# An Optoelectronic Multi-Terabit CMOS Switch Core for Local Area Networks

Honglin Wu, Amir Gourgy, and Ted H. Szymanski
*High Performance Networking Group, Department of Electrical and Computer Engineering*
*McMaster University, Hamilton, Ontario L8S 4K1, Canada*
*wuh4@mcmaster.ca, amir@grads.ece.mcmaster.ca, and teds@mail.ece.mcmaster.ca*

## Abstract

*Optoelectronic integrated circuits can support thousands of integrated optical laser diodes and photodetectors bonded to a high-performance CMOS substrate, and can be used in the design of Multi-Terabit optical Local Area Networks. This paper describes the design of an integrated optoelectronic CMOS crossbar switch to interconnect approx. 128 parallel fiber ribbon optical links, each with 12 channels clocked at 2.5 Gigabit/sec, to achieve a Local Area Network (LAN) with an aggregate capacity of 3.84 Terabits/second. A prototype switch core has been designed in 0.18µm CMOS technology. Logic optimization and synthesis was performed using the Synopsis logic optimization tools, and VLSI layout was performed using the Cadence 2002 tools. It is shown that using 0.18µm CMOS technology, a 3.84 Terabit crossbar switch for an optoelectronic LAN occupies approx. 1.78 sq. cm of real estate, and consumes approx. 90 watts of power.*

## 1. Introduction

The U.S. Accelerated Strategic Computing Initiative (ASCI) program aims to triple large-scale scientific computing performance every 18 months for the next decade. IBM estimates that a multiprocessing system designed to meet the ASCI performance target of 100 TeraFlops in the year 2004 will require an optical Local Area Network (LAN) with an aggregate bandwidth in the neighborhood of 300 Terabits per second [1]. Such a computing system will challenge existing approaches to LANs, and will almost certainly require highly reliable optical interconnection technologies.

A Multi-Terabit optical LAN that interconnects 128 workstations is shown in Figure 1 [2]. Each workstation has a parallel optical datalink operating at approx. 30 Gbits/sec to a centralized optoelectronic switch core. The centralized switch core interconnecting 128 processors is required to support an aggregate bandwidth of approx. 3.84 Terabit/sec. In principle, a large scale computing system to meet the ASCI target could exploit a large number of these optoelectronic LANs to achieve aggregate data rates in the hundreds of Terabits/sec.

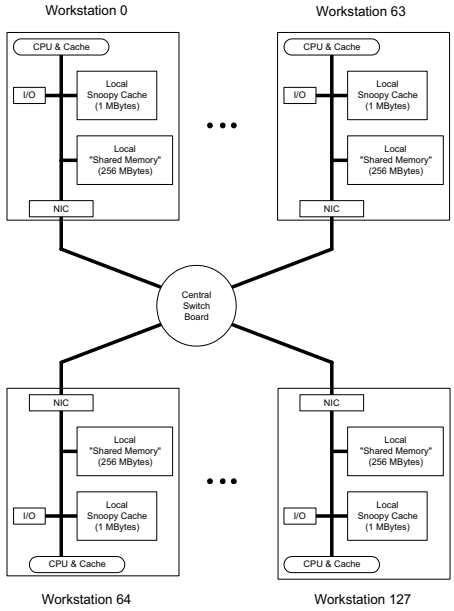



**Figure 1. Multi-Terabit local area network [2]**

In this paper, we present the detailed design for a single chip optoelectronic switch supporting approx. 3.84 Terabits/sec of bandwidth using 0.18µm[1] CMOS technology. The designs are validated using the Synopsis logic synthesis tools and the Cadence integrated circuit layout tools. All key components have been laid out in 0.18µm CMOS. The designs will scale to support

[1] All the designs, synthesis and simulation results reported in this paper were implemented using TSMC 0.18µm CMOS technology.

Proceedings of the 28th Annual IEEE International Conference on Local Computer Networks (LCN'03)
0742-1303/03 $ 17.00 © 2003 IEEE

Terabits of bandwidth using faster and smaller CMOS technologies.

Several Terabit range switches have been described in the literature recently, including the Tiny-Tera [3], the Cyclone [4], and the Saturn switches [5]. These prior switches do not exploit direct optical-to-electronic conversion, data switching, and electrical-to-optical conversion on a single substrate. The proposed switch exploits the emerging optoelectronic integrated circuit technology, which merges laser diodes and Photodetectors (PD) directly onto a silicon integrated circuit, and eliminates the need for costly and slow off-chip electronic signaling. This paper presents an analysis of the area and power requirements of a single chip optoelectronic switch, and establishes that a single chip Multi-Terabit switch is feasible.

This paper is organized as follows. Section 2 presents an overview of the single-chip optoelectronic switch core. Sections 3, 4, and 5 describe the design and implementation of the optical I/O, arbitration logic and crossbar array. Finally, section 6 provides concluding remarks.

## 2. Overview

As shown in Figure 2, the single-chip optoelectronic CMOS switch comprises a Photodetector (PD) array, input modules, arbitration logic circuits, a crossbar array, output modules, and a Vertical-Cavity Surface-Emitting Laser (VCSEL) array. Each component operates at a specific clock rate that determines its area requirement and power consumption.

