, should be comprehensible to computer architects interested in optoelectronic technology and should encourage the design of novel architectures and/or algorithms that were not possible with the bandwidth-limited electrical interconnects.

6.1 Conclusions

Constant progress in processor performance as a result of better semiconductor technology and advanced architectural techniques has put pressure on the bandwidth requirement for conventional electrical interconnects. Optical interconnects, an emerging technology with a potential to provide high-bandwidth low-latency interconnects based on optoelectronic network routers, can be the solution. Unlike previous optoelectronic chip designs, the optoelectronic network router is much more sophisticated and larger in size. This leads to three interesting questions: How can the optoelectronic router improve

network performance? What are the issues related to optoelectronic network router designs? And, finally, can optoelectronic routers be implemented?

In this dissertation, a cost and performance evaluation of an optical network is performed and compared to a similarly configured electrical network. It is shown that optical interconnects can significantly reduce the network latency via wider channels and higher off-chip clock rates. In addition, high-speed signaling in optical interconnects scales well with faster, next-generation routers. Altogether, an optical network balances between network configurations and network performance very appropriately. Optoelectronic device's and packaging parameters are also considered and shown to affect the network performance adversely. However, such effects are not significant enough to change the results using current technology, and should diminish as progress in optoelectronic and packaging technologies continue.

By observing the layout of the WARRP circuits and the array of optoelectronic devices, it can be concluded that wiring between structured array of optoelectronic devices and randomly placed VLSI circuit's I/O ports is a problem. This problem is also applicable to other complex designs because they are likely to share a similar I/O placement style. The wiring problem results in significantly longer wires, reduced transistor density, and reduced achievable on-chip clock rates. The introduction of optoelectronic compatible CAD tools like EPOCH/EGGO has enabled automatic integration of optoelectronic devices and VLSI circuits thereby optimizing those effects. A semi-empirical model based on wiring cost and resources was established and used to predict the effect of the wiring problem on next-generation complex CMOS/SEED chips. The results show that an optoelectronic chip can provide an order of magnitude more I/O pin-outs while sacrificing approximately 30% achievable on-chip clock rates, compared to its pure CMOS counterpart. Hence the aggregate bandwidth outpaces that which is provided by high-performance BGA packaging in pure CMOS chips. Although transistor

density can be reduced by as much as 40% in complex CMOS/SEED chips, this effect is not major because it exists only on the area beneath the SEED array and transistors are getting cheaper with time.

The direct approach to prove the technological feasibility of optoelectronic routers is to implement them. Based on two optoelectronic/VLSI integration technologies, the WARRP core and the WARRP II optoelectronic router chips were implemented. Both chips are scaled down versions of the WARRP router [41] and incorporate fully adaptive deadlock recovery routing. Due to the lack of sufficient and accurate simulations, most of the chip functions do not work. This result does not discourage the design of complex optoelectronic chips at all because several comparably complex designs have been successfully implemented [16, 27, 69]. In fact, implementation of these chips has led to the last question—whether an optoelectronic router can effectively be implemented.

Overall this dissertation shows that optoelectronic network routers may be technologically feasible and can offer certain architectural advantages in multiprocessor systems. Nevertheless, the success of this emerging technology depends on three major requirements: small and robust packing at all chip and system levels, optoelectronic compatible CAD tools and efficient optoelectronic/VLSI integration techniques, and uniform and reliable optoelectronic devices.

6.2 Future work

So far I have shown that optoelectronic network routers are feasible and potentially provide significant performance boost over the conventional electronic counterparts. To the best of my knowledge, however, all current state-of-the-art network routers have been designed without the assumption of optical interconnects. While the optoelectronic version of such routers is possible, it does not fully utilize the high-bandwidth low-latency capability provided by optical interconnects. It should be noted that optical

interconnects are not cheap and must not be taken for granted. What is suggested as future work, here, is to further optimize the link interface such that the achievable link bandwidth, link utilization, and system performance can be further improved. This would include the development of efficient channel configurations, asynchronous token-based channel arbitration, flit-bundling external flow controls, and an efficient buffer management scheme.

6.2.1 Efficient Channel Configurations

The large number of pin-outs in optical interconnects enables wider communication channels and possibly a wider-than-internal datapath. There are two major configurations to be addressed here. Single Wide Channels (SWC), which allow only one channel per router port, results in virtually constant network latency that is less sensitive to network topologies. On the other hand, Multiple Narrow Channels (MNC), which allow several small channels per router port, can strengthen the network fault-tolerance suitable for critical operations (e.g., in space or remote areas) at the cost of higher per-message latency. While the number of virtual channels is not directly affected by the design choices, the buffer configurations are—SWC requires wider buffers whereas MNC requires deeper buffers. To simplify the discussion, I assume that external channels are wider than the internal datapath. Although a very wide internal datapath is possible, it is not likely due to limited wireability of a complex chip design.

6.2.1.1 Single Wide Channel (SWC) Configurations

In the SWC configuration, as shown in Figure 36, physical channel width is an integer multiple of the internal datapath. Each physical channel is shared by several virtual channels which are arbitrated by a round-robin, first-come-first-served policy. All data coming from virtual channel controllers will be multiplexed onto a wide physical channel and will be demultiplexed at the other side on the next router. Virtual channels

are used to improve channel utilization such that whenever the flit gets blocked, the others can use the physical link. Once connected, the virtual channel controller drives data through the switch and physical channel to the next virtual channel controller in the receiving router. By observing a feedback signal from the receiving node (shown as flow control signal in Figure 36) the sender knows when to stop sending the data. Each virtual channel has its own flow control. Because of different internal and external data width, SWC requires both deserializer and serializer circuits to match the widths.

