

A Progress in Developing High-performance Multiprocessor Network Routers based on Optoelectronic Technologies[¶]

Mongkol Raksapatcharawong[§]

Electrical Engineering Dept., Kasetsart University, Jatujak, Bangkok 10900, THAILAND

E-mail: fengmkr@ku.ac.th http://www.eng.ku.ac.th/~fengmkr/

Abstract

Computer architects have realized that interconnection bandwidth has become a critical limitation to the development of high-performance multiprocessor systems. Major reason is that the progress of processor performance has increasingly outpaced that of the interconnection networks, thereby limiting the usefulness of multiprocessor systems. This work presents a comprehensive study and the development of optoelectronic-based network routers. Optoelectronic technology can potentially provide ample bandwidth required by multiprocessor systems but at the same time can raise some critical issues that are discussed here such as on-chip wiring and chip packaging. We also proposed new architectural techniques suitable for the development of optoelectronic-based network routers to increase the network bandwidth utilization.

Keywords: Interconnection network, Multiprocessor, Optoelectronic, Router.

1. Introduction

The constant progress in semiconductor technology has enabled the design of more advanced microprocessor on a single die. As a result, microprocessor performance doubles at approximately every 18 months. This has two contradict effects in computer architect's point of view. On one hand, a more powerful computer system can be designed and implemented. This is particularly true in uniprocessor systems. On the other hand, all sub-systems must be improved at a similar rate. Unfortunately, this is not the case for the interconnect sub-system [1], especially in distributed multiprocessor systems where global communications among processing nodes are handled by the underlying interconnection networks, as shown in Figure 1. Hence, multiprocessor system cannot gain much performance from the more powerful processors unless its interconnection network is sufficiently improved.

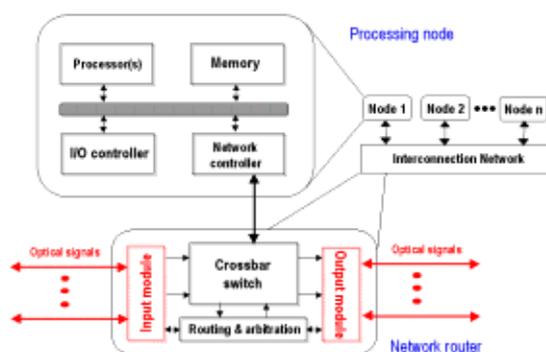


Figure 1. A distributed multiprocessor system with an optical interconnection network.

Insufficient network bandwidth is the problem. As processors are faster, they tend to communicate with each other more often, and, thus, require more bandwidth from the network. Current electrical interconnects have various inherent limitations such as skin effect, crosstalks, EMI/RFI interferences, and etc. which do not allow high-bandwidth communication at a useful range. Although semiconductor technology allows faster routers to be developed, electrical interconnect does not. This can be seen in state-of-the-art network routers used in commercial multiprocessor systems including SGI [2], CRAY T3E [3], and Intel Teraflop [4]. Such routers operate at a humble 375MHz or less clock rates and 20-bit-wide or less datapath which yield less than 1GB/s of raw bandwidth per direction per port.

Similar to 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, shown in Figure 1, feature high-bandwidth optical interconnects by means of a large number of I/O pin-outs (up to 47,000 I/Os are expected [5]), each capable of operating at very high speeds (up to 2.48Gb/s has been reported [6]). 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 which will be discussed collectively in this work.

The rest of the paper is organized as follow: Section 2 evaluates the usefulness of optoelectronic routers in terms of network performance using an analytical model. Section 3 discusses design issues of complex optoelectronic chips and estimates their effects on chip performance. Section 4 presents WARRP II, an optoelectronic network router, and explains the system demonstration of the

[¶] This work was supported by a research fellowship from the Royal Thai Government and is now being supported by the Engineering Research Fund, grant no. 41/08/EE; Kasetsart University Research and Development Institute, grant no. KURDI-6.42 and the Thailand Research Fund, grant no. PDF42-mongkol.