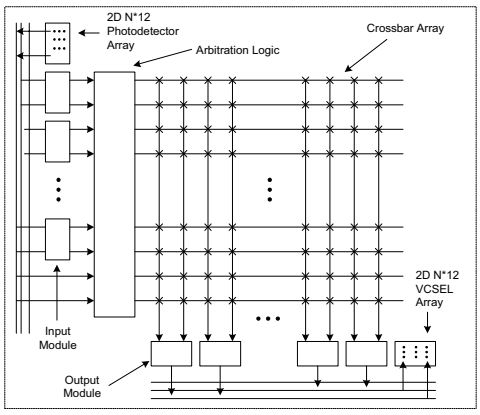



**Figure 2. Single-chip optoelectronic CMOS switch**

Each input port connects to an Agilent parallel fiber ribbon with 12 channels clocked at 2.5 Gigabit/sec for a port bandwidth of 30 Gbps. Incoming optical signals are converted to electrical signals by a 2D photodetector array. With 128 input ports, the Photodetector array requires $128 \times 12 = 1536$ Photodetectors, and the switch achieves an aggregate capacity of 3.84 Terabit/sec.

Each output port connects to an Agilent parallel fiber ribbon with 12 channels clocked at 2.5 Gigabit/sec for a port bandwidth of 30 Gbps. Outgoing electrical signals are converted to optical signals by a 2D VCSEL array. With 128 output ports, the VCSEL array also requires $128 \times 12 = 1536$ VCSELs.

The switch uses a combined input and output queueing (CIOQ), where each input module contains an input queue and each output module contains an output queue.

The switch core is based on a conventional multi-channel broadcast-and-select architecture [2] such that each input port has a direct path to every output port. The arbitration logic circuits assign contention-free paths from input ports to output ports across the crossbar array.

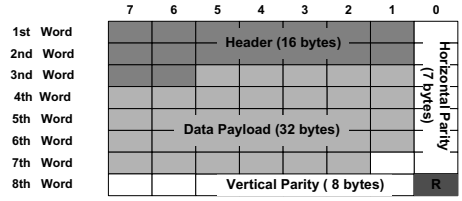



**Figure 3. Typical 512 bits packet format [2]**

By default, the switch uses a fixed-length 64-byte packet format, as shown in Figure 3, but can also be reconfigured to use variable-length Ethernet packets. A 16-byte header contains all datalink protocol information.

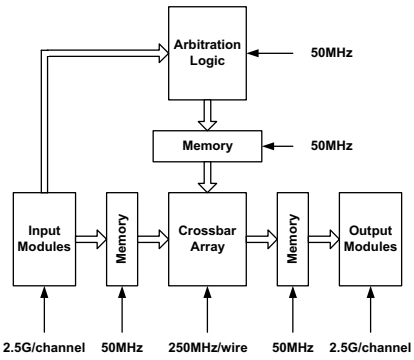



**Figure 4. Clock rate requirement for components**

At 30 Gbps rate per input port, the time to process a 64-byte packet is at most 17.1 ns. Therefore, the real clock rate required to maintain the 30 Gbps rate per input module is approximately 58.5 MHz, however, 50 MHz clock rate is used for simplicity in the rest of the paper. As shown in Figure 4, the switch is pipelined such that a memory stage is inserted between all major components. The buffer memory and the arbitration logic circuits are

clocked at 50MHz. Note that the crossbar array is clocked at 250 MHz/wire as explained next.

The crossbar array must support an aggregate bandwidth of 30 Gbps per input port; for a $128 \times 128$ switch, this totals an aggregate 3.84 terabit/sec. There is a trade off in the crossbar design between area and speed requirements. On the one hand, clocking the crossbar at a slow speed (e.g., 50 MHz) would require large amount of wiring to support the required bandwidth (30 Gbps per input port); On the other hand, clocking the crossbar at higher clock rate, reduces the amount of wiring (area) at the expense of imposing stringent speed requirement. We designed a crossbar array with 120 wires per input port clocked at 250 MHz, which meets the bandwidth requirements of 30 Gbps per input port. The design of the crossbar array is explained in detail in section 5.

The memory stages were implemented using positive edge DFFs; specifically, DFFPQ1 available in TSMC 0.18µm technology standard cell library [6] was used.

Referring to Figure 4, the memory stage between input modules and the crossbar array stores the data (256 bits for each packet) of the 128 packets and requires $128 \times 256$ high speed DFFs, the memory stage between the arbitration logic and the crossbar array stores the control information (128 bits for each output module) and requires $128 \times 128$ high speed DFFs, and the memory stage between the crossbar array and output modules requires $128 \times 256$ high speed DFFs. Each DFFPQ1 has an area of approximately $58\,um^2$ [6] and consumes 4.7 uW when clocked at 50 MHz according to Synopsys synthesis results. Therefore, the total area required to implement 3 memory stages in Figure 4 is $4.8\,mm^2$. The memory power consumption is 0.39 W (at 50 MHz). Based on Synopsys synthesis results, the 50 MHz clock rate requirement is easily satisfied for all the memory stages.

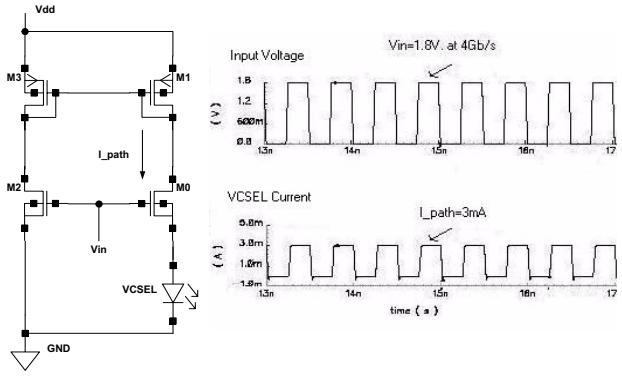
The design uses combined input and output queueing (CIOQ), with input queues of size 8 packets per input module, output queues of size 4 packets per output module, to avoid buffer overflow. These queues require additional memory. The input side queues require $30.4\,mm^2$ area and consume 2.5 W (at 50 MHz). The output side queues require $15.2\,mm^2$ area and consume 1.2 W (at 50 MHz). Therefore, in total, the CIOQ packet memory requires $45.6\,mm^2$ area and consumes 3.7 W (at 50 MHz).