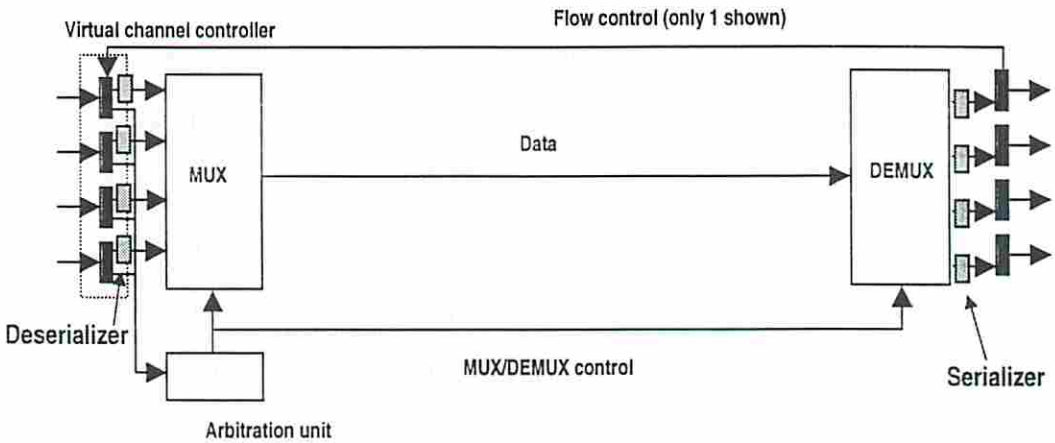


Figure 36. Block diagram of SWC configuration.

6.2.1.2 Multiple Narrow Channel (MNC) Configuration

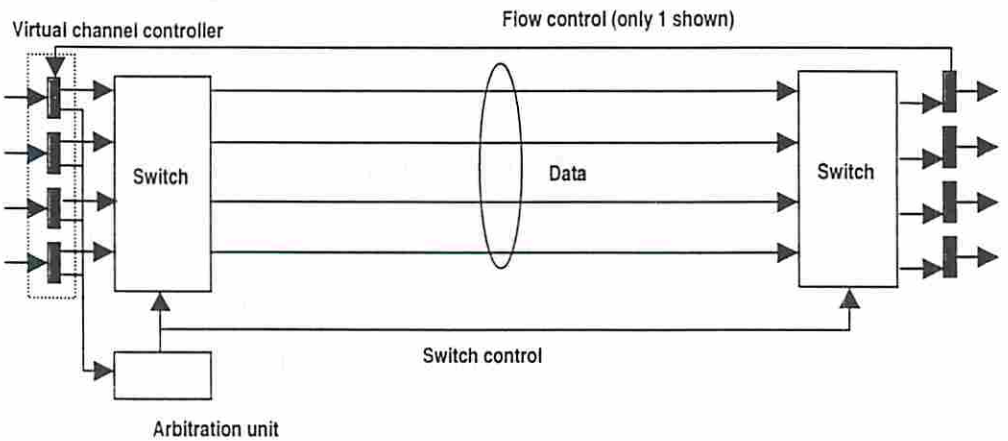


Figure 37. Block diagram of MNC configuration.

In an MNC configuration, as shown in Figure 37, the wide physical channel is equally split into narrower channels; each carries different data packet and operates in parallel. I assume physical channel width is equal to the internal datapath width, thereby eliminating the need for serializer and deserializer circuits. Similar to the SWC, flow control is done on each virtual channel.

6.2.2 *Asynchronous Token-Based Channel Arbitration*

High-bandwidth optical interconnects imply heavy sharing among large number of virtual channels. An efficient channel arbitration scheme is required to reduce the arbitration latency and to improve the channel utilization. More importantly, that scheme must be scalable with increased on-chip and/or off-chip bandwidth at a reasonable implementation cost. Furthermore, it should support simultaneous arbitration as required by MNC configuration.

I propose the use of an asynchronous-based token scheme that employs a circulating token to grant exclusive access to a physical channel in round-robin fashion similar to [70]. Due to its simplicity, it can operate at very high speeds such that the token can asynchronously circulate through all virtual channels within a few (on-chip or off-chip) clock cycles. For instance, an asynchronous token circuit proposed in [24] yields only two gate delays. Since the arbitration latency is very low, virtually constant with the number of virtual channels, and the implementation cost is linear with the number of virtual and physical channels, this scheme is well scalable. Due to its small implementation cost, it can be replicated to support MNC configuration as shown in Figure 38. The arbitration latency can be further reduced by bidirectionally circulating the tokens under MNC configuration.

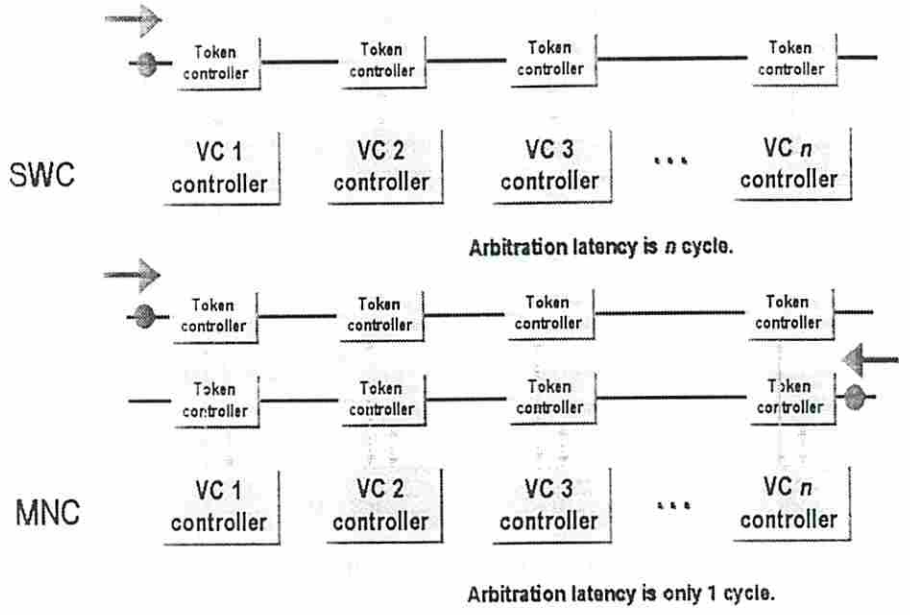


Figure 38. Asynchronous token-based arbitration latency for SWC and MNC.

