Power Complexity of Multiplexer-Based Optoelectronic Crossbar Switches

Ted H. Szymanski, *Member, IEEE*, Honglin Wu, and Amir Gourgy, *Student Member, IEEE*

***Abstract*—The integration of thousands of optical input/output (I/O) devices and large electronic crossbar switching elements onto a single optoelectronic integrated circuit (IC) can place stringent power demands on the CMOS substrates. Currently, there is no sufficiently general analytic methodology for power analysis and power reduction of large-scale crossbar switching systems. An analysis of the power complexity of single-chip optoelectronic switches is presented, assuming the classic broadcast-and-select crossbar architecture. The analysis yields the distribution of power dissipation and allows for design optimization. Both unpipelined and pipelined designs are analyzed, and a technique to reduce power dissipation significantly is proposed. The design of a 5.12 Tbit single-chip optoelectronic switch using 0.18-μm CMOS technology is illustrated. The pipelined switch design occupies $<$ 70 mm^2 of CMOS area, and consumes $<$85 W of power, which compares favorably to the power required in electrical crossbar switches of equivalent capacity.**

***Index Terms*—Broadcast, CMOS, crossbar, optical, switch, optoelectronic, power, Terabit, VCSEL, VLSI.**

I. Introducti/on

THE last decade has witnessed the development of optoelectronic integrated circuits (OEICs) with the potential for thousands of vertical cavity surface emitting lasers (VCSELs) and photodetectors (PDs), bonded onto a CMOS substrate with millions of transistors [1], [2]. Crossbar switches are ubiquitous components in computer and communication systems. Single-chip optoelectronic crossbar switches exploiting this new technology thus have the potential to provide exceptionally high multiTerabit bandwidths and low latency communications.

However, the integration of thousands of optical input/output (I/O) devices and very large electronic crossbar switches onto the same OEIC can place stringent power demands on the CMOS substrate. Conventional electronic crossbar switch integrated circuits (ICs) are limited in their aggregate capacity typically a few terabits per second (Tb/s), due to constraints on the density, bandwidth, and power consumption of conventional electronic I/O signaling technologies. In contrast, single-chip OEICs can support potentially 50 000 emitter/receiver pairs [1], and can have I/O bandwidths in the range of potentially 10–50 Tb/s [2]. These exceptional bandwidths may well exceed the logical processing and power dissipation capabilities of the CMOS substrate, potentially limiting the impact of the OEIC technology.

Power is emerging as a dominant design constraint in silicon VLSI [3], [4]. Inspite of the ubiquitous nature of crossbar switches, to date there is no sufficiently general published analytic design methodology for power minimization of crossbar switches. In this paper, an analytic methodology for the power analysis of single-chip electrical or optoelectronic crossbar switches, using any target standard cell library, is presented. The switch components which dissipate considerable power have been identified, and a technique to significantly reduce power is described. The analytic methodology allows a designer to thoroughly explore the design space and select a crossbar switch design which meets the power, VLSI area and timing constraints early, before detailed circuit level design begins.

High-capacity single-chip electrical crossbar switches, with aggregate capacities of hundreds of gigabits per second, are commercially available. Triquint Semiconductor has a 64×33 crosspoint switch chip (TQ8033), with a peak capacity of 50 Gb/s and a power dissipation of 4.9 W [5]. Vitesse Semiconductor has a 144×144 crossbar switch (VSC3040), with a peak capacity of 518 Gb/s and a power dissipation of 16 W [6]. Single-chip switches are scalable to larger sizes, by interconnecting multiple chips in a two-dimensional (2-D) array or in a nonblocking 3-stage CLOS network. Using the Triquint chips, the largest three-stage CLOS network (with 99 chips) will achieve a maximum capacity of 1.65 Tb/s, and will dissipate 495 W. Using the Vittese chips, a three-stage CLOS network (with 80 chips) will achieve a capacity of 10.4 Tb/s and will dissipate 1.28 kW of power [6]. There are several other examples of very high capacity electrical crossbar switches [7]–[12].

With the recent development of VCSEL technology, several Terabit capacity optoelectronic switches/systems have been proposed. Several research teams have demonstrated single-chip optoelectronic switches with several hundred optical I/O bonded onto a CMOS substrate, i.e., see [13] and [14]. An essentially-nonblocking multistage crossbar switch architecture which scales to hundreds of terabits of capacity using OEIC technology was proposed in [15]. A ten-year Canadian project to design a free-space multiterabit optical backplane is described in [16], [17]. A multiterabit local area network (LAN) using optoelectronic crossbars is described in [18], and transistor-level switch designs for optoelectronic networks are considered in [19].

While other power models for crossbar switches have been presented, i.e., [20]–[22], they have not used parameters from

Manuscript received December 31, 2003, revised October 2004. This research was supported by the NSERC Canada Grant OPG-0121601, the L. R. Wilson/Bell Canada Enterprises Chair in Data Communications, and a CITO-Ontario Center of Excellence grant.

T. H. Szymanski, H. Wu, and A. Gourgy are with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada (e-mail: teds@mail.ece.mcmaster.ca).

Digital Object Identifier 10.1109/TVLSI.2005.844285

real CMOS standard cell libraries, and therefore do not provide designers with critical data on their designs implemented in a particular CMOS technology. The analytic methodology developed in this paper is general and useful for the accurate architectural evaluation of electrical and optoelectronic crossbar switches, given a specific standard cell library. Designers can mathematically model their switches and analytically optimize the CMOS area, clock rate, and power dissipation for any target submicron CMOS technology (i.e., 0.18-, 0.13-, 0.09-μm CMOS), before undertaking a detailed CMOS VLSI design, thereby saving potentially hundreds of hours of computer-aided design (CAD) time, as is illustrated next.

Traditionally, IC design teams follow a detailed "design-flow," with the extensive use of CAD tools, to complete a design. Such design-flows have a several major steps, including high-level design specification, register-transfer level (RTL) design, functional simulation, logic synthesis, logic simulation, prelayout timing and power analysis, transistor netlist generation, transistor level simulation, clock tree generation, place-and-route, physical parameter extraction, post-layout timing and power analysis, layout-versus-schematic verification, design rule verification, and tape-out [4], [23], [44]. Design-flows currently rely upon iteration to meet design constraints, i.e., if timing or power constraints are not met after the post-layout analysis, a design team may be sent back to the initial RTL design stage, to perform a high-level redesign. These loops are often reiterated several times, until the design is guaranteed to meet all design constraints. By using the proposed methodology design teams can save significant design time, by mathematically optimizing the design at the early stages of the flow and minimizing iterations through the flow. In addition, the design techniques proposed in this paper will reduce the power consumption of optoelectronic switches.