## 3. Design of Optical I/O

Referring to the $N \times N$ switch shown in Figure 2, the input modules require an array of $N \times 12$ Photodetectors (PDs), pre-amplifiers, amplifiers and threshold detectors, and the output modules require an array of $N \times 12$ VCSELs and VCSEL drivers.

The optical PDs can be arranged in a tightly packed 2D array located at one corner of the die. Optical channels are arranged in a 2D array with X and Y pitches of $31.25\,um$ and $125\,um$ respectively. For $N = 128$, the PD array requires approx. $6\,mm^2$, arranged as $16\,mm \times 0.375\,mm$. The VCSEL array will require a similar area [2].

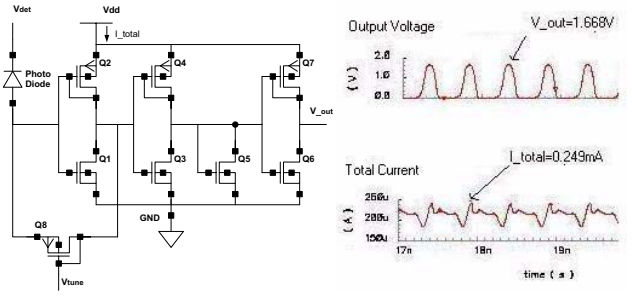
To evaluate the power consumption of the proposed switch, the VCSEL driver circuit in [7] was scaled down to TSMC 0.18µm CMOS technology and implemented using the Cadence Spectre™ CAD tool.




(a) (b)

**Figure 5. (a) 0.18µm VCSEL driver circuit [7] and (b) performance at 4Gbps**

The driver circuits in Figure 5 (a) are designed to supply 3 mA of current to each VCSEL at 4 Gbps. Simulations of the driver operation at 4 Gbps is shown in Figure 5 (b). The driver exhibits excellent performance at 4 Gbps using 0.18µm CMOS technologies, exceeding the requirement of 2.5Gbps. The total circuit current is 3.34 mA per VCSEL at 1.8 V, and the power dissipated per VCSEL is approx. 6 mW. The area is $12.2\,um^2$ per VCSEL driver circuit.




(a) (b)

**Figure 6. (a) 0.18µm PD receiver circuit [8] and**

**(b) performance at 4Gbps**

The transimpedance amplifier circuits presented in [8] were also scaled to 0.18µm CMOS and implemented with the Cadence Spectre™ CAD tool. The transimpedance amplifiers in Figure 6 (a) are designed to detect 8 uA of current received by Photodetector, and generate digital logic signals at 4 Gbps. Simulations of the amplifier operation at 4 Gbps is shown in Figure 6 (b). The amplifier exhibits excellent performance at 4 Gbps, exceeding the requirement of 2.5 Gbps. The circuit current is 0.5 mA at 1.8 V, and the power dissipated per PD receiver is approx. 1 mW. The area is $1.2\,um^2$ per PD receiver.

These designs indicate that each VCSEL/PD pair will dissipate approx. 7 mW and require $7.8\text{k}\,um^2$ area. For $N = 128$ input/output ports, there are 1,536 VCSEL/PD pairs that consume an area of $12\,mm^2$ and dissipate a power of 10.8 Watts.

The optoelectronic I/O interface results in a substantial area and power savings compared to traditional LVDS electronic I/O. A LVDS output pad clocked at 2.5 Gbps in 0.18µm technology typically requires at least $20\text{k}\,um^2$ area and dissipates 10 mw per Gbps. A LVDS input pad typically requires at least $20\text{k}\,um^2$ area and dissipates 10 mw per Gbps. Using LVDS electrical I/O signaling, the switch would require at least $61\,mm^2$ area and would dissipate approx. 76.8 W power.

## 4. Design of Arbitration Logic

In this section we describe the design of arbitration logic, analyze its computational complexity, and report simulation results to corroborate our analysis.

Under the traditional logic gate model, where all processing is done with binary (2-input) logic gates, the cost complexity is expressed in terms of binary gates, and the delay complexity is expressed in binary gate delays. The power dissipation is then derived from the cost complexity for a specific clock rate.

Although the traditional logic model works well for technologies above 0.1µm [9], it ignores wire RC delay that becomes significant for technologies below 0.35µm (Figure 7). Unfortunately, Synopsys synthesis tools don't support wire RC delay models. In addition, estimating wire delay from Cadence tools is difficult. To overcome these shortcomings, we use conservative wire delay estimates available from Semiconductor Industry Association (SIA) Roadmap [9]. This approach was also followed in [10]. According to [9] and [10], in 0.18µm CMOS technology, the average delay due to copper wiring with low dielectric constant ($\kappa$) materials in interlayer dielectrics is approximately 100% of the gate delay (Figure 7). Therefore, It can be estimated that the average capacitance of the wire is equal to the average capacitance of a gate, and the power of a circuit considering the wiring capacitance approx. doubles.