Arbitration latency of the asynchronous token scheme can be described as follows. Virtual channel controller n (VC n) is requesting to use a physical channel. In SWC, it would take n arbitration cycles (an arbitration cycle is defined as the time it takes for a token to propagate through a VC) in the worst case or $n/2$ cycles on the average. With bidirectional arbitration in MNC, it would take only single arbitration cycle on the average. Note that the arbitration cycle is much faster than on-chip and off-chip clock cycles since it operates asynchronously.

6.2.3 Flit-bundling Transfer Technique

I further assume that external flow control is fully pipelined by means of a wave pipelining scheme [71] and sufficient amount of buffers at both ends. The latter isolates the effect of the buffer management technique on the virtual channel switching from flow control. Currently, electrical interconnects are not capable of high-speed operation, which requires the design of external flow control to be aware of average message latency. Since the off-chip clock rate cannot be very fast, a flow control scheme that includes both channel arbitration and data thru latencies is widely employed. This scheme

features fairness to all active virtual channels and reasonably small average message latency. Each virtual channel takes turns in transmitting a single flit on every clock cycle thereby evenly distributing latency on all message lengths and reducing average message latency. Here, this flow control scheme is referred to as "single-flit transfer technique."

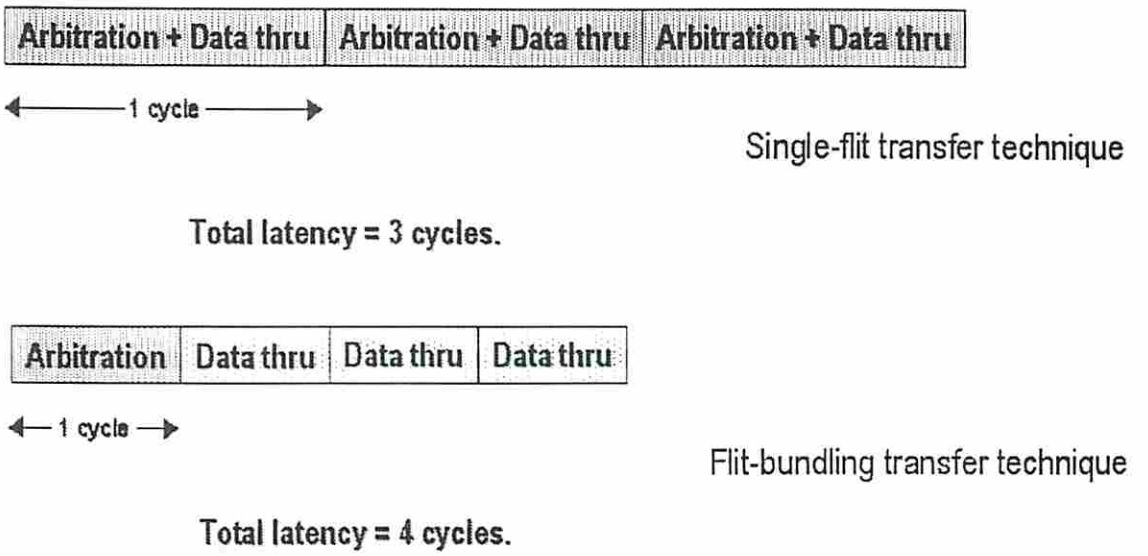


Figure 39. Message latency for single-flit and flit-bundling transfer techniques.

With low-latency optical interconnects, off-chip clock rates can be very fast and, hence, embedding arbitration into data thru cycle becomes very inefficient as it reduces the achievable bandwidth. In this work, I propose the use of flit-bundling transfer technique that decouples the arbitration cycle from the data thru cycle. By doing arbitration only once and transferring as many flits as possible, we can better utilize the channel and increase the off-chip clock rate, as shown in Figure 39. Depending on message lengths, the average message latency may not be significantly increased due to faster off-chip clock speed and lower arbitration overhead.

6.2.4 Delayed Buffer: An Efficient Buffer Management

Evidently, flit-bundling transfer technique works well when there are several flits available in the output buffer to be transmitted continuously (and no blocking on the

other side). However, this is unlikely if the off-chip clock is much faster than on-chip clock. In that case, the router core cannot fill the flits to the output buffer as fast as it is delivered to the channel. Thus, the channel will be switched to another active virtual channel, wasting the useful bandwidth during the arbitration cycle. Design faster router core is one solution but it may not be always achievable. An alternate solution is to overlap the arbitration cycle with the data thru cycle. This can be done by releasing the token as soon as the channel has been granted. While this technique can hide the arbitration latency it cannot hide the switching latency (which is usually included in the arbitration cycle). Including the switching latency in the data thru cycle would unnecessarily reduce the achievable off-chip clock rates. A simpler yet efficient solution is to use a buffer management scheme called "Delayed Buffer."

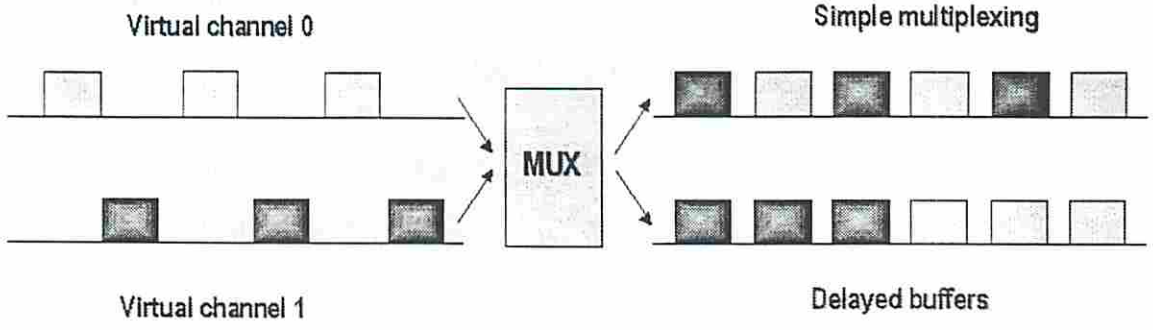


Figure 40. Simple multiplexing and delayed buffer schemes comparison.