[§] The author is an instructor and researcher associated with the *SCORPion* (Superior *CO*munications *R*esearch and *P*rototyping for *co*mmercialization) group at Kasetsart University.

chip. Section 5 features continuing research (the proposed architectural techniques to improve the router performance) that are being conducted by the SCORPion group at Kasetsart University. Section 6 concludes this work.

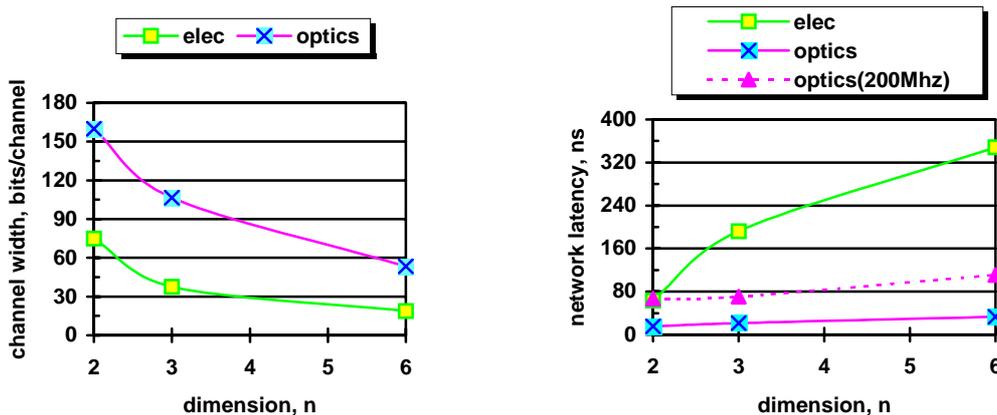
2. Performance of Optoelectronic-based Networks

An important question that must be answered regarding the use of optical interconnects in multiprocessor systems is, “How effective is it?” Clearly, with high-speed signaling and dense optical I/Os provided by optoelectronic technology, network routers can be designed to operate at much higher clock rates and can have more number of ports with each having wider channels. Consequently, packets can be transmitted to the destinations at faster clock speeds and less number of clock cycles. In this section, we investigate such effects assuming a widely known k -ary n -cube class of networks which is utilized in several multiprocessor systems such as the CRAY T3E or Intel Teraflops. Topologies for this class of networks have channels which span n dimensions and have k nodes connected in each dimension (radix).

The network analysis here assumes wormhole switching [7] which pipelines the transfer of flits¹ 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

$$T_{lat} = T_C \cdot \left(D + L_F \cdot \frac{F}{W} \right) + T_{contention}, \quad (1)$$

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 flits, 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 [8]. 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 includes the decision time to route the header flit and the switching time to switch a flit from input to output buffers (pass-thru delay)².



(a) Channel width.

(b) Network latency.

Figure 2. Channel width and network latency of a 64-node system.

Given Eq. (1), we further model the network performance based on DROI [9] (Diffractive Reflective Optical Interconnects) to find the number of optical links (connections) that can be established on a given volume. In addition, we model the channel cycle time (i.e. external router delay) using current technologies. Details on modeling can be found in [10]. To make the analysis more perceivable, we compare a 64-node system with state-of-the-art electrical PCB interconnects by HP [11] and optical interconnects based on DROI scheme.

The channel width for both schemes is depicted in Figure 2(a). As the figures show, optical interconnects can achieve 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 2(b).

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 2(b) assuming channel cycle time is fixed by internal router delay

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

² Current high-performance network routers use the maximum between the pass-thru and routing delays to determine the internal clock cycle and, thus, require multiple cycles to route and pass the header flit whereas subsequent flits will require only one cycle. This should not affect the results of this study as external clock cycle is assumed to dominate, unless otherwise specified, rather it strengthens our point that high-speed interconnects such as optical interconnects are urgently needed.