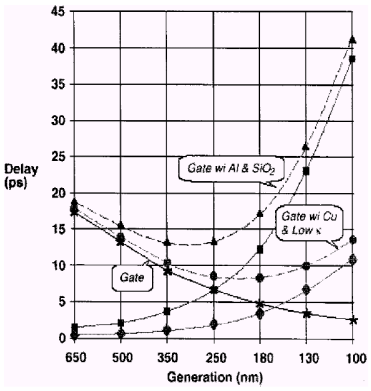



**Figure 7. Gate and wire delay versus technology [9]**

In the following, we derive cost complexity under the traditional logic gate model, report Synopsys synthesis results, and show that the arbitration delay still meets the 20 ns requirement when scaled by 100% to account for wiring delay.

The arbitration logic circuits comprise address filter arrays, Prefix Computation Rankers (PCR), and comparator arrays. Arriving packets are served in a FIFO order, such that arbitration logic circuits resolve output port contention using a simple ranking mechanism.

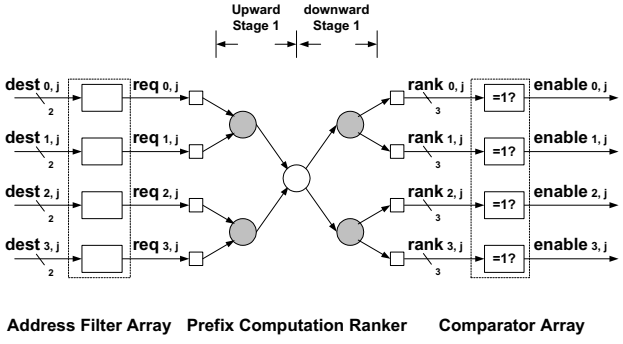



**Figure 8. An arbitration logic circuit for single output port in a $4 \times 4$ switch**

Figure 8 illustrates how output port contention is resolved for a *single* output port by an arbitration logic circuit in a $4 \times 4$ switch. In a $N \times N$ switch, N arbitration logic circuits, with the same structure shown in Figure 8, would be used. The destination address (7 bits) of the packet at head-of-line of input port $i$ ($0 \le i \le 127$), denoted $\text{dest}_i$, is broadcasted to all address filters for all output ports. The address filters generate request signals,

denoted $req_{i,j}$ from input port $i$ ($0 \le i \le 127$) to output port $j$ ($0 \le j \le 127$).

Referring to Figure 8, the packet at the head of each input module is processed by an address filter, which converts a 7 bit address to a single '1' bit request signal denoting the described destination $j$ ($0 \le j \le 127$) by '0' or '1'. It can be verified that each address filter requires approximately 9 binary gates and is 4 gates deep. Each output port has 128 address filters that require 1.15k binary gates.

In TSMC 0.18µm CMOS technology, the average power of a binary gate is $30nW/MHz$ (assuming a single standard load) [11], the average propagation delay is 82 ps, and area is approx. $16\,um^2$ [6]. Therefore, the area of an address filter array for one output module is $0.018\,mm^2$, the delay is 0.33 ns, and the power is 1.7 mW (at 50MHz). Therefore, the address filters for all 128 output modules require $2.3\,mm^2$ area, have 0.33 ns delay and consume 0.22 W (at 50MHz) power.

For illustration purposes, assume that all packets are destined to the same output port. The request signals $req_{i,j}$ are then ranked by ranker circuits. The ranker circuit assigns a unique number (rank) to each request signal $req_{i,j}$ such that smaller numbers correspond to higher priority. A prefix computation ranker [12] is used. The prefix computation ranker assigns a rank denoted $rank_{i,j}$ to each request signal $req_{i,j}$ from input port $i$ ($0 \le i \le 127$) to output port $j$ ($0 \le j \le 127$) as

$$rank_{i,j} = (\sum req_{0,j} \cdots req_{i,j})\ And\ (req_{i,j} = 1) \qquad (1)$$

In Figure 8, the ranker is implemented using a binary adder tree. The computations performed at each node are shown in Figure 9.

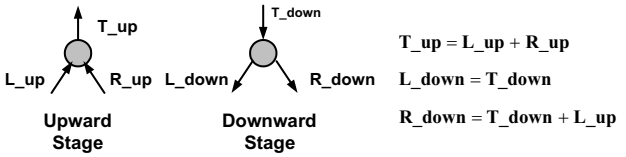



**Figure 9. Node of prefix computation ranker**

Let $C_+(M,2)$, $D_+(M,2)$, and $P_+(M,2)$ denote the binary logic gate cost, binary logic gate delay, and binary logic power dissipation respectively of the add operator + with a $M$ bit result. Let $C_{PCR}(N,M,2)$, $D_{PCR}(N,M,2)$, and $P_{PCR}(N,M,2)$ denote the cost, depth, and power dissipation respectively of the Prefix Computation Ranker (PCR) of degree 2, with $N$ elements in vectors $X$ and $Z$, each with $M$ bits. The computation is easily generalized to nodes with degrees higher than 2; for example, a 4-degree computation would use a tree with $Log_4 N$ levels, at the expense of increased computation within the nodes.

For $N = 2^k$, $k$ an integer, the parallel prefix function $P_+^{(n)} : A^n \mapsto A^n$ on an $N$ element vector with add operator + can be implemented by a circuit with the following cost, depth, and power upper bounds[2] derived from [13] measured in average area, delay, and power of a binary gate in the unit of $um^2$, ps, and nW/MHz respectively:

$$\begin{aligned} C_{PCR}(N,M,2) &\le (2N - LogN - 2) \cdot C_+(M,2) \\ D_{PCR}(N,M,2) &\le 2LogN \cdot D_+(M,2) \\ P_{PCR}(N,M,2) &\le (2N - LogN - 2) \cdot P_+(M,2) \end{aligned} \qquad (2)$$