The delayed buffer, instead of "greedily" requesting the channel whenever there is a flit to send, waits for a certain number of flits (called "delay threshold") to be buffered before asserting a channel request signal. The hardware required is just a small counter associated with each virtual channel but it may require larger buffer depending on the on-chip to off-chip clock ratio. This technique would make the flit-bundling transfer technique more effective by reducing the frequency of virtual channel switching and the arbitration overhead. An illustration of flit transmission using delayed buffers is shown in Figure 40.

To achieve conclusive results, some parameters must be further investigated which include communication behavior, traffic load rate, message latency, throughput, channel switching frequency, optimal delay threshold, and optimal number of channels. Regardless of the results, the suggested architectures here are expected to achieve higher performance than the non-optimized router architectures and should be further pursued.

Appendix A: Gaussian Beam Propagation through a Lens

This work assumes an optical beam with a Gaussian irradiance profile which is represented by

$$I = I_o e^{\left(-2r^2/w^2\right)}, \tag{33}$$

where I_0 is the intensity at the center of the beam, e is the base of the natural logarithm ≈ 2.718, r is the distance from the beam center, and w is the *spot radius* of the beam where its intensity drops to $1/e^2$ of its peak value, I_0 [34].

As light travels in free-space, the spot size of the Gaussian beam increases:

$$w(z) = w_o\left[1+\left(\frac{\lambda z}{\pi w_o^2}\right)^2\right]^{1/2} \tag{34}$$

where w(z) represents the spot radius at a distance z along propagation axis from the beam waist w_0 at $z = 0$ (where its wavefront was flat, e.g., at the source windows). Gaussian beam propagation through a microlens is illustrated in Figure 41.

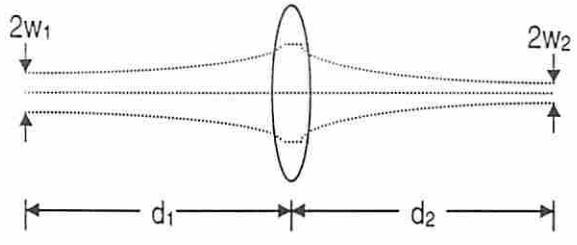


Figure 41. Gaussian beam propagation through a microlens.

This propagation complies with the lens law [34] in which the distances and sizes of object and image are shown to be

$$d_1 = f + \frac{w_1}{w_2}\sqrt{f^2 - \left(\frac{\pi}{\lambda} w_1 w_2\right)^2}, \tag{35}$$

$$d_2 = f + \frac{w_2}{w_1}\sqrt{f^2 - \left(\frac{\pi}{\lambda} w_1 w_2\right)^2}, \tag{36}$$

where d_1 and d_2 are the object and image distances, f is the focal length of the microlens, w_1 and w_2 are the beam radii of the object and image.

Appendix B: Connection Capacity in a DROI System

Each interconnect in a DROI system is described simply by an imaging system with two microlenses and two subholograms as shown in Figure 42.

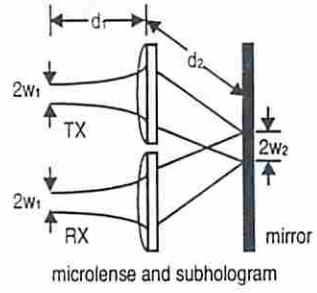


Figure 42. Gaussian beam propagation in a DROI system.

Recall from Section 3.2.4 that connection capacity is a function of the area over which interconnects can be established and the maximum light beam area along the propagation path (Eq.[13]). This maximum light beam area can be observed at the microlenses as shown in Figure 42. Due to symmetry of the system, both transmitter and receiver microlenses happen to be the same (e.g., same diameter, focal length, and f-number).

Theoretically, in a system with no volume constraints, the interconnect area, A_{system}, has no limit. The only limitation in this case is processing technologies (e.g., free-space system packaging, VLSI fabrication process, etc.) In contrast, in a volume-limited system, the maximum interconnect area is

$$A_{system} = \frac{2v}{R_{\max}\cos\theta}, \tag{37}$$

where v is the system volume, R_{max} is the maximum connection path (Eq.[11]), and θ is the maximum hologram deflection angle.

I assume for this DROI system that each optical beam is deflected by a linear blazed grating DOE [21] as illustrated in Figure 43.

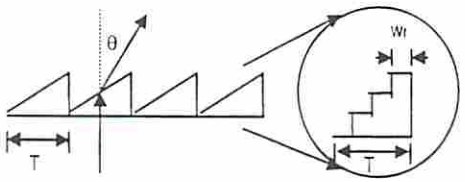


Figure 43. Linear blazed grating DOE structure (four-level binary optics).

Suppose that the linear blazed grating is implemented using binary optics in L_b levels with a feature size equal to w_f and a grating period T. After passing through the grating, if the optical signal propagates through material with refractive index n_x, then the angle of deflection, θ, can be written as [21]

$$\theta = \sin^{-1}\left[\frac{\lambda / n_x}{L_b w_f}\right]. \tag{38}$$

Due to the discrete features of a DOE, several diffraction orders are generated once the light beam propagates through the blazed grating. Only the first diffraction order is used for interconnection. The hologram efficiency is therefore defined as the ratio of the first diffraction order power to the total input power and is given by [21]

$$\eta_{+1} = \left[\frac{\sin(\pi / L_b)}{\pi / L_b}\right]^2 = \operatorname{sinc}^2(1/L_b). \tag{39}$$

Hologram efficiency is a major contribution to link efficiency, which determines the signal conversion time (Section 3.2.2). Progress in VLSI fabrication will help increase both deflection angle and hologram efficiency resulting in better overall system performance.

To find the spot size at a microlens I employ Eq.'s[34-36] with parameters corresponding to those shown in Figure 42. I first find the object distance d_1 and the image distance d_2 which is half of $R_{\max}$ determined by Eq.[11]. My calculation shows the

object distance is approximately $555 \sim 560\mu m$. Once the object distance is found, I use Eq.[34] to find the spot size at the microlens. In order to collect 99.5% of beam power, the microlens diameter, M_D, must be about four times larger than the spot radius at the microlens (using Eq.[33]). Note that I assume that an off-axis hologram does not change the beam radii; rather, it elongates the image or object distances.