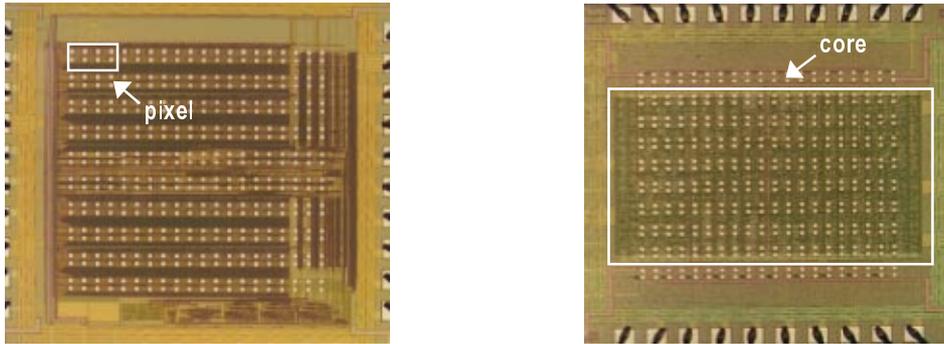
($T_c=5\text{ns}$ as in Intel Teraflop [4], for example). In this case, optics is still more than two times faster than electronics for the hypercube ($n=6$) configuration.

In addition to enabling the development of higher-speed routers and/or having wider channels, our results clearly state that optical interconnects can yield virtually constant network latency across various topologies. This suggests that optical-based networks provide more design flexibility, supporting larger domain of applications than that of the conventional electronic-based networks.

3. Design Issues and Performance Implications

The design and implementation of complex CMOS/SEED chips such as network routers are unconventional. 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, as shown in Figure 3(a). Hence, pixel-based designs are very useful for massively parallel applications which require simple functions such as signal processing [12], bit-slice arithmetic logic unit (ALU) [13], and simple switch [14]. 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 (see Figure 3(b)), this design paradigm is called “core-based.” Examples of core-based design are the WARRP core [15], the WARRP II router chip [16], the AMOEBA switch chip [17], and a 64-bit microprocessor core [18].

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 pure-CMOS chips. This section therefore evaluates these performance tradeoffs to validate the expected performance of optoelectronic chips.



(a) A pixel-based design
(the TRANSPAR chip—courtesy A. Sawchuk et al. [19]).

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

Figure 3. Comparison of CMOS/SEED chip design paradigms.

We first synthesize a core-based design using optoelectronic-compatible CAD tools to measure the negative effects of SEED integration. The *WARRP* (Wormhole Adaptive Recovery-based Routing via Preemption) router architecture [20] was chosen as a representative due to its architectural flexibility (this eases the syntheses of various WARRP router configurations—up to 50,000 transistors were synthesized to achieve more accurate results.) Second, we analytically model the wiring cost in terms of number of metal layers required by all SEEDs to connect with the core circuitry. By combining both steps, we can find the relationship between the wiring cost and its negative effects and also can extrapolate the results to further estimate the chip performance for larger designs using future technologies. The elaboration of this semi-empirical model can be found in [10].

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
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
# BGA 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

Using the model, we can predict the performance of core-based optoelectronic chips in comparison with pure-CMOS chips based on published semiconductor trends from SIA [21] and SEED integration trends from Krishnamoorthy [22], which are listed in Table 1.

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.

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 2. 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 0.07 μm technology because an additional metal layer is available. It is worth noting that the reduced transistor density effect is not critical as transistors are getting cheaper in time. 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. In effect, core-based chips have a potential to deliver its promises (large number of I/O pin-outs and high-speed signaling) at a nominal cost of reduced clock rate. For switch or network router implementations, large number of I/O pin-outs is more critical than the achievable clock rates. For example, a design that requires more than 4100 I/O pin-outs cannot be implemented by BGA packaging technology as suggested in Table 1.