The addition function can be realized using a ripple-carry adder with the following cost, depth, and power dissipation upper bounds as established in [13] measured in average area, delay, and power of a binary gate in the unit of $um^2$, ps, and nW/MHz respectively:

$$\begin{aligned} C_+(M,2) &\le 5M - 3 \\ D_+(M,2) &\le 3M - 2 \\ P_+(M,2) &\le 5M - 3 \end{aligned} \qquad (3)$$

The notation 'upper bound' $f(N) = O(g(N))$ is used to denote that there exist constants c and $N_0$ such that $f(N) \le cg(N)$ for all $N \ge N_0$. Each node of the $N$ input ports prefix computation ranker contains $M = LogN + 1$ bits ripple-carry adder. The $N$ input ports prefix computation ranker has the following cost, depth, and power dissipation bounds correspondingly:

$$\begin{aligned} C_{PCR}(N,M,2) &= O(NLogN) \\ D_{PCR}(N,M,2) &= O(Log^2 N) \\ P_{PCR}(N,M,2) &= O(NLogN) \end{aligned} \qquad (4)$$

Equations (2) and (3) can be used to estimate the area, delay, and power of the prefix computation ranker, using $N = 128$, $M = LogN + 1$, and the TSMC data for the typical binary gate. This estimation will ignore wiring delays and wiring capacitance. The effect of wire delays and wire capacitance will be modeled separately later.

According to equations (2) and (3), the area of one ranker is $0.15\,mm^2$, the delay is 25.3 ns, and the power is 13.7 mW (50MHz). All 128 rankers for the entire switch require $19.2\,mm^2$ area, have 25.3 ns delay and consume 1.75 W power without consider wire effect.

After the ranking is completed, each output port must examine up to 128 ranks and select the rank $rank_{i,j}$ with

[2] All logarithms are to the base 2 unless otherwise indicated

the highest priority, or the smallest ranking (>0). This task can be accomplished with 128 simple digital comparators. Each comparator tests $\text{rank}_{i,j}$ for a rank of 1, and requires approximately 8 binary logic gates with 4 logic levels. Using the TSMC 0.18µm data for a typical gate, it can be easily verified that the 128 comparators required for each output port, have an area of $0.016\,mm^2$, a delay of 0.33 ns, and a power of 1.5 mW when clocked at 50 MHz. All 128 comparator arrays for the entire switch require $2\,mm^2$ area, have 0.33 ns delay and consume 0.19 W power when clocked at 50 MHz, without considering wire effect.

Based on the previous complexity analysis of the address filter array, prefix computation ranker, and comparator array, for 128 input ports and one single output port, the area of the arbitration logic circuit is $0.18\,mm^2$, the delay is 26 ns, and the power dissipation is 17 mW (50MHz). For the complete switch, the area of 128 arbitration logic circuits is $23\,mm^2$, delay is 26 ns, and power dissipation is 2.2 W without considering optimization and wire capacitance.

**Table 1. Synopsys synthesis results of an arbitration logic circuit, clock at 50MHz**

| N | Without Optimization | | | Optimization for Speed | | |
|---|---|---|---|---|---|---|
| | Area (um*um) | Power (mW) | Delay (ns) | Area (um*um) | Power (mW) | Delay (ns) |
| 4 | 1451 | 2.54 | 4.82 | 1935 | 2.50 | 1.53 |
| 8 | 3049 | 3.63 | 5.39 | 5326 | 3.98 | 1.65 |
| 16 | 9953 | 5.28 | 9.7 | 13795 | 5.81 | 3.82 |
| 32 | 24263 | 6.92 | 12.05 | 26784 | 7.31 | 4.86 |
| 64 | 55833 | 8.61 | 16.4 | 61407 | 10.29 | 5.57 |
| 128 | 126017 | 12.13 | 26.42 | 119947 | 15.01 | 7.03 |
| 256 | 278167 | 20.01 | 27.19 | 259723 | 26.39 | 8.16 |

To corroborate the complexity analyzes, we implemented the arbitration logic circuit in the VHDL hardware description language by using Synopsys design tools for $N = 4,8,...256$. Table 1 provides Synopsys synthesis results of the area, the power dissipation (at 50 MHz), and the delay for an arbitration logic circuit as reported by Synopsys design tools. It must be noted that Synopsys synthesis doesn't model wire delay and wire capacitance. The Synopsys Design Analyzer™ takes several hours to synthesize arbitration logic (optimized for speed) for one output module on a SUN Blade 1000™ server.

The area and delay comparisons between the Synopsys synthesis results without optimization and the complexity analysis results for an arbitration logic circuit ($N = 8,16,...128$) are shown in Figure 10. The analytic area results are 33% larger than the Synopsys synthesis results since the complexity cost model is built on an upper bound. The analytic area results are within 5% of the Synopsys synthesis results (ignoring wire delay and wire capacitance). These results shows that the complexity models used in complexity analysis are reasonable and practical.

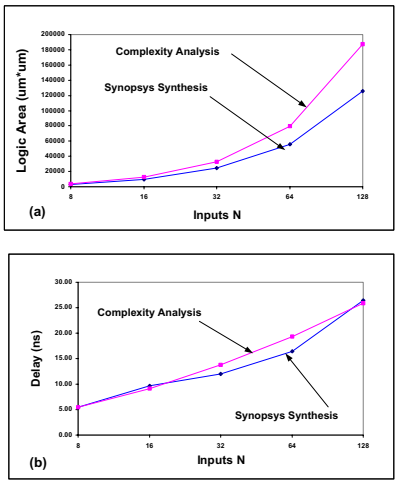



**Figure 10. Synopsys synthesis and analysis for (a) area and (b) delay for an arbitration logic circuit**

Referring to Table 1, the Synopsys synthesis results when optimized for speed use almost same area, and are much faster. Based on the optimized Synopsys synthesis results, the area of an arbitration logic circuit for 128 input ports and a single output port is $0.12\,mm^2$. The speed of arbitration is mainly determined by the speed of the prefix computation ranker. Synopsys optimized synthesis indicates that the maximum delay along the critical path is 7 ns (ignoring wire delay and wire capacitance). The power consumption is 15 mW clocked at 50 MHz. The previous results are for a single output port. Therefore, all 128 arbitration logic circuits have an area of $15.36\,mm^2$, a delay of 7 ns, and a power dissipation of 1.92 W without considering wiring effect.