Without power dissipation considerations, connection capacity would be limited only by the microlens diameter and the maximum interconnect area. Hence, the maximum connection capacity of a system is simply

$$C = \frac{A_{system}}{2M_D^2}, \tag{40}$$

where C is the maximum connection capacity, A_{system} is the maximum interconnection area, and M_D is the microlens diameter. The factor of two in the denominator takes into account that both transmitter and receiver are in the same plane. (A practical value of connection capacity might be different from Eq.[40] due to not all of A_{system} being used. Moreover, the transmitter/receiver circuit area could be larger than the microlens itself and, hence, determine the system connection capacity.)

By employing the above and parameters assumed in Table 6, I find that a microlens with a diameter of $125\mu m$, a focal length of about $460 \sim 467\mu m$, and an f-number of ~ 3.8 is required for the assumed hypothetical system with no volume constraint. These values are practical, which confirm the validity of my results.

Appendix C: Optoelectronic/VLSI Integration Technologies

C.1 A Monolithic GaAs MESFET/LED/OPFET Integration Technology

The MIT/NCIPT Epitaxy-on-Electronics (E-O-E) is an experimental monolithic integration technology developed at MIT to achieve superior speed, device density, system reliability, ultimate complexity, and manufacturability compared to hybrid integration. The GaAs-based circuitry is fabricated by a standard foundry service such as MOSIS. The LEDs are later grown using the EonE technique at MIT. The process detail is shown in Figure 44.

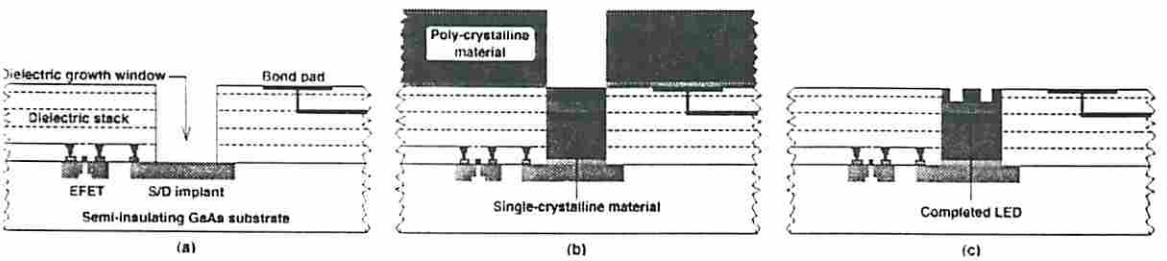
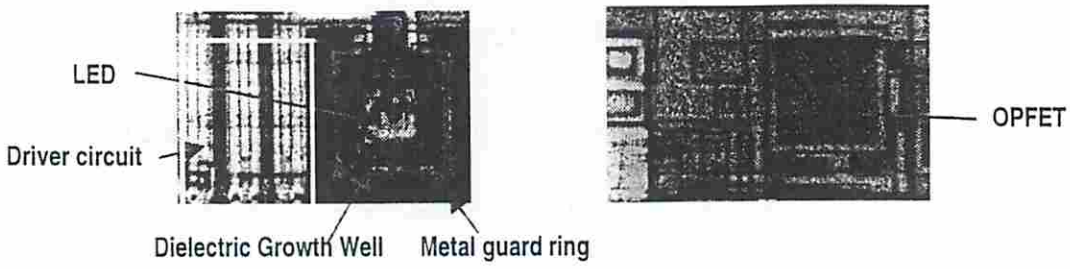


Figure 44. Cross-sectional views of an epitaxy-on-electronics (E-O-E) process.

In Figure 44(a), dielectric insulation and metal layers are removed from the GaAs IC wafer in certain regions to expose the underlying GaAs wafer surface where the optoelectronic devices (which is an LED in this case) will be created. In Figure 44(b), after molecular beam epitaxy (MBE) growth with epitaxial heterostructure on the exposed substrate and polycrystalline deposit on the dielectric layers. A completed LED after polycrystalline deposit is removed and interconnected with electronic bond pad is shown in Figure 44(c).

Figure 45(a) shows the dielectric growth wells where the LEDs will be fabricated on. The OPFET photodetectors are GaAs OPFET (Optical FET) modeled as EFET with the optical power input converted to an equivalent gate bias. Its source input is connected to a diode-connected DFET load to make an optical-in/DCFL-out inverter. Since the detectors are indeed regular GaAs MESFETs, they are fabricated with the VLSI circuitry.



(a) LED and driver circuit. (b) OPFET photodetector.

Figure 45. Microphotographs of LED and OPFET of the WARRP core.

C.2 A Hybrid CMOS/SEED Integration Technology

Hybrid SEED technology [59] has been proven to be the most efficient optoelectronic/VLSI integration technology to-date. This technique not only allows easy conversion between electronics and optics and vice versa but also allows large numbers of devices, which is a crucial solution for leveraging the need for off-chip bandwidth demanded by high-performance silicon-based (CMOS) chips. The integration fits well with silicon electronics without requiring any substantial changes to that technology, and should be capable of operating at the speeds and voltages of silicon circuits. Recent experiments show that more than 16,000 SEEDs can be successfully integrated on a silicon circuitry at 99.878% yield [33]. Each device can operate up to 2.48Gb/s in dual-rail mode [35].

The hybrid SEED relies on quantum well diode modulators [72]. Multi quantum well diode, made of 50 to 100 alternating very thin (e.g., 10nm) layers of two different semiconductor materials, show a large change of optical absorption when an electric field is applied perpendicular to the layers [73]. An efficient way to control the optical absorption and minimize the power dissipation is to reverse biasing the diode. Such diodes need only be a few microns thick overall yet still would have sufficient change in optical transmission to make useful modulators. This allows two-dimensional arrays of such diode modulators to be made using standard semiconductor growth and processing

techniques. The modulators are very fast, limited in practice only by the electrical drive circuits and parasitic capacitance, and can operate with compatible voltages (e.g., 3.3 or 5V). The modulator diodes also function as good photodiodes, so the same device can be used for optical modulators and detectors.