This paper is organized as follows. Section II describes the single chip OEIC, and summarizes the optical device and CMOS standard cell library parameters. Section III presents the analytic methodology for the VLSI area, delay, and power analysis of the conventional switch. Section III also proposes a design technique to minimize power dissipation significantly. Section IV presents an analysis for a pipelined switch. Section V presents a design example of a single-chip 5.12-Tb/s OEIC switch. Section VI provides concluding remarks.

II. Optoelectronic Switch Architecture

The single-chip OEIC switch shown in Fig. 1 will interconnect N incoming and N outgoing parallel multimode fiber ribbons [18]. Each computer is connected to the crossbar switch through a parallel optical fiber ribbon, with 12 channels each operating at 4 Gb/s, similar to existing commercially available fiber ribbon transceivers [24].

Assume that one fiber in each fiber ribbon is used to transmit a high-quality low-jitter clock, which is recovered with a delay locked loop (DLL). The clock is then replicated and each version is phase shifted to perform near optimal data sampling on each data stream. Circuits to perform these tasks are described in [25], [26]. One fiber in each ribbon is used for parity check and control, and the ten remaining fibers in each ribbon are used for data. To maintain dc balance at each receiver, each serial stream can be encoded with an 8 B/10 B encoder, or a disparity counting circuit can be used with no effective loss of datarate. Therefore, each fiber ribbon supplies 40 Gb/s of data. To achieve a capacity of 5.12 Tb/s, 128 fiber ribbons must enter and exit from the single-chip switch. Referring to Fig. 1, the switch will therefore require a 2-D array of $128 \times 12 = 1\,536$ PDs for data reception, a 2-D array of 1536 VCSELs for data transmission, and will require a CMOS switch with an aggregate capacity of 5.12 Tb/s, which would represent the largest single chip optoelectronic or electronic crossbar switch constructed to date.

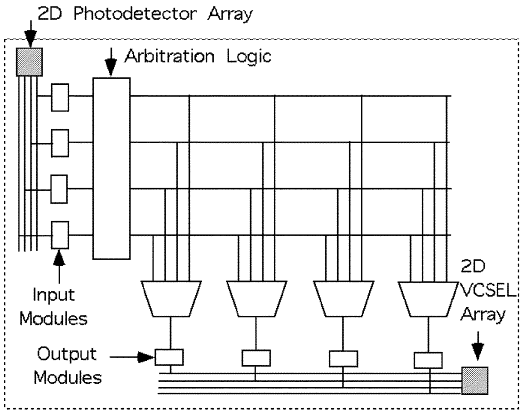


Fig. 1. Single-chip CMOS/VCSEL optoelectronics switch.

The input and output modules in Fig. 1 contain the traditional physical and data link layer functions for managing transmissions on a link, including buffering and error control.

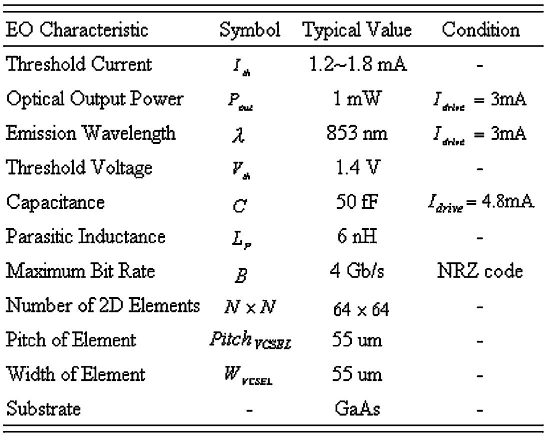
A. Optical I/O

To develop an OEIC, optical devices (VCSELs and PDs) are fabricated on a GaAs wafer specifically designed to be flip-chip bonded to a CMOS substrate, thereby providing the means to connect thousands of optical devices in the GaAs wafer directly to the silicon substrate. The OEIC technology roadmaps proposed in [1], [2] project the development of OEICs with 4–8 K optical devices and with aggregate throughputs of tens of terabits by 2007, and provide the motivation for this paper.

Recently a GaAs IC with 4 096 VCSELs and PDs was fabricated and tested, and the relevant physical parameters are summarized in Tables I and II, [27]. These devices were not designed to be flip-chip bonded onto a silicon substrate, which requires the placement of metal bonding pads on the surface and the use of "bottom emitting" VCSELs. The optical devices in [27] are relatively slow and have a maximum modulation bandwidth of 2 GHz, or a datarate of 4 Gb/s using NRZ signaling, as shown in Table I.

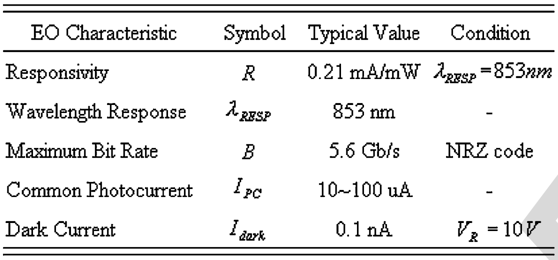
The optical devices in [27] were arranged in two 2-D 64×64 arrays, occupying 12.3 mm^2 each, for a density of 32 K optical devices per square centimeter. The VCSELs each consumed 8.9 mW of electrical power and generated 1.6 mW of optical power, for a conversion efficiency of 18%. The GaAs IC in [27] would offer an aggregate outgoing bandwidth of 16 Tb/s while dissipating 36 W for the VCSELs, if all VCSELs were simultaneously enabled. The receiver circuits for the PDs were

TABLE I
PARAMETERS FOR 2-D VCSELS [27] AND [28]



EO Characteristic	Symbol	Typical Value	Condition
Threshold Current	I_{th}	1.2~1.8 mA	-
Optical Output Power	P_{out}	1 mW	I_{drive} = 3mA
Emission Wavelength	λ	853 nm	I_{drive} = 3mA
Threshold Voltage	V_{th}	1.4 V	-
Capacitance	C	50 fF	I_{drive} = 4.8mA
Parasitic Inductance	L_p	6 nH	-
Maximum Bit Rate	B	4 Gb/s	NRZ code
Number of 2D Elements	$N \times N$	64 × 64	-
Pitch of Element	$Pitch_{VCSEL}$	55 um	-
Width of Element	W_{VCSEL}	55 um	-
Substrate	-	GaAs	-