Referring to Figure 2, the 7 bits destination of the packet at the head of each input module is broadcasted to 128 address filters. Buffers are required to drive the signals. It can be verified that a buffer for one wire requires approximately $880\,um^2$ area, has 0.88 ns delay, and consumes 39 uW (at 50 MHz) power. The entire switch needs $128 \times 7$ buffers. The total area is $0.79\,mm^2$, the delay is 0.88 ns, and the power is 35 mW (at 50 MHz).

Considering the buffers, the 128 arbitration logic circuits have an area of $16.15\,mm^2$, a delay of 7.88 ns, and a power dissipation of 1.96 W without considering wiring effect.

By scaling the logic gate delay by 200% to account for wire delays, the estimated arbitration delay increases from 7.88 ns to 15.76 ns. This delay easily satisfies our requirement for arbitration delay of 20 ns.

## 5. Design of Crossbar Array

In this section we describe the design of the crossbar array, analyze its area, power, and delay complexity, and report simulation results. Each crossbar column is implemented using tri-states, as shown in Figure 11 (a).

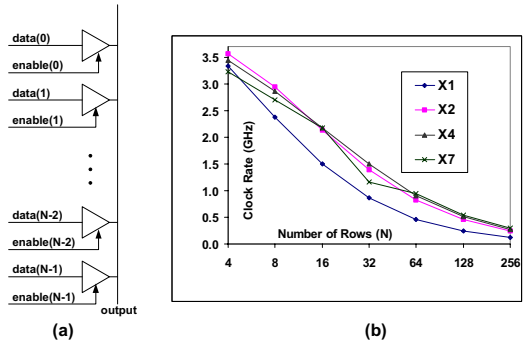



**Figure 11. (a) A crossbar column structure and (b) delay comparison of Synopsys synthesis results**

Each input port maintains a data rate of 30 Gbps. Assume for now that the crossbar switch supports a data rate of 500 Mbps per 'wire' (i.e. a single bit datapath), without considering wire delays and wire capacitance. Using the data from [9] and [10], we can estimate the effects of wiring to double the delay and power. Therefore, the crossbar switch can support a data rate of 250 Mbps per wire.

To achieve high speed, the tri-state buffers were implemented using standard cell BUFTD7 (3-state buffer with active-low enable and driving strength 7) available in TSMC 0.18 CMOS technology [6]. This selection is based on the Synopsys synthesis results shown in Figure 11 (b). Figure 11 (b) verifies that the crossbar can operate at 500 Mbps without wire delay and 250 Mbps with wire delay with driving strength of 7.

The area of a single tri-state buffer implemented using BUFTD7 is approx. $53\,um^2$ [6]. Because each column requires 128 tri-state buffers in Figure 11 (a), the area per column (1 wire) is $6.8\text{k}\,um^2$ which is perfectly matched with Synopsys synthesis result.

There are two main sources of power dissipation in the crossbar array: switching and internal power. Switching power is dissipated by charging and discharging the total load capacitance of a gate. The switching power ( $P_c$ ) of a gate can be calculated using equation (5) from [15]

$$P_c = \frac{V_{dd}^2}{2} \sum_{\forall nets(i)} (C_{load_i} \times TR_i) \qquad (5)$$

where $C_{load_i}$ is the capacitive load of net i; $TR_i$ is toggle rate of net i in transitions per second (normally set as half of clock rate $f_{clk}$); and $V_{dd}$ is the power supply voltage, which is 1.6 V for 0.18µm CMOS technology.

The switching power of the crossbar is determined by the $N$ rows (120 wires each) and $N$ columns (120 wires each). For a crossbar column in Figure 11 (a), there are 128 input nets, 128 enable nets, and 1 output net. Each 3-state buffer has an input port capacitance of approximately 7.9 fF and an output port capacitance of approximately 22.8 fF [6].

Referring to Figure 2 and section 2, each input module drives 120 row wires, where each row wire drives the input capacitance of 128 tri-state buffers. The total row capacitance for one wire is $7.9\,\text{fF} \times 128 = 1.0\,\text{pf}$. Each output module receives 120 column wires, where each column wire drives the output port capacitance of 128 tri-state buffers. The total column capacitance for one wire is $22.8\,\text{fF} \times 128 = 2.9\,\text{pf}$. To include the effect of wiring, these capacitances are doubled.

Clock rate and toggle rate are set as 250 MHz and 0.5/ns for calculation. For one column wire, substituting the previous parameters in equation (5), the switching power for a column (one wire) is 0.464 mW without considering wire capacitance and the switching power for a row (one wire) is 0.16 mW without considering wire capacitance.