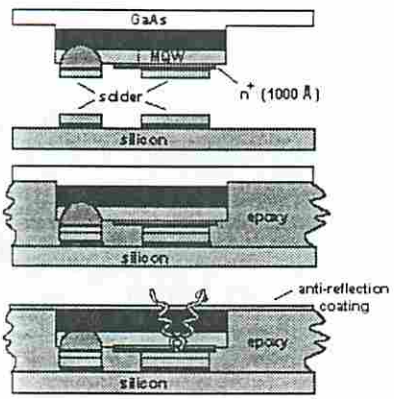


Figure 46. Illustration of flip-chip bonding process used to bond arrays of MQW diode modulators and detectors to silicon CMOS circuitry in the hybrid SEED process. With the final removal of the epoxy, individual modulators are left connected to the silicon circuitry in a 2D array.

The hybrid SEED process is shown in Figure 46. Arrays of MQW diodes are fabricated with reflecting metal on the top. The array is then turned over and solder-bonded to the silicon circuit as shown in top and middle. The GaAs substrate is removed chemically from the quantum well diodes to leave isolated diodes bonded to the silicon circuit. The resulting array can be used as reflection modulators or photodetectors, depending on the silicon circuits to which they are connected. The use of reflection modulators is convenient since it means that light does not have to pass through the silicon circuits, so conventional silicon circuit mounting can be used, and the double pass of the light beam through the modulator increases the amount of modulation of the light beam.

The only additional processing of the silicon circuit which is required is to deposit some metals and solder. This can be done after the usual fabrication of the silicon wafer,

and no change is required in the usual fabrication process. As a result, this technique can be used with silicon circuits from any fabrication process.

Figure 47 shows a picture of a silicon circuit with attached quantum well diodes. In this case, the diode size is $15x40\mu m^2$, and the solder bond size is $15x15\mu m^2$. Interestingly this is much smaller than a regular bonding I/O pad (see Figure 5).

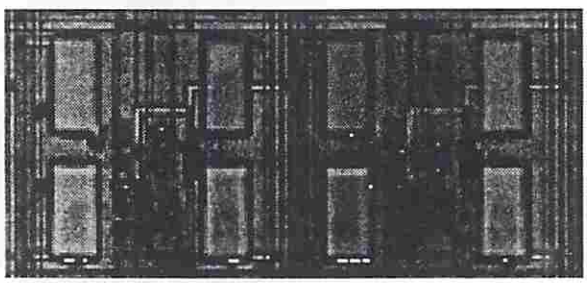


Figure 47. Picture of part of a hybrid SEED chip. The quantum well diodes are the regular array of rectangles, and each is $15x45\mu m^2$ in area and a few microns thick. Underneath is active silicon circuitry.

References

[1] David Patterson et al., "A Case for Intelligent RAM," *IEEE Micro*, **17**(2), 34-44 (1997).

[2] Jim Turley, "Mitsubishi Mixes Processor, Memory: M32R/D Combines 32-Bit RISC Core with 2 Mbytes of On-Chip DRAM," *Microprocessor Report*, 10-12 (May 27, 1996).

[3] Anjan Venkatramani and Timothy M. Pinkston, "DISHA: A Deadlock Recovery Scheme for Fully Adaptive Routing," *Proceedings of the 9th International Parallel Processing Symposium*, 537-543 (1995).

[4] Yungho Choi and Timothy M. Pinkston, "Crossbar Analysis for Optimal Deadlock Recovery Router Architecture," *Proceedings of the 11th International Parallel Processing Symposium*, 583-588 (1997).

[5] Mike Galles, "SPIDER: A High-Speed Network Interconnect," *IEEE Micro*, **17**(1), 34-39 (1997).

[6] Joseph Carbonaro and Frank Verhoorn, "Cavallino: The Teraflops Router and NIC," *Proceedings of Hot Interconnects IV*, 157-160 (1996).

[7] Steven L. Scott and Gregory M. Thorson, "The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus," *Proceedings of Hot Interconnects IV*, 147-156 (1996).

[8] Karl-Heinz Brenner and Frank Sauer, "Diffractive-reflective optical interconnects," *Applied Optics*, 4251-4254 (1988).

[9] Matthew Haycock and Randy Mooney, "A 2.5Gb/s Bidirectional Signaling Technology," *Proceedings of Hot Interconnects IV*, 149-156 (1997).

[10] William J. Dally and John Poulton, "Transmitter Equalization for 4-Gbps Signalling," *IEEE Micro*, **17**(1), 48-56 (1997).

[11] The OETC project web page http://www-phys.llnl.gov/H_Div/photonics/OptInt.html

[12] The Jitney Project web page http://atp.nist.gov/www/comps/briefs/93010151.htm

[13] The POLO Project web page
http://www.usc.edu/dept/engineering/eleceng/Adv_Network_Tech/pol.html

[14] Motorola Optobus web page http://design-net.com/logic/optobus.homepage.html

[15] John Lehman, "An Introduction to the ChEEtah Project," *Proceedings of Hot Interconnects V*, 125-132 (1997).

[16] Ashok V. Krishnamoorthy et al., "The AMOEBA Chip: An Optoelectronic Switch for Multiprocessor Networking Using Dense-WDM," *Proceedings of the 3rd International Conference on Massively Parallel Processing using Optical Interconnects*, 94-100 (1996).

[17] Haldun M. Ozaktas and Joseph W.Goodman, "Lower bound for the communication volume required for an optically interconnected array of points," *Journal of the Optical Society of America*, 2100-2106 (1990).

[18] Michael R. Feldmand and Clark C. Guest, "Interconnect density capabilities of computer generated holograms for optical interconnection networks using diffractive analysis," *Applied Optics*, 4052-4064 (1995).

[19] Ahmed Louri and Stephen Furlonge, "Feasibility study of a scalable optical interconnection network for massively parallel processing systems," *Applied Optics*, 1296-1308 (1996).

[20] William J. Dally, "Performance Analysis of k-ary n-cube Interconnection Networks," *IEEE Transaction on Computers*, 775-785(1990).