TABLE II
PARAMETERS FOR 2-D RCPDS [27] AND [29]



EO Characteristic	Symbol	Typical Value	Condition
Responsivity	R	0.21 mA/mW	$\lambda_{RESP}=853nm$
Wavelength Response	λ_{RESP}	853 nm	-
Maximum Bit Rate	B	5.6 Gb/s	NRZ code
Common Photocurrent	I_{PC}	10~100 uA	-
Dark Current	I_{dark}	0.1 nA	V_R = 10V

not fabricated [27], but our own designs (in Section 2.B) indicate a power of 1.75 mW per PD receiver. Therefore, the VCSEL/PDS in [27], coupled with our VCSEL driver and PD receiver circuits in 0.18-μm CMOS reported in Section II-B, together demonstrate a bandwidth-power efficiency of 2.7 mW per Gb/s, consistent with [2] which projected efficiencies of between 0.84 and 3.75 mW per Gb/s. The challenge addressed in this paper is to determine if CMOS electronic switching can match this efficiency.

In Fig. 1, the PDs are arranged in a 2-D 32 × 48 array at one corner of the die. The pitches in the x/y directions are 75 and 43.4 μm, respectively, for reasons discussed next, for a total area of 5 mm^2. The optical PDs have a similar density to those in [27].

The 128 incoming and outgoing parallel fiber ribbons are interfaced to the switch in Fig. 1 using the technique described in [18]. The ends of 128 fiber ribbons can be arranged into a 2-D array with 32 rows of 4 ribbons, with a polished end-surface area of 14.4 × 12.5 mm. The optical signals leaving the surface undergo an image compression by a factor of 6, using a high-quality camera lens, with a resulting image size of ≈ 5 mm^2. The optical signals will align to within a few microns of the centers of the PDs, and can be focused onto the PDs using a 32 × 48 microlens or holographic array. In our laboratory experiments, optical signals can be accurately positioned to within one micron of a target, with as little as a 10-μm pitch between signals, considerably more resolution and spatial density than is required. The inverse process, i.e., an image expansion by a factor of 6, is used to map the VCSEL signals into the outgoing parallel fiber ribbons. The signals can be coupled into the fibers using a 32 × 48 array of polymer or diffractive microlenses. See [16] for a discussion of imaging techniques used to interconnect 512 free-space optical signals between OEICs, in a free-space optical backplane.

B. Optical Transmitter/Receiver Circuits in 0.18-μm CMOS

To evaluate the power consumption of the proposed switch, a simple and robust VCSEL driver circuit in [31], [32] (HP 0.6-μm CMOS), and [33] (HP 0.8-μm CMOS) was scaled down to the TSMC 0.18-μm CMOS technology, implemented using the Cadence Virtuoso™ design tool, and simulated using the Cadence Affirma™ Spectre analog circuit simulator. The driver circuit in Fig. 2(a) is designed to supply each VCSEL with 3 mA of current and 2 V at a 4-Gb/s datarate, resulting in a peak optical output of 1 mW and an average optical power of 0.5 mW. In Fig. 2(b), the average current is 2.5 mA per VCSEL at a 3.3-V supply voltage, and the total power dissipated per VCSEL and driver is ≈ 8.25 mW.

A transimpedance amplifier (TIA) circuit in [34], [35], and [36] (Lucent 0.35-μm CMOS) was also scaled down to TSMC 0.18 μm CMOS technology. These circuits were designed to interface with resonant cavity photodetectors (RCPDs). The circuit is shown in Fig. 3(a), and the recovered logic signal is illustrated in Fig. 3(b). The load was 40 fF, equivalent to six standard loads. The TIA exhibits excellent low power performance at 4 Gb/s using 0.18 μm CMOS technology. The current per receiver varies between 0.14 and 1.4 mA, with an average value of 0.54 mA, with a supply voltage of 1.8 V. The PDs themselves are reversed-biased and dissipated negligible power. Therefore, the total power dissipated per PD and receiver is 1.75 mW.

These designs indicate that each VCSEL/PD pair, along with the VCSEL drivers and PD receivers, will dissipate ≈ 10 mW. A 5.12 Tb/s capacity switch with 1 536 VCSEL/PD pairs will have a power dissipation of 15.4 W for the optical I/O. A clock-and-data recovery circuit with 2 DLLs, to lock and sample a 4 Gb/s stream requires ≈27 mW per serial channel [25]. A single DLL to sample one data stream will require 13.5 mW. Therefore, the CDR for the entire switch will require ≈ 20.7 W. The total optical I/O, including all VCSEL, PD, and CDR circuitry, will require 36 W. The incident optical power from 1536 fibers will also add ≈ 1.5 W, much of which will dissipate as heat. This 36-W figure is reasonable, and compares very well with the power needed for electronic I/O pins in all-electronic switches (to be discussed ahead). It will be shown that the digital switching logic forms the dominant power constraint in the optoelectronic switch.

C. Optical Power Budget and Bit Error Rate

The VCSELs generate 0.5 mW or − 3 dBm of optical power. The VCSEL-to-fiber coupling efficiency is typically high, better than − 0.45 dB [37]. Multimode fiber has an attenuation of < 3 dB/km [37], and 50 m of fiber ribbon will have an attenuation of 0.15 dB. Our proposed imaging system uses two optical lenses and one microlens array, each contributing < 1-dB loss (similar to data reported in [16]). Allowing a further 3-dB loss for bending, aging, and temperature effects, the optical signal power at the receiver will be −9.6 dBm, or