Table 2. 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

4. System Demonstration

The last question regarding optoelectronic network router might be, “Can it really be implemented?” WARRP II—a fully adaptive deadlock-recovery multiprocessor network router—is our answer. This chip was implemented by the SMART Interconnects group at University of Southern California in collaboration with the Optoelectronic group at Lucent Technology. It features a scaled-down, fully functional version of the WARRP router architecture [20] integrating an array of 20x10 SEEDs on a 2x2mm² CMOS circuitry, via flip-chip bonding. The CMOS core circuitry was fabricated by MOSIS and later on flip-chip bonded by Lucent.

Each Self Electro-optic Effect Device (SEED) is 20x60 μm^2 with a horizontal pitch of 62.5 μm and a vertical pitch of 125 μm , respectively, and operates at 850nm wavelength. Recent experiments have shown that this promising technology can provide more than 47,000 devices on a 3.7x3.7mm² area in the near future [5], and each can currently operate at up to 2.48Gb/s with only 300 μW optical power input in dual-rail mode [6]. Using the HP14B CMOS process (a 0.5 μm , 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 4 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, flow controller, and channel preemption logic). 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 was extensively simulated using switch-level IRSIM (due to its complexity, exhaustive SPICE simulations were not possible given the limited design time frame and CPU resources). Maximum operation speed is estimated to be 25MHz.

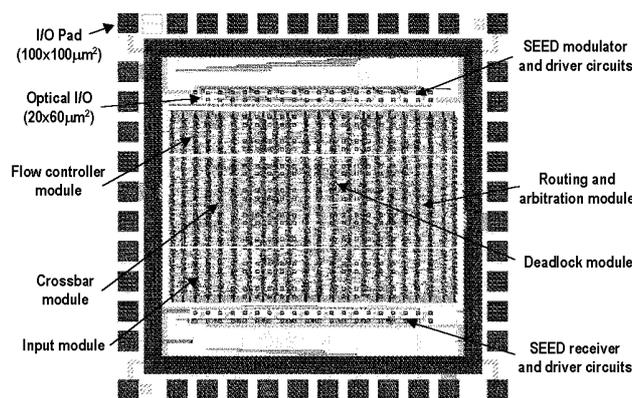


Figure 4. The WARRP II floorplan.

The demonstration system is divided into an electrical system and an optical system. The WARRP II testing board is central to the electrical system and is consisted of the WARRP II router chip, an FPGA chip, and a microcontroller board. We simplify the network interface functions (e.g., packetize/depacketize and collect some statistics) by using the FPGA chip from ALTERA Corporation. Similarly, a commercially available 68HC11-based microcontroller board with 32-Kbyte RAM running at 2 MHz is programmed to perform as a simple processing node (i.e., to generate and receive packets). This microcontroller allows us to develop the node controller with high-level languages such as BASIC or C. The PC interface is provided to exchange data with the FPGA and microcontroller, which will be displayed on the PC monitor.

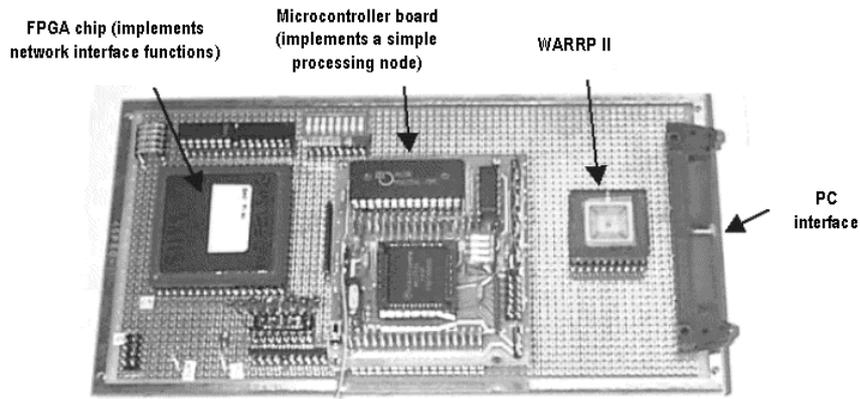


Figure 5. The WARRP II testing board.