[21] A. H. Sayles, B. L. Shoop, and E. K. Ressler, "A novel smart pixel network for signal processing applications," *Proceedings of the LEOS 1996 Summer Topical Meeting on Smart Pixels Technical Digest*, 86-87 (1996).

[22] D. S. Wills et al., "A Fine-Grain, High-Throughput Architecture Using Through-Wafer Optical Interconnect," *Journal of Lightwave Technology*, 1085-1092 (1995).

[23] F. B. McCormick et al., "Five-stage free-space optical switching network with field-effect transistor self-electro-optic effect devices smart-pixel arrays," *Applied Optics*, 1601-1681 (1994).

[24] Timothy M. Pinkston and Charles Kuznia, "Smart-pixel-based network interface chip," *Applied Optics*, 4871-4880 (1997).

[25] Timothy M. Pinkston, Mongkol Raksapatcharawong, and Yungho Choi, "WARRP Core: Optoelectronic implementation of network router deadlock handling mechanisms," *Applied Optics*, 276-283 (1998).

[26] Timothy M. Pinkston, Mongkol Raksapatcharawong, and Yungho Choi, "WARRP II: an optoelectronic fully adaptive network router chip," *Optics in Computing Technical Digest of the 1998 International Tropical Meeting*, 311-315 (1998).

[27] F. E. Kiamilev et al., "Design of a 64-bit, 100 MIPS microprocessor core IC for hybrid CMOS-SEED technology," *Proceedings of the 3rd International Conference on Massively Parallel Processing using Optical Interconnects*, 53-60 (1996).

[28] Ashok Krishnamoorthy et al., "Photonic page buffer based on GaAs MQW modulators bonded directly over active silicon CMOS circuits," *Applied Optics*, 2443-2448 (1996).

[29] Fouad Kiamilev, Richard Rozier, and Ashok Krishnamoorthy, "Smart Pixel IC Layout Methodology and its Application to Photonic Page Buffers," *International Journal of Optoelectronics*, 199-216 (1997).

[30] Richard Rozier, Ray Farbarik, Fouad Kiamilev, Jeremy Ekman, Premanand Chandramani, Ashok Krishnamoorthy, and Richard Oettel, "Automated Design of ICs with Area-Distributed I/O Pads," *1998 IEEE/LEOS Summer Topical Meetings—Smart Pixels Session*, 25-26 (1998).

[31] R. R. Tummala and E. J. Rymaszewski, *Microelectronics Packaging Handbook.* New York: Van Nostrand Reinhold, 1989.

[32] The National Technology Roadmap for Semiconductor (NTRS) document available on the WEB at http://www.sematech.org:80/public/roadmap/index.htm

[33] T. L. Worchesky, K. J. Ritter, R. Martin, and B. Lane, "Large arrays of spatial light modulators hybridized to silicon integrated circuits," *Applied Optics*, 1180-1186 (1996).

[34] Ashok V. Krishnamoorthy, "Scaling Optoelectronic-VLSI Circuits into the 21st Century: A Technology Roadmap," *IEEE Journal of Selected Topics in Quantum Electronics*, 55-76 (1996).

[35] T. K. Woodward, A. L. Lentine, K. W. Goossen, J. A. Walker, B. T. Tseng, S. P. Hui, J. Lothian, R. E. Leibenguth, "Demultiplexing 2.48-Gb/s Optical Signals with a CMOS Receiver Array Based on Clocked-Sense-Amplifier," *IEEE Photonics Technology Letters*, 9(8), 1146-1148 (1997).

[36] Stenven P. Vander Wiel and David J. Lilja, "When Caches Aren't Enough: Data Prefetching Techniques," *IEEE Computer*, **30**(7), 23-30(1997).

[37] Bernard K. Gunther, "Multithreading with Distributed Functional Units," *IEEE Transactions on Computers*, 399-411 (1997).

[38] John L. Hennessy and David Patterson, *Computer Architecture A Quantitative Approach*, 2nd edition, San Francisco: Morgan Kaufmann, 1996.

[39] Jose Duato, "A new theory of deadlock-free adaptive routing in wormhole networks," *IEEE Transactions on Parallel and Distributed Systems*, 1320-1331 (1993).

[40] Timothy M. Pinkston and Sugath Warnakulasuriya, "On Deadlocks in Interconnection Networks," *Proceedings of the 24th International Symposium on Computer Architecture*, 38-49 (1997).

[41] Timothy M. Pinkston, Yungho Choi, and Mongkol Raksapatcharawong, "Architecture and Optoelectronic Implementation of the WARRP Router," *Proceedings of Hot Interconnects V*, 181-189 (1997).

[42] D. A. B. Miller and H. M. Ozaktas, "Limit to the Bit-Rate Capacity of Electrical Interconnects from the Aspect Ratio of the System Architecture," *Journal of Parallel and Distributed Computing*, 42-52 (1997).

[43] Y. Liu, M. Hibbs-Brenner, B. Morgan, J. Nohava, B. Walterson, T. Marta, S. Bounnak, E. Kalweit, J. Lehman, D. Carlson, P. Wilson, "Integrated VCSELs, MSM Photodetetors, and GaAS MESFETs for Low Cost Optical Interconnects," *In Spatial Light Modulators Technical Digest, OSA 1997 Spring Topical Meeting*, 22-24 (1997).

[44] Shinji Matsuo, Kuota ateno, and Takashi Kurokawa, "VCSEL-Based Smart Pixel," *In Spatial Light Modulators Technical Digest, OSA 1997 Spring Topical Meeting*, 19-21 (1997).

[45] Information can be found at http://web.mit.edu/fonstad/optochip/opto.home.html

[46] Timothy M. Pinkston, Mongkol Raksapatcharawong, and Yungho Choi, "WARRP II: An Optoelectronic Fully Adaptive Network Router Chip," *OSA Topical Digest for Optics in Computing 1998*, 311-315 (1998).

[47] Kazuhiro Aoyama and Andrew A. Chien, "The Cost of Adaptivity and Virtual Lanes in a Wormhole Router," *Journal of VLSI Design* (1994).