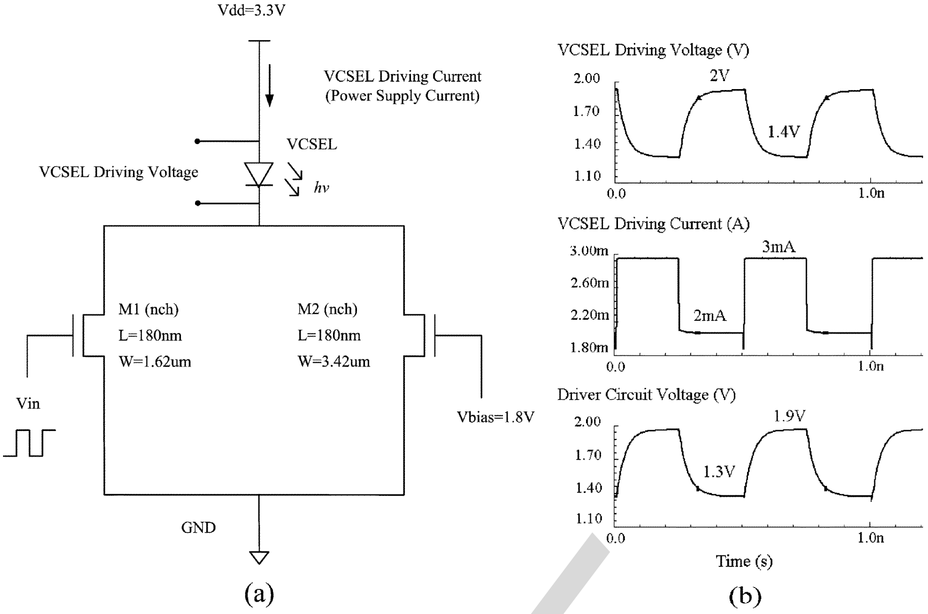


Fig. 2. (a) Proposed 0.18-μm VCSEL driver circuit from Cadence Virtuoso tool. (b) Simulation of the proposed VCSEL driver circuit at 4 Gb/s from Cadence Affirma Spectre simulator.

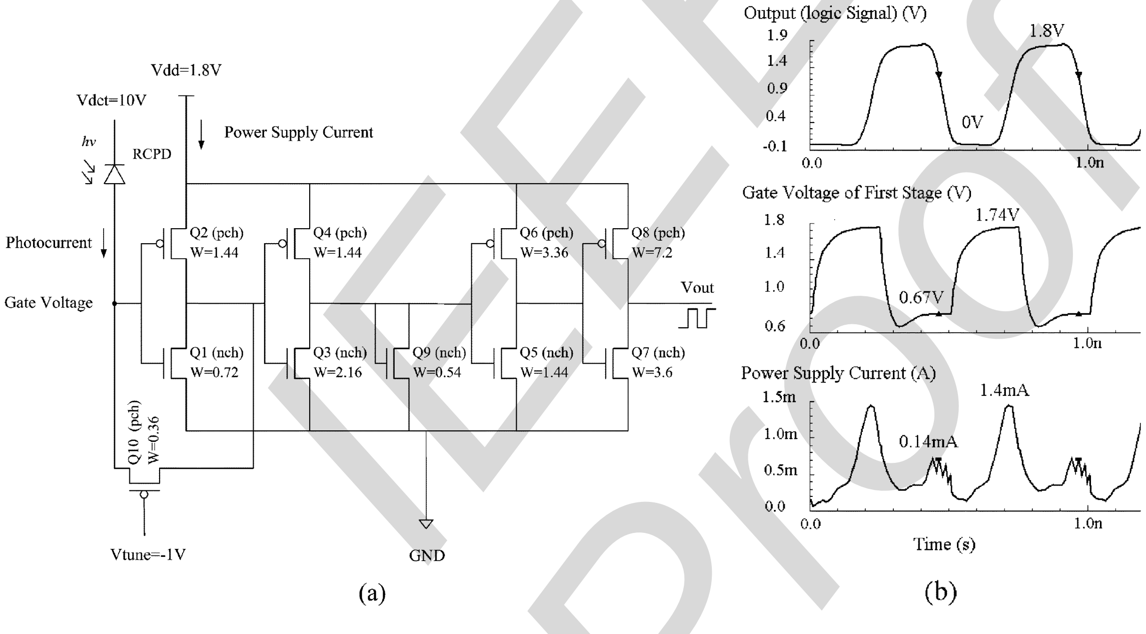


Fig. 3. (a) Proposed 0.18-μm RCPD receiver circuit from Cadence Virtuoso tool; W is in microns. (b) Simulation of the proposed RCPD receiver circuit at 4 Gb/s from Cadence Affirma Spectre simulator.

110 μW. The sensitivity of a PD, defined as the received signal power required to ensure a bit-error rate (BER) of 10^{-12}, is reported to be -16 dBm for similar GaAs PDs operating at 2 Gb/s [24]. Assuming the same sensitivity for our PDs, the proposed system operates with a power margin of $+6.4$ dBm (at a BER of 10^{-12}).

While a BER of 10^{-12} is often considered a reasonable target in traditional computer systems, it will correspond to five bit errors per second in the proposed system, given the 5.12 Tb/s bandwidth. It has been argued that future multiTerabit systems will require much lower BERs, or strong error detecting/correcting codes [38]–[40] with high throughput and reasonable hardware complexity, to limit the effects of bit errors.

Following the analysis in [39], the BER in the proposed system can be determined by assuming a thermal-noise limited system with white Gaussian noise, and the results are summarized. A receiver sensitivity of -16 dBm at 2 Gb/s, implies a BER of 10^{-12}, which further implies a signal-to-noise ratio (SNR) of 23 dB. It can be verified that the noise equivalent power (NEP) or $\mathrm{NEP} = 0.003$ nW/Hz$^{1/2}$, the noise power$=$ 190 nW at 4 Gb/s, the signal power $=$ 110 μW, the $\mathrm{SNR} = 27.6$ dB, and the $\mathrm{BER} = 2.3 * 10^{-33}$. This BER will imply 10^{-20} bit errors per second, or a mean time to error (MTTE) of 10^{20} seconds. The same methodology can be followed to determine the BER given any other reasonable

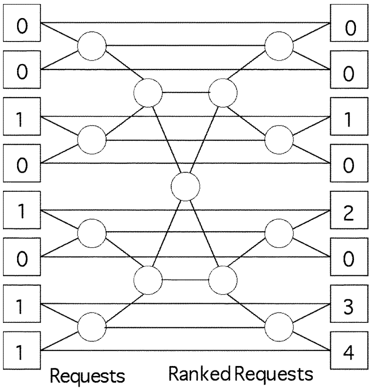


Fig. 4. Logic diagram of input arbitration circuit.