The internal power of a driving cell is the power dissipated in charging and discharging of any existing capacitances internal to the cell. The internal power ( $P_{\text{int}}$ ) can be calculated using equation (6) from [15]:

$$P_{\text{int}} = E_Z \times TR_Z \qquad (6)$$

$$E_Z = f\left[C_{load_i}, WeightAvg_{(trans)}\right]$$

$$WeightAvg_{(trans)} = \frac{\sum_{i=A,B} TR_i \times Trans_i}{\sum_{i=A,B} TR_i}$$

where $E_Z$ is internal energy for output Z as a function of input transitions and output load, which is defined in the technology library; $TR_Z$ is the toggle rate of output pin Z; $TR_i$ is the toggle rate of input pin i; $Trans_i$ is the transition time of input i; and $WeightAvg_{(trans)}$ is the weighted average transition time for output Z.

The parameters in equation (6) were obtained from Synopsis library, and internal power dissipated in one BUFTD7 is 9.8 uW (at 250 MHz). Therefore, the total internal dissipated is 1.25 mW for one crossbar column (1 wire).

Based on our analysis, the total power dissipation in a $128 \times 128$ crossbar array with 120 wires is 28.78 W (at

250 MHz), including internal and switching power, which is extremely close to the results obtained from Synopsys synthesis derived from Table 2. This result doesn't include wire capacitance, which will double the power.

To compute the delay across the crossbar array, a CMOS timing generic model was used [16] such that

$$D_{Total} = D_1 + R_{driver}(C_{wire} + C_{pin}) \qquad (7)$$

For BUFTD7, $D_1$ is approx. $0.164ns$ for rising delay [6]. $C_{wire} + C_{pin}$ is the total capacitance on the net, and $R_{driver}$ is the output resistance of the cell. For rising delay, the rising resistance is about $0.5\,\mathrm{K}\Omega$ [16] and the total capacitance is 3 pf without considering wire capacitance. Therefore, the transition delay is 1.7 ns, which is within 8 percent of synthesis (1.85 ns) in Table 2. The wire capacitance will double the delay.

**Table 2. Synopsys synthesis results of a crossbar column, clock at 250MHz**

| X7 | without optimization=mininum area optimization | | | |
|---|---|---|---|---|
| N | Cell Area (um*um) | Power Comsumption (mW) | Speed | |
| | | | Delay (ns) | Clock (GHz) |
| 4 | 211.41 | 0.140 | 0.31 | 3.226 |
| 8 | 422.82 | 0.195 | 0.37 | 2.703 |
| 16 | 845.65 | 0.252 | 0.46 | 2.174 |
| 32 | 1691.30 | 0.463 | 0.86 | 1.163 |
| 64 | 3382.59 | 0.925 | 1.06 | 0.943 |
| 128 | 6765.18 | 1.849 | 1.85 | 0.541 |
| 256 | 13530.37 | 3.698 | 3.44 | 0.291 |

Table 2 shows the Synopsys synthesis results clocked at 250 MHz without considering wire effect. The delay of the crossbar increases linearly as N increases. Conversely, the delay decreases as the driving strength of Tri-State buffers increase as shown in Figure 11 (b).

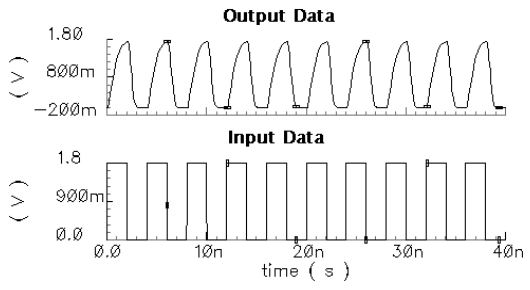



**Figure 12. Performance of conventional crossbar column at 500MHz, N=128 from Cadence**

Figure 12 shows the simulation result (after layout) of a crossbar column from Cadence tools. Both Synopsys synthesis (Table 2) and Cadence simulation (Figure 12) shows that 500 MHz clock rate can be achieved for each crossbar wire without considering wire delay and wire capacitance. Therefore, a 250 MHz clock rate should be achievable after considering the wiring effect.

To maintain the 30 Gbps data rate per input port, given 250 Mbps per crossbar wire, the crossbar data path must include 120 wires, and the area and power of crossbar will increase 120. Therefore, the $128 \times 128$ switch requires 128 crossbar columns, each 120 bits wide. Using the Synopsys synthesis results from Table 2, the 128 crossbar columns will have the area of $103.9\,mm^2$ and the power dissipation of 28.4 W. The power dissipation will be doubled when consider wiring effect.

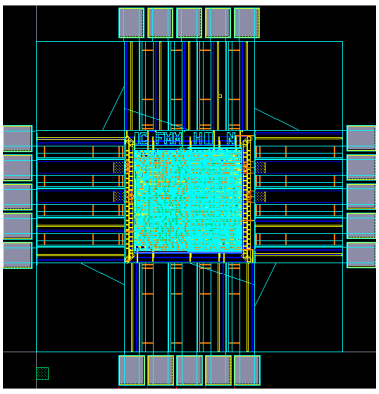


**Figure 13. Layout of a prototype crossbar switch core in 0.18µm CMOS with Cadence layout tools**

The layout of a single $128 \times 1$ prototype crossbar chip is shown in Figure 13, which has been submitted to Canadian Microelectronics Corporation for fabrication.

## 6. Summaries and Conclusions

The design of an optoelectronic CMOS/VCSEL single chip $128 \times 128$ switch core that interconnects 128 parallel fiber ribbons, each with 12 fibers clocked at 2.5 Gbps, to switch approx. 3.84 Terabits/sec was described. This switch was designed and implemented using of the state-of-art TSMC 0.18µm 1P6M CMOS technology. The Synopsys and Cadence CAD tools were used.