[48] Bradley D. Clymer and Joseph W. Goodman, "Optical Clock Distribution to Silicon Chips," *Optical Engineering*, 1103-1108 (1986).

[49] Neil H. E. Weste and Kamran Eshraghian, *Principle of CMOS VLSI Design: A Systems Perspective 2nd edition*, Addsion-Wesley 1993.

[50] Sandia National Laboratories web page at http://www.sandia.org/

[51] Michael R. Feldman, Sadik C. Esener, Clark C. Guest, and Sing H. Lee, "Comparison between optical and electrical interconnects based on power and speed considerations," *Applied Optics*, 1742-1751 (1988).

[52] Jürgen Jahns and Sing H. Lee, *Optical Computing Hardware*, Academic Press 1994.

[53] Timothy J. Drabik, "Optoelectronic Integrated Systems Based on Free-Space Interconnects with an Arbitrary Degree of Space Variance," *Proceeding of the IEEE*, 1595-1622 (1994).

[54] James Buchanan, *CMOS/TTL Digital Design*, Mcgraw-Hill 1990.

[55] Kevin Bolding, "Chaotic Routing: Design and Implementation of an Adaptive Multicomputer Network Router," *Ph.D. Thesis*, University of Washington (July, 1993).

[56] T. B. Alexander, K. G. Robertson, D. T. Lindsay, D. L. Rogers, J. R. Obermeyer, J. R. Kelly, K. Y. Oka, and M. M. Jones, "Corporate Business Servers: An Alternative to Mainframes for Business Computing," *HP Journal*, 8-33 (June 1994).

[57] Philippe J. Marchand, Ashoj V. Krishnamoorthy, Sadik C. Esener, and Uzi Efron, "Optically Augmented 3-D Computer: Technology and Architecture," *Proceedings of the First International Workshop on Massively Parallel Processing using Optical Interconnects*, 133-139 (1994).

[58] W. Stephen Lacy, Christophe Camperi-Ginestet, Brent Buchanan, D. Scott Wills, Nan Marie Jokerst, and Martin Brooke, "A Fine-Grain, High-Throughput Architecture Using Through-Wafer Optical Interconnect," *Proceedings of the First International Workshop on Massively Parallel Processing using Optical Interconnects*, 27-36 (1994).

[59] K. W. Goossen et al., "Demonstration of a dense, high-speed optoelectronic technology integrated with silicon CMOS via flip-chip bonding and substrate removal," *1995 Spring Topical Meeting—Optical Computing Section*, 142-144 (1995).

[60] Gye M. Yang, Michael H. MacDougal, and P. Daniel Dapkus, "Low threshold native-oxide defined SQW VCSELs with AlAs/GaAs DBRs," *OSA Topical Digest for CLEO '95/QELS*, CPD4-1 (1995).

[61] M. K. Hibbs-Brenner, R. A. Morgan, R. A. Walterson, J. A. Lehman, E. L. Kalweit, S. Bounnak, T. Marta, and R. Gieske, "Performance, Uniformity, and Yield of 850-nm VCSEL's Deposited by MOVPE," *IEEE Photonics Technology Letters*, 7-9 (1996).

[62] M. C. Wu, "Micromachining for Optical and Optoelectronic Systems," *Proceedings IEEE*, 1833-1856 (1997) (invited paper).

[63] M. H. Ayliffe, D. Kabal, P. Khurana, F. Lacroix, A. G. Kirk, F. P. A. Tooley, and D. V. Plant, "Optomechanical, electrical and thermal packaging of large 2D optoelectronic device arrays for free-space optical interconnects," *OSA Topical Digest for Optics in Computing 1998*, 502-505 (1998).

[64] C.-H. Chen, B. Hoanca, C. B. Kuznia, A. A. Sawchuk, and J.-M. Wu, "Architecture and Optical System Design for TRANslucent Smart Pixel Array (TRANSPAR) Chips," *OSA Topical Digest for Optics in Computing 1998*, 316-319 (1998).

[65] Anjan K. V., Timothy M. Pinkston, and José Duato, "Generalized Theory for Deadlock-Free Adaptive Wormhole Routing and its Application to DISHA Concurrent," *Proceedings of the 10th International Parallel Processing Symposium*, 815-821 (1996).

[66] K. W. Goossen, "Optoelectronic/VLSI," *1997 OSA Spring Topical Meeting—Spatial Light Modulators Technical Digest*, 2-5 (1997).

[67] U. Koelle et al., "Integration of VCSEL Arrays with Silicon Chips for Free-Space Optical Interconnects," *1998 IEEE/LEOS Summer Topical Meetings—Smart Pixel Session*, Postdeadline Papers PD002.

[68] L. M. F. Chirovsky et al., "Bottom-Emitting I^2-VCSEL's for Flip-Chip Bonding to Smart Pixel IC's," *1998 IEEE/LEOS Summer Topical Meetings—Smart Pixel Session*, Postdeadline Papers PD003.

[69] Dirk A. Hall et al., "Experimental Demonstration of OPTOCHIP: A GaAs E-O-E Smart Pixel Neural Array for Digital Image Halftoning," *1998 IEEE/LEOS Summer Topical Meetings—Smart Pixel Session*, Postdeadline Papers PD001.

[70] James D. Allen, Patrick T. Gaughan, David E. Schimmel, and Sudhakar Yalamanchili, "Ariadne—An Adaptive Router for Fault-tolerant Multicomputers," Georgia Institute of Technology, Technical Report TR-GIT/CSRL-93/10.

[71] J. Duato, P. Lopez, F. Silla, and S. Yalamanchili, "A High Performance Router Architecture for Interconnection Networks," *Proceedings of the 25th International Conference on Parallel Processing*, 61-68 (1996).

[72] D. A. B. Miller, "Quantum well optoelectronic switching devices," *International Journal of High Speed Electronic*, 19-46 (1990).

[73] D. A. B. Miller, D. S. Chemla, T. C. Damen, A. C. Gossard, W. Wiegmann, T. H. Wood, and C. A. Burrus, "Electric field dependence of optical absorption near the bandgap of quantum well structures," *Phys. Rev. B 32*, 1043-1060 (1985).