TABLE III
PARAMETERS FOR AMIS 0.35-μm CMOS

Gate (0.35um CMOS)	Area (std area)	Delay (nsec) 1 std ld	Delay Slope (nsec/gate)	Input Cap. (std ld)	Intrinsic Cap. (std ld)
Inverter (INV1)	0.8	0.080	0.029	1	0.4
2-1 MUX (MX21)	2.0	0.230	0.039	1	4.0
4-1 MUX (MX41)	4.2	0.500	0.065	1	10.9
8-1 MUX (MX81)	9.0	0.530	0.059	1	25.9
3-State Buffer (ITB1)	1.2	0.140	0.050	1	1.8
D Flip Flop (DF111)	5.5	0.350	0.049	1	14.9

TABLE IV
PARAMETERS FOR SCALED 0.18-μm CMOS

Gate (0.18um CMOS)	Area (std area)	Delay (nsec) 1 std ld	Delay Slope (nsec/gate)	Input Cap. (std ld)	Intrinsic Cap. (std ld)
Inverter (INV1)	0.8	0.038	0.014	1	0.4
2-1 MUX (MX21)	2.0	0.110	0.019	1	4.0
4-1 MUX (MX41)	4.2	0.240	0.031	1	10.9
8-1 MUX (MX81)	9.0	0.254	0.029	1	25.9
3-State Buffer (ITB1)	1.2	0.672	0.024	1	1.8
D Flip Flop (DF111)	5.5	0.168	0.024	1	14.9

set of device and operating parameters. Given the large power margin, the end results should be similar: a very low BER.

D. Arbitration Logic

In a LAN environment, the switch must employ a fast on-chip arbitration algorithm. The study of arbitration or scheduling algorithms is an active research area, i.e., see [41], [42]. These algorithms consume VLSI area and power, and will affect the performance of the switch. In this paper we focus on the crossbar switch, and assume a simple parallel prefix ranker circuit described in [15] is used to determine the switch control settings. A diagram of the arbitration logic is shown in Fig. 4. Our analysis indicates that the arbitration logic uses considerably less VLSI area and power than the crossbar switch. In the rest of the paper, we will focus on the 2-D crossbar switch array, where the majority of power is consumed.

E. CMOS 0.18-μm Technology

The switch will be designed using a 0.18-μm CMOS technology. A popular standard cell library is the VST Technology library for 0.18-μm CMOS, supported by the TSMC foundry service [43]. However, access to this library requires a non-disclosure agreement. In order to present realistic design data for 0.18-μm CMOS technology, a scaled version of a publically available 3.3-V 0.35-μm standard cell library from American Semiconductor [44] is used.

Table III illustrates several parameters for this 0.35-μm library. The area of a NAND gate defines the "standard gate area", and its input capacitance defines the "standard load." The area of a standard gate was not reported, and is assumed to be 40 μm^2, yielding a gate density of 25 kgates/mm^2, to match that of other libraries. The standard load in the 0.35-μm library is reported to be 27.7 fF.

Using the MOSIS scalable CMOS design rules [45], the 0.35-μm library has been scaled down by parameter $s = 0.18/0.35 = 0.51$, to yield data for a 0.18-μm library shown in Table IV. All cell dimensions are scaled down by a factor of s, and the cell area is therefore scaled down by the factor s^2. The 0.18-μm library has a standard gate area of 10 μm^2, for a gate density of 100 kgates/mm^2, and a standard load of 7 fF. (This figure includes 8 μm^2 for the cell and an extra 25% for wiring overhead.) These figures are essentially identical to publically reported figures for the VST 0.18-μm library, indicating an excellent agreement. The areas of more complex gates in Tables III and IV are reported as "equivalent gates," as multiples of the standard gate area.

The time constant of a standard logic gate is determined by the resistance and capacitance of a minimum size transistor, $\tau = R * C$. The pull-up and pull-down resistances of a gate are both scaled down by the factor s, due to the linear decrease in the transistor gate length. The capacitive load includes the wire capacitance and the gate capacitance, and to a first order approximation in submicron CMOS, the capacitive load is also scaled down linearly. Therefore, given a constant supply voltage, the gate delay is scaled down by the factor s^2. In our scaled library, the supply voltage is also scaled down from 3.3 to 1.8 V, increasing the time constant linearly by $\approx$ 1.83. In Table III, the delays of the 0.35-μm cells were scaled down by a factor of 0.48, to determine the delays for the 0.18-μm cells in Table IV. More detailed second-order analyzes can be used to yield slightly different timing data for the scaled library [46].

All standard cells come in several drive strengths (ds), typically denoted in libraries as $x = 1, 2, 4$. When the drive strength of a cell increases, the capacitive load which the cell can drive in a given time typically increases linearly, its area also increases since the output transistors must be larger, and its intrinsic (internal) capacitance increases. All capacitances in the Table IV are expressed in terms of the standard load C_{std} of 7 fF. The delay of a cell can be modeled by its "intrinsic" or inherent switching delay, and the linear delay associated with the capacitive load being driven [23], in a linear delay model as follows:

$$\text{delay(std cell)} = \text{delay}_{\text{int}} + \Delta_d \cdot C_{\text{load}}/(C_{\text{std}} \cdot \text{ds}). \quad (1)$$

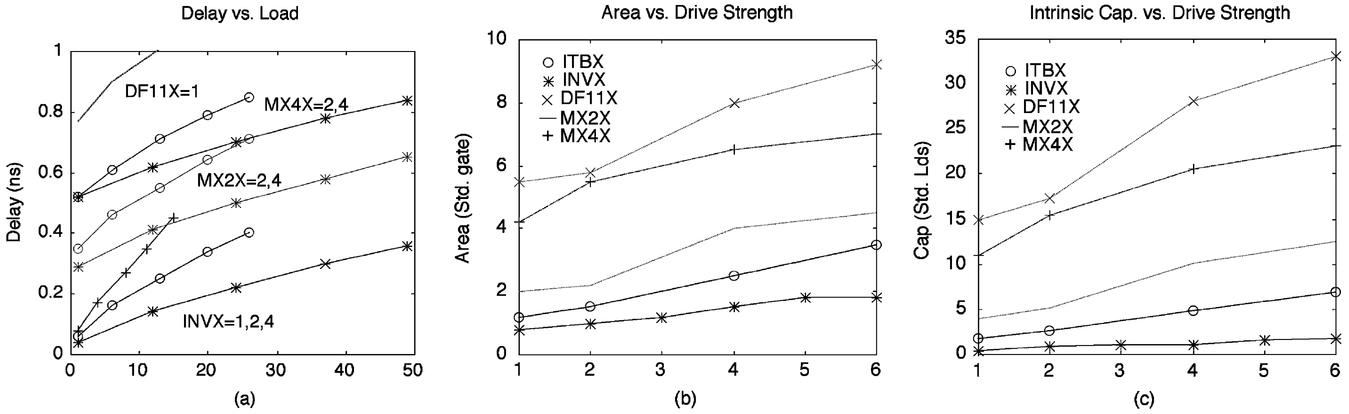


Fig. 5. Parameters of AMIS standard cell library. (a) Delays versus load. (b) Area versus drive strength. (c) Intrinsic capacitance versus drive strength.