The optical setup for the WARRP II chip is a 2-D system in which all components are placed on the base plate of size 12inx15in, providing unidirectional communication between chips as shown in Figure 6. This setup is designed and built in collaboration with the SIPI group at USC. The external light source is a semiconductor laser with adjustable external cavity which operates at a wavelength of $848\pm 5\text{nm}$. The generated wavelength is measured by an on-site monochromator to source a suitable wavelength of $\sim 850\text{nm}$ for the SEED arrays. An array of 20×10 light beams is generated via a diffractive optical element (DOE) which was fabricated through the DARPA/Honeywell CO-OP foundry run. This beam array is horizontally polarized through a $\lambda/4$ waveplate. Almost all power passes straight through a polarizing beam splitter (PBS) and is circularly polarized by another $\lambda/4$ waveplate. It is then focused by a Cooke triplet onto a SEED modulator array. The modulated beam array experiences the reverse optical path until they are vertically polarized by the $\lambda/4$ waveplate. This time, it changes its course of propagation to the right when it passes through the PBS. Consequently, this beam array will reach a SEED detector array of the next chip. Figure 6 shows the setup of two WARRP II chips communicating via free-space. Note that only simple functions of the WARRP II chip such as reset were successfully tested. We believe that transistor-level simulation on the WARRP II is indispensable but was not conducted due to our limited computing resources and, thus, led to the failure in some parts of the circuitry which could not be identified and corrected beforehand. However, this is by no means conclude that the complex optoelectronic chips are not feasible as other designs fabricated through the same foundry run were successfully tested [14, 17, 18].

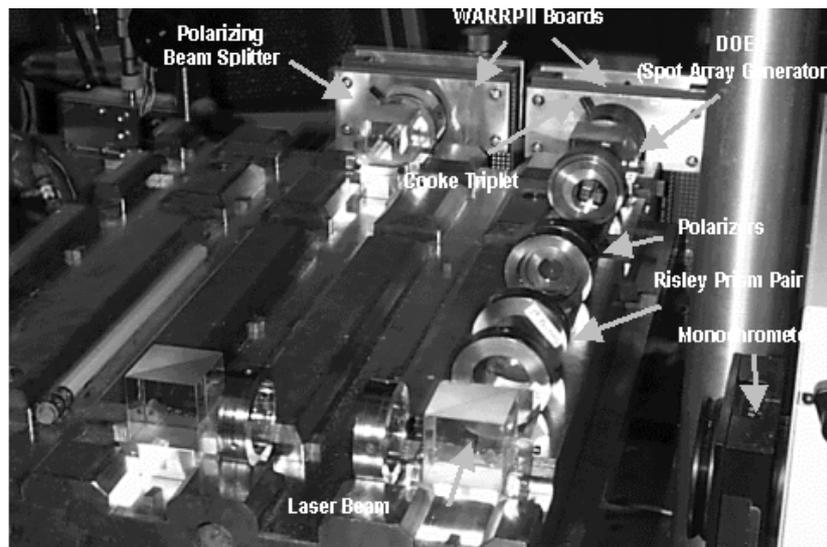


Figure 6. Optical setup for the WARRP II testing (courtesy Jen-Ming Wu et al., USC).

Although optoelectronic network routers based on CMOS/SEED integration technology show the potentials to alleviate the bandwidth problem currently limiting the performance of multiprocessor systems, they also raise some concomitant issues that must be effectively addressed (in addition to those described in Section 3) before this approach can flourish.