The clock requirements of the components and area/power results from synopsys synthesis and Cadence simulation tools of our proposed design are summarized in Table 3. The required clock rate for each component is derived in section 2. The area and power results of pipelined memory and CIOQ memory are derived in section 2. The area and power results of optical I/O are derived in section 3. The area and power results of arbitration logic are derived in section 4. The area and power results of crossbar array are derived in section 5. It

is clear that the digital switching logic forms the dominant power and area constraints.

**Table 3. Optoelectronic CMOS switch chip summary**

| Components | Required Clock Rate MHz | Optmizated Clock Rate MHz | Clock Rate with Wire Delay MHz | Area mm*mm | Power at Required Clock W |
|---|---|---|---|---|---|
| Pipelined Memory | 50 | 3000 | 1500 | 4.8 | 0.39 |
| CIOQ Memory | 50 | 3000 | 1500 | 45.6 | 3.7 |
| Opticla I/O | 2500 | 4000 | 4000 | 12.0 | 10.8 |
| Arbitration Logic | 50 | 127 | 64 | 16.15 | 1.96 |
| Crossbar Array | 250 | 500 | 250 | 103.9 | 28.4 |
| **Total** | | | | **178** | **45** |

The switch core can fit on a $1.34\,cm \times 1.34\,cm$ die (Table 3). When considering wire delay and wire capacitance, the power result of each component in Table 3 is doubled approximately. Therefore, the switch core consumes 90 W totally at the required clock rates when considering the wire delay and capacitance. These area and power figures are quite feasible under the state-of-art CMOS technology [9].

The centralized switch core can interconnect 128 workstations using an optoelectronic Local Area Network with an aggregate capacity of 3.84 Terabits/sec. A prototype switch core design that contains all major components of the complete design has been submitted to Canadian Microelectronics Corporation for fabrication.

## 7. Acknowledgement

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant # 121602, by the L.R. Wilson/Bell Canada Enterprises Chair in Data Communications at McMaster University, by the Nortel Networks Scholarships (OGSST), and by the Canadian Microelectronics Corporation (CMC) through computing equipment and CMOS IC fabrication services.

## 8. References

[1] A. Benner, E. Schenfeld, J. Sauer, L. Rudolph, T. Sterling, and T. H. Szymanski, "Design Options for Interconnecting a 100+ TFlop/sec Parallel Supercomputer in 2004", Paper and Panel Discussion, *Fifth International Conference on Massively Parallel Processing using Optical Interconnects*, Las Vegas, Nevada, USA, June 15-17 1998.

[2] T. H. Szymanski, A. Au, M. Lafrenière-Roula, V. Tyan, B. Supmonchai, J. Wong, B. Zerrouk, and S. T. Obenaus, "Terabit Optical Local Area Networks for Multiprocessing Systems", *Applied Optics*, *Special Issue on Massively Parallel Optical Interconnects for Multiprocessor Systems*, Vol. 37, No. 2, Jan. 1998, pp. 264-275.

[3] N. McKeown, M. Izzard, A. Mekkittikul, W. Ellersick, M. Horowitz, "The Tiny Tera: A Packet Switch Core", *IEEE Micro.*, Vol. 17, No. 1, Jan.-Feb. 1997, pp. 26-33.

[4] K.Y. Yun, "A Terabit Multiservice Switch", *IEEE Micro*., Vol. 21, Issue 1, Jan.-Feb. 2001, pp. 58-70.

[5] J. Chao, "Saturn: A Terabit Packet Switch Using Dual Round-Robin," *IEEE Communications Magazine*, Vol. 38, Issue 12, Dec. 2000, pp. 78-84.

[6] *Native-18 Standard Cell Library 0.18u TSMC Process*, Rev. 1.0, Virtual Silicon Technology Inc, Sunnyvale, CA, USA, Sep. 1999.

[7] C. T. Chan, J. S. Hwang, and O. T. –C. Chen, "A 2.5 Gb/s CMOS Optoelectronic Transceiver for Optical Communications", *Proceedings of the 44th IEEE 2001 on Circuit and Systems*, Vol. 1, 2001, pp. 381-384.

[8] T. K. Woodward and A. V. Krishnamoorthy. "1-Gb/s Integrated Optical Detectors and Receivers in Commercial CMOS Technologies", *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 5, No. 2, March/April 1999, pp. 146-156.

[9] *The National Technology Roadmap For Semiconductors*, Semiconductor Industry Association, San Jose, CA 95110, 1997.

[10] R. Ho, K. W. Mai, and M. A. Horowitz, "The Future of Wires", Invited Paper, *Proceedings of the IEEE*, Vol. 89, No. 4, Apr. 2001, pp. 490 –504.

[11] *0.18 Micron CMOS Process Technology*, TSMC, Hsin-Chu, Taiwan 300, R.O.C., March 2002.

[12] R. Manohar and J. A. Tierno, "Asynchronous Parallel Prefix Computation", *IEEE Transactions on Computers*, Vol. 47, No. 11, Nov. 1998, pp. 1244-1252.

[13] J. E. Savage, *Models of Computation: Exploring the Power of Computing*, Addison-Wesley, Reading, MA, USA, 1998.

[14] T. H. Szymanski, "Design Principles for Practical Self-Routing Nonblocking Switching Networks with O(NlogN) Bit-Complexity", *IEEE Transactions on Computers*, Vol. 46, No. 10, Oct. 1997, pp. 1057-1069.

[15] *Power Compiler™ Reference Manual*, Version 2000.11, Synopsys, Inc., Mountain View, CA, USA, Nov. 2000.

[16] *Design Compiler™ Reference Manual Optimization and Timing Analysis*, Version 2000.11, Synopsys, Inc., Mountain View, CA, USA, Nov. 2000.