The delay slope Δ_d (nanosec per standard load) in (1) was determined from the 0.35-μm library by linear interpolation, as recommended by American Semiconductor [44]. The delay slope in Table IV was determined by scaling the results in Table III. Fig. 5(a) plots the delay versus driven load, for three families of standard cells (the D flip-flip cell DFF11X, the degree 2 and 4 multiplexer cells $\mathrm{MX2X} = 2,4$ and $\mathrm{MX4X} = 2,4$, and the basic inverter cell $\mathrm{INVX} = 1,2,4$), from the AMIS 3.3-V 0.35-μm library. Fig. 5(a) illustrates that the delay can be accurately modeled as linear.

However, there is no prior published data on how a standard cell area or intrinsic capacitance varies according to its drive strength. Fig. 5(b) plots the cell area versus the drive strength, for the same standard cells. Fig. 5(b) indicates that the cell area as a function of the drive strength can be reasonably accurately represented by a linear model

$$\text{area (std cell)} = \text{area}_{\text{int}} + \Delta_a \cdot (\text{ds} - 1) \qquad (2)$$

where the area slope Δ_a represents the area increase per drive strength increment. Fig. 5(c) plots the intrinsic (internal) capacitance of a cell as a function of the drive strength, and it too can be reasonably accurately modeled as linear. We point out that a system designer using our analytic methodology should use the real standard cell data provided by their standard cell library vendor, and use linear approximations for projections to larger drive strengths.

The power dissipation of a standard cell is given by

$$P(\text{std cell}) = \frac{1}{2} f_{\text{clk}} \cdot f_{\text{duty}} \cdot (C_{\text{int}} + C_{\text{load}}) \cdot (V_{\text{dd}})^2 \qquad (3)$$

where C_{int} reflects the intrinsic capacitance of the cell, and C_{load} is the capacitive load being driven by the cell. With 0.18-μm CMOS, the power dissipated per standard gate, with a 7 pF intrinsic capacitance, driving one standard load with a 0.5 toggle rate, is 11.34 nW/MHz, which is very close to the values reported in our 0.18-μm standard cell library.

F. Electrical I/O Technologies

High-speed electrical signalling technologies include positive emitter coupled logic (PECL), common mode logic (CML), and low voltage differential signalling (LVDS). The typical power dissipation of an LVDS transmitter/receiver pair is ≈ 70 mW per Gb/s [47], [48]. This figure does not include clock and data recovery. Therefore, a single-chip electrical crossbar switch with 128 input and output ports, with 12 serial channels each clocked at 4 Gb/s, will dissipate $\approx$5 120 Gb/s $*$ 70 mW/Gb/s $=$ 360 W. In contrast, the optical I/O described earlier dissipates $\approx$ 15.4 W.

III. Analysis of the Conventional Crossbar

Referring to Fig. 1, the basic crossbar switch architecture consists of N input modules, each broadcasting one virtual circuit (VC) over a broadcast bus, where each bus contains w bits. We can generalize the design so that each input/output port supports multiple VCs simultaneously, sharing the w bits, and the analysis does not change substantially.

There are Nw driver cells required to drive the Nw horizontal wires. There are N multiplexer trees, one for each output module. Each multiplexer tree spans all N horizontal broadcast busses, and will concentrate and forward one VC to its output port. Each intersection point in Fig. 1 represents a logical connection between the horizontal row bus and vertical column bus each w-bits wide, i.e., there is an array of w metal via connections. Each N-to-1 multiplexer tree can be constructed with $(N-1)/(m-1)$ smaller multiplexer standard cells with m inputs each, arranged in a tree topology, as shown in Fig. 6. Let all busses be w bits wide, let all multiplexer standard cells have the same drive strength, and let $A(\text{mux})$ and $A(\text{buf})$ denote the area of an m-to-1 multiplexer cell and a buffer cell with a given drive strength, expressed in microns, which can be determined as described in Section II. (In our analysis, we use the data for the MX4 and INVX cells from the AMIS library). Assume there are sufficient layers of metal so that the design is not constrained by wiring. The area required for entire switch is therefore

$$A(\text{switch}) = Nw \cdot A(\text{buf}) + Nw \cdot \frac{(N-1)}{(m-1)} \cdot A(\text{mux}). \quad (4)$$

In an optimized layout generated by CAD tools, the drive strength and the size of the multiplexer standard cells used in the switch will vary according to the capacitive load being driven. Equation (4) overestimates the area and power of the switch, by setting all standard cells to use the same drive strength. An optimization tool such as the synopsis logic synthesis engine will

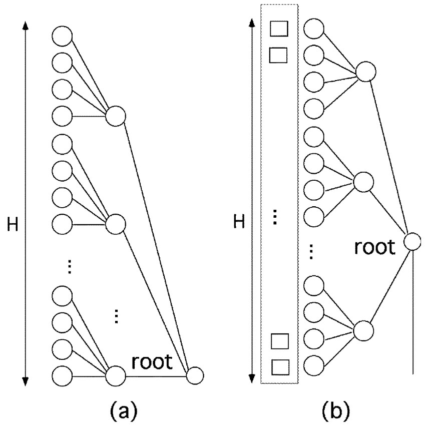


Fig. 6. Multiplexer layouts to reduce wire length. (a) Worst-case. (b) Best-case.

adjust the strength of the standard cells according to an optimization criteria specified by the designer. To minimize VLSI area, small standard cells will be selected, resulting in a larger delay. To minimize delay, large standard cells will be selected, resulting in a larger VLSI area. It is straight-forward to adopt (4) to consider nonhomogeneous drive strengths, by using a weighted average to determine the cell areas, or by considering each level separately. For example, a tree-based multiplexer as shown in Fig. 6 may increase the drive strength at each level of the tree, where the lowest (N/m) standard cells use drive strength 1. In this case, the weighted average area of a multiplexer standard cell is given by

$$A'(\mathrm{mux}) = \left(\frac{m-1}{N-1}\right) \sum_{k=1}^{\log_m N} \left(\frac{N}{m^k}\right) \cdot A(\mathrm{mux}, \mathrm{ds}=k). \quad (5)$$