- Very dense optical I/Os integrated on top of CMOS circuitry significantly increase the on-chip power dissipation. As a result, chip packaging becomes more complicated because it requires very efficient cooling techniques. However, if not all the I/Os are utilized the packaging can then be simplified accordingly.
- The need for an external light source for SEEDs also increases the complexity of chip packaging. For instance, Diffractive Optics Elements (DOEs) are required to split and focus the light source to all the devices. Nevertheless, experiments in system packaging have made impressive progress, e.g., free-space optical module package [23] and DWDM module for the AMEOBA switch [17]. 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.
- 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 $300\mu\text{W}$ optical power per SEED (to operate at 2.48Gbps [6]), 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 [24, 25]—with modulation speed up to 800Mbit/s.

- The most serious issue is that the design and package of free-space optical system for multiprocessor networks would be extremely complex, even with regular topologies such as the k-ary n-cube class of networks. Remember that the system setup in Figure 6 only supports two WARRP II chips with simplex communication. Not to be discouraged, the researchers at University of California at Los Angeles have recently demonstrated the fabrication and packaging of micro-optic apparatus on a chip called Micro-Opto-Electro-Mechanical Systems (MOEMS) [26]. This technique has shed the light for small and rigid optoelectronic chip packaging for each network router which then can be connected together via fibers.

5. Ongoing Research

Designing network routers based on very high-bandwidth optical interconnects is unprecedented. Therefore, the internal router architectures must be carefully investigated to improve the link utilization as the increased bandwidth comes at high cost and, thus, must not be taken for granted. The WARRP architecture explores fully adaptive deadlock-recovery routing technique, the most efficient bandwidth utilization scheme to-date [27]. This section presents additional architectures that might further improve the link utilization over the WARRP architecture using fast arbitration among large number of virtual channels, flit-bundling technique, and efficient output buffer management scheme.

a) 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. Here, we 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 [28], as shown in Figure 7. 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 clock cycles.

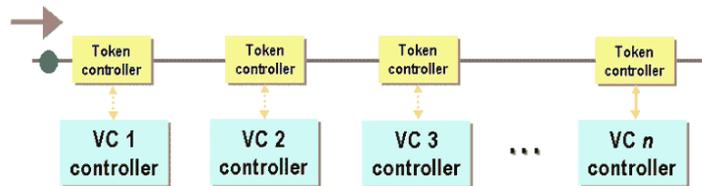
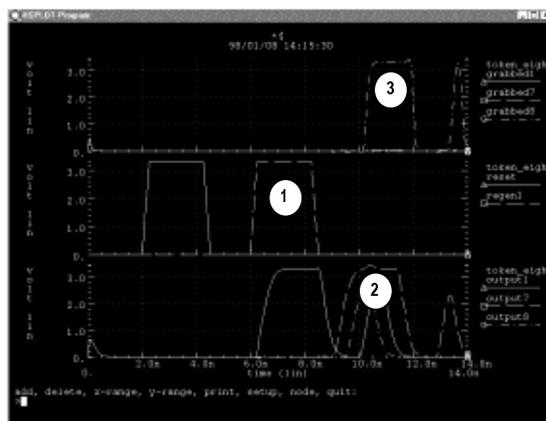


Figure 7. An illustration of asynchronous token arbitration scheme.

We design an 8-stage asynchronous token controller using HP14B 0.5 μ m CMOS process and simulate its functionality using a transistor-level simulator (HSPICE). Figure 8 shows that a token can propagate through all the virtual channel controllers (8VCs) in less than 4nS (approximately 0.5nS per stage). The token can then be grabbed by any VC that wants to access the physical channel within 0.75nS. Assuming a 500MHz-or-less network router, the virtual channel arbitration process would take only two cycles on the average, under this scheme.



- 1) A token is generated at the input of virtual channel controller #1 (VC1).
- 2) The token propagates to the input of VC8 after 3.2nS.
- 3) VC8 successfully grabs the token after 0.75nS.

Figure 8. Waveforms of an 8-stage asynchronous token circuit.

b) Flit-bundling Transfer Technique

We can further improve the router performance by fully pipelined the external flow control by means of a wave pipelining scheme [29] 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.”