The total number of control signals needed to configure the switch is $N \log_2 N$. The area of the D flip-flop (DFF) memory elements needed to store the configuration bits and the DFFs associated with the horizontal busses must be added to the area to yield

$$\begin{aligned} A(\mathrm{switch}) &= Nw(A(\mathrm{DFF}) + A(\mathrm{buff})) \\ &+ Nw \cdot \frac{(N-1)}{(m-1)} \cdot A(\mathrm{mux}) + N \cdot \log_2(N) \cdot A(\mathrm{DFF}). \quad (6) \end{aligned}$$

Assuming a unity aspect ratio, the square root of the area yields the horizontal and vertical dimensions H and V (in microns), respectively

$$H = V = \Omega(N\sqrt{w/m}) \approx N\sqrt{(w/m)A(\mathrm{mux})}. \quad (7)$$

Equation (7) is valid if the switch has sufficient perimeter to allow all $Nw + N\log_2 N$ wires to enter, and Nw wires to exit. Assume there are M layers of metal available for routing wires, in each of the horizontal and vertical directions, and that all wires are long distance wires with 6λ width and 10λ pitch, a conservative estimate. In a worst-case scenario, Nw wires will enter on one side of the rectangular layout, leading to the following *horizontal wire* routing constraint

$$Nw10\lambda/M \leq \sqrt{A(\mathrm{switch})} = H = V = \Omega(N\sqrt{w/m}). \quad (8)$$

Equation (8) is conservative, since layout tools will generally route wires out of all four sides of the layout. If (8) is met, then the layout has sufficient perimeter to allow all wires to enter and exit in a worst-case scenario. Otherwise, the VLSI area of the switch layout must be increased by adding unused silicon. VLSI layouts are typically characterized as "corelimited" or "I/O-limited" [23]: In a corelimited design, the layout area is determined by the aggregate area of all the internal standard cells, and the wiring is not a factor. In an I/O-limited design, the layout area is determined by the I/O cells or I/O wiring, and there is considerable "white-space" or unused silicon added, to increase the area for routing. Our analysis for 0.18-μm CMOS indicates that the constraint (8) is easily met. In a 0.18-μm process, the worst-case pitch of a wire is 0.9 μm and one millimeter of perimeter is sufficient for 1111 wires per layer of metal. Our crossbar switch may require 5120 wires (at 1 Gb/s each), which can be met with $\approx$ 2 mm per side, with two layers of metal for routing in the horizontal direction. Equation (6) will require considerably more area, generating considerably more area than needed to route the wires.

The power consumed by all the multiplexer standard cells is given by

$$P(\mathrm{muxes}) = Nw \cdot \frac{(N-1)}{(m-1)} \cdot \frac{1}{2} f_{\mathrm{clk}} f_{\mathrm{duty}} \cdot C(\mathrm{mux}) \cdot (V_{\mathrm{dd}})^2 \quad (9)$$

where $C(\mathrm{mux})$ is the total capacitance switched in a multiplexer standard cell. In our case, we set $C(\mathrm{mux}) = m * \mathrm{C}_{\mathrm{mx-in}} + \mathrm{C}_{\mathrm{mx-int}}$, since each input port in each multiplexer is active. In the AMIS library. the multiplexer input-port capacitance $\mathrm{C}_{\mathrm{mx-in}}$ is fixed at one standard load, independent of the drive strength of the cell.

Each horizontal broadcast bus drives the input capacitance of N input ports, one to each of N distinct multiplexer-trees, and has a wire length of H μm. The total capacitance of a horizontal wire of length H μm, including the capacitance of the multiplexer input ports and the wire capacitance is, therefore

$$C(\mathrm{H\ wire}) = N \cdot C_{\mathrm{mx-in}} + H \cdot C_{\mu\mathrm{m}} \quad (10)$$

where $C_{\mu\mathrm{m}}$ is the capacitance per micron length of a long distance metal wire $(\mathrm{width} = 6\lambda, \mathrm{pitch} = 10\lambda)$. In the MOSIS 0.18-$\mu$m technology, the wire capacitance is $\approx$ 0.184 fF/μm.

Referring to (10), the power consumed by N horizontal broadcast busses, each with w bits, is therefore $(f_d = f_{\mathrm{duty}})$

$$\begin{aligned} &P(\mathrm{H\ busses}) \\ &\quad = (Nw) \cdot \frac{1}{2} f_{\mathrm{clk}} f_d \cdot [N \cdot C_{\mathrm{mx-in}} + H \cdot C_{\mu\mathrm{m}}] \cdot (V_{\mathrm{dd}})^2. \end{aligned} \quad (11)$$

Equation (11) is useful to determine the power consumed by the broadcast busses, including the power on the input ports of the multiplexer cell on the busses, which is also counted in (9). The power consumed by the horizontal bus wires alone is given by

$$P(\mathrm{H\ wires}) = (Nw) \cdot \frac{1}{2} f_{\mathrm{clk}} f_d \cdot [H \cdot C_{\mu\mathrm{m}}] \cdot (V_{\mathrm{dd}})^2. \quad (12)$$

As shown in Fig. 6, each N-to-1 multiplexer tree must intersect all N horizontal busses, and must span a vertical distance of $V = H$ μm. Each multiplexer tree can be arranged in VLSI so that the root node is placed near the bottom (or top), as shown in Fig. 6(a). However, this layout increases the aggregate wire length, and is a worst-case. Alternatively, each multiplexer tree can be arranged in VLSI to minimize the cumulative wire length, which is the best-case, as shown in Fig. 6(b). CAD tools will perform the placement and routing of multiple multiplexer trees, and will generally optimize the layouts for maximum performance. We will assume an average case layout for every tree, where the wire lengths for both layouts are averaged. The total capacitance of wires in each multiplexer tree can be determined by summing up the wire lengths in each level of the tree. For the best case, the total length of wires in a multiplexer tree is given by

$$\begin{aligned}\rho_{\text{best}-\text{case}} &= \sum_{k=1}^{\log_m N} \left(\frac{H}{N} \cdot m^{k-1}\right) \left(2 \sum_{j=1}^{m/2} \left(j - \frac{1}{2}\right)\right) w \\ &= \left(\frac{H}{N}\right) \left(\frac{N-1}{m-1}\right) \left(\frac{m^2}{4}\right) w \approx \frac{wm^2H}{(m-1)4}. \qquad (13)\end{aligned}$$