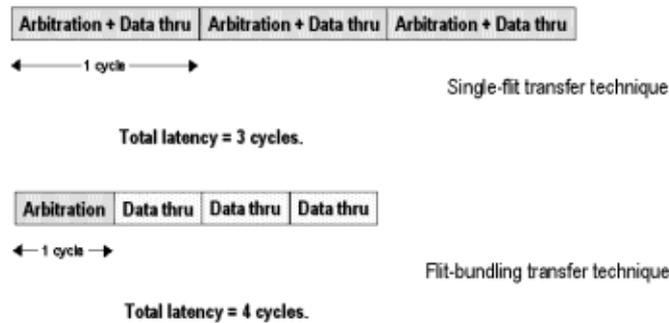


Figure 9. Message latency for single-flit and flit-bundling transfer techniques.

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, we 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 9. Depending on message lengths, the average message latency may not be significantly increased due to faster off-chip clock speed and lower arbitration overhead.

c) Delayed Buffer

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 we propose here is to use a buffer management scheme called “Delayed Buffer.”

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 required hardware 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 10.

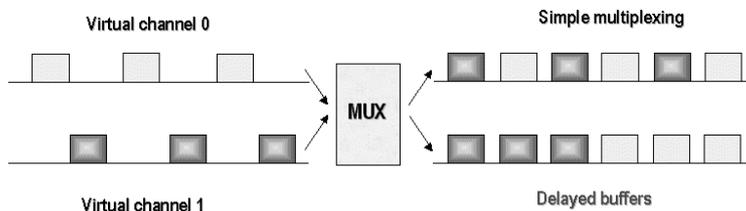


Figure 10. Simple multiplexing and delayed buffer schemes comparison.

To systematically evaluate the performance of the proposed techniques, we are developing a multiprocessor network simulator that incorporates both the WARRP architecture and the new techniques. The simulator will allow us to further investigate some important parameters and their performance implications, e.g., communication behavior, traffic load rate, message latency, throughput, channel switching frequency, optimal delay threshold, and optimal number of channels. Concurrently, we are implementing *SPIDERS* (Scalable-Performance Interconnects based on Distributed and Enhanced Routing Schemes)—a network router core employing the proposed techniques and the WARRP architecture based on FPGA technology. By doing so, we can assess the cost of implementation as well as estimate the router performance, before we can go further to develop a full-custom optoelectronic version. Note that the *SPIDERS* architecture is so flexible that it can be further developed as a proprietary local area network switch similar to the Myrinet switch [30]. Regardless of the results, we believe that *SPIDERS* should achieve higher performance than the non-optimized router architectures and should provide some guidelines useful for the development of high-performance multiprocessor network routers based on optoelectronic technologies.

6. Conclusion

Interconnection bandwidth provided by the conventional electrical interconnects is becoming more and more critical performance limiting factor in multiprocessor systems. For this reason, optical interconnects based on optoelectronic devices are being actively investigated as an alternative owing to its dense optical I/O pin-outs and high-speed signaling capabilities. Both features are preponderant to the development of high-performance multiprocessor network routers, a key solution to unleash the full processing power of the systems. In this work, we investigate the usefulness of optoelectronic router in multiprocessing environment and found that more network design flexibility can be achieved with optical networks. However, the design of complex optoelectronic chip such as network routers can affect the chip performance due to increased on-chip wiring complexity. As a result, transistor density and achievable on-chip clock rates of optoelectronic chips are lower compared to that of the pure-CMOS chips. Again, such negative effects are small price to pay compared to the increased bandwidth (which can be up to ten times higher). We eventually investigate the technology feasibility by implementing the WARRP II chip using CMOS/SEED integration technology. Although the WARRP

architecture is considered the most advanced of its kind, we notice that some improvements can be added. Therefore, we propose three new architectural techniques including asynchronous token arbitration scheme among large number of virtual channels, flit-bundling technique, and efficient output buffer management scheme which are suitable for high-speed optical links. Such techniques, while the increased performance is not yet conclusive, are at least believed to provide guidelines towards the development of high-performance multiprocessor network routers.

7. Acknowledgements

First and foremost, I would like to thank my former advisor, Dr. Timothy Mark Pinkston, whose support, guidance and encouragement have made the WARRP development possible. I also thank the SMART group fellows including Yungho Choi, Joon-Ho Ha, Wei Hong Ho, and Sugath Warnakulasuriya. Information regarding SEED technology from Dr. Ashok Krishnamoorthy and his group at Lucent Technology and the optical setup provided by Dr. Charles Kuznia and SIPI group at USC are greatly appreciated. The Altera FPGA tools donated by Joe Hanson is truly acknowledged. Last but not least, I deeply thank Miss Watcharee Veerakachen for her constant encouragement and review of this paper.

8. References

- [1] David Patterson et al., "A Case for Intelligent RAM," *IEEE Micro*, 17(2), 34-44 (1997).
- [2] Mike Galles, "SPIDER: A High-Speed Network Interconnect," *IEEE Micro*, 17(1), 34-39 (1997).
- [3] 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).
- [4] Joseph Carbonaro and Frank Verhoorn, "Cavallino: The Teraflops Router and NIC," *Proceedings of Hot Interconnects IV*, 157-160 (1996).
- [5] K. W. Goossen, "Optoelectronic/VLSI," *1997 OSA Spring Topical Meeting—Spatial Light Modulators Technical Digest*, 2-5 (1997).
- [6] 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).
- [7] William J. Dally, "Performance Analysis of k-ary n-cube Interconnection Networks," *IEEE Transaction on Computers*, 775-785 (1990).
- [8] Kazuhiro Aoyama and Andrew A. Chien, "The Cost of Adaptivity and Virtual Lanes in a Wormhole Router," *Journal of VLSI Design* (1994).
- [9] Karl-Heinz Brenner and Frank Sauer, "Diffractive-reflective optical interconnects," *Applied Optics*, 4251-4254 (1988).
- [10] Mongkol Raksapatcharawong, "Analysis and Implementation of Optoelectronic Network Routers," Ph.D. dissertation, CENG 98-20, University of Southern California, Los Angeles, December 1998.
- [11] 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).
- [12] 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).
- [13] D. S. Wills et al., "A Fine-Grain, High-Throughput Architecture Using Through-Wafer Optical Interconnect," *Journal of Lightwave Technology*, 1085-1092 (1995).
- [14] 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).
- [15] Timothy M. Pinkston, Mongkol Raksapatcharawong, and Yungho Choi, "WARRP Core: Optoelectronic implementation of network router deadlock handling mechanisms," *Applied Optics*, 276-283 (1998).
- [16] 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).
- [17] 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).
- [18] 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).
- [19] 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).
- [20] Timothy M. Pinkston, Yungho Choi, and Mongkol Raksapatcharawong, "Architecture and Optoelectronic Implementation of the WARRP Router," *Proceedings of Hot Interconnects V*, 181-189 (1997).
- [21] The NTRS document available on the WEB at <http://www.sematech.org:80/public/roadmap/index.htm>.
- [22] 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).
- [23] M. H. Ayliffe et al., "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).
- [24] 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.
- [25] L. M. F. Chirovsky et al., "Bottom-Emitting I²-VCSEL's for Flip-Chip Bonding to Smart Pixel IC's," *1998 IEEE/LEOS Summer Topical Meetings—Smart Pixel Session*, Postdeadline Papers PD003.
- [26] M. C. Wu, "Micromachining for Optical and Optoelectronic Systems," *Proceedings IEEE*, 1833-1856 (1997) (invited paper).
- [27] 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).
- [28] 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.
- [29] 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).
- [30] Nanette J. Boden et al., "Myrinet—A Gigabit-per-Second Local Area Network," *IEEE Micro*, 15(1), 1995.