The first summation reflects the height of a subtree, as one moves up the tree toward the root. The second summation reflects the lengths of the m wires entering a multiplexer standard cell, expressed in multiples of the height of the subtree, assuming the destination standard cell is placed at the median height. For the worst-case tree, the total length of wires in a multiplexer tree is given by

$$\rho_{\text{wc}} = \sum_{k=1}^{\log_m N} \left(\frac{H}{N} \cdot m^{k-1}\right) \left(\sum_{j=1}^{m} \left(j - \frac{1}{2}\right)\right) w \approx \frac{wm^2H}{(m-1)2}. \qquad (14)$$

Therefore, the total length of wires in the average-case tree is

$$\rho_{\text{ave}-\text{case}} \approx \frac{3wm^2H}{8(m-1)} \approx \frac{3wm^2H}{8}. \qquad (15)$$

Assuming all wires are long distance wires with a 6λ width and 10λ pitch, a worst-case assumption, then the following vertical wire routing constraint must also be recognized: the total area used to route vertical wires must be realizable on M layers of metal given the area of the switch, i.e.,

$$N \frac{3wmH}{8} 10\lambda / M \leq A(\text{switch}) = \Omega(N^2 w/m). \qquad (16)$$

Our analysis indicates that this *vertical wiring* routing constraint is met for the 0.18-μm CMOS process with six or more layers of metal, for switches of degree 2. However for degree 4 or 8 multiplexers, the area consumed by routing vertical wires increases, and this constraint may not be met, in which case the area of the switch must be increased until (16) holds.

The power consumed by all vertical wires in all N multiplexers is, therefore

$$P(\text{V wires}) = N \cdot \frac{1}{2} f_{\text{clk}} f_d \cdot \left[\frac{3wmHC_{\mu\text{m}}}{8}\right] \cdot (V_{\text{dd}})^2. \qquad (17)$$

Combining (9) and (12), and (17), the total power consumed by the 2-D switch is

$$\begin{aligned}P = {}& (Nw) \cdot \frac{1}{2} f_{\text{clk}} f_d \frac{(N-1)}{(m-1)} \cdot C(\text{mux}) \cdot (V_{\text{dd}})^2 \\ &+ (Nw) \cdot \frac{1}{2} f_{\text{clk}} f_d \cdot [H \cdot C_{\mu\text{m}}] \cdot (V_{\text{dd}})^2 \\ &+ (Nw) \cdot \frac{1}{2} f_{\text{clk}} f_d \cdot \left[\frac{3m^2HC_{\mu\text{m}}}{8(m-1)}\right] \cdot (V_{\text{dd}})^2. \qquad (18)\end{aligned}$$

Equation (18) has an intuitive interpretation. The first term represents the power required by switching the total capacitance of the multiplexer standard cells. The second term represents the power required by switching the capacitive load of the horizontal broadcast wires. The third term represents the power required by switching the capacitive load of the vertical wires in the multiplexer trees.

The maximum clock frequency of a VLSI circuit is determined by the critical path, defined as the worst-case path from an input port to an output port (D flip-flops). In this design, the critical path is the time required to drive a horizontal bus, plus the time to propagate through the vertical multiplexer tree. Referring to Table IV, the delay formula in nanoseconds for a single horizontal wire driven by the inverter standard cell (INVX) is

$$\text{D(H bus)} = 0.038 + 0.014 \frac{\text{(capacitive load)}}{(\text{C}_{\text{std}} \cdot \text{ds})}. \qquad (19)$$

The capacitive load of a horizontal wire from (10) is used in (19) to yield

$$\text{D(H bus)} = 0.038 + 0.014 \frac{(N \cdot C_{\text{mx}-\text{in}} + H \cdot C_{\mu\text{m}})}{C_{\text{std}} \cdot \text{ds}}. \qquad (20)$$

The delay through a multiplexer tree is computed in a similar manner. There are $\log_m N$ levels in the multiplexer tree, and the length of the wires being driven increases as one moves up the tree toward the root. The delay through a multiplexer tree is accurately approximated by summing the intrinsic delay of all $\log_m N$ stages and the delay of any one cell driving the cumulative capacitive load of one vertical wire of length H, and for $m = 4$ is given by

$$\text{D(mux}-\text{tree)} \approx (0.240) \log_m N + \frac{(0.031)HC_{\mu\text{m}}}{C_{\text{std}} \cdot \text{ds}}. \qquad (21)$$

In a very accurate model, the delay of the source DFF and the capacitive load of the destination DFF that define the critical path should be considered in (21). The clock period in nanoseconds is given by the sum of (20) and (21), to yield

$$\begin{aligned}f_{\text{clk}} \approx {}& \left(0.038 + \frac{(0.014)(N \cdot C_{\text{mx}-\text{in}})}{(\text{ds}) \cdot \text{C}_{\text{std}}}\right. \\ &\left. + 0.240 \log_m N + \frac{(0.014 + 0.031)H \cdot C_{\mu\text{m}}}{(\text{ds}) \cdot \text{C}_{\text{std}}}\right)^{-1}. \qquad (22)\end{aligned}$$

For sufficiently large N, the clock rate is given by

$$f_{\text{clk}} \approx \frac{(\text{ds}) \cdot \text{C}_{\text{std}}}{(0.014)(N \cdot C_{\text{mx}-\text{in}} + 2.2H \cdot C_{\mu\text{m}})}. \qquad (23)$$

The previous equations are based on the assumption where all cells have equal drive strength. The power dissipated in a wire is independent of the drive strength of the cell driving the wire, hence this assumption does not affect the power dissipated in the wires.

The total power can be determined by using the clock frequency from (23), in the power expression in (18), rewritten in standard form as

$$\begin{aligned} P(\text{switch}) &= \frac{1}{2}(Nw) f_d f_{\text{clk}} (V_{\text{dd}})^2 \\ &\quad \cdot \left(\frac{(N-1)}{(m-1)} \cdot (mC_{\text{mx-in}} + C_{\text{mx-int}}) \right. \\ &\quad \left. + [H \cdot C_{\mu\text{m}}] + \left[\frac{3m^2}{8(m-1)} \cdot HC_{\mu\text{m}} \right] \right). \end{aligned} \tag{24}$$

The energy per bit switched can be determined by dividing the total power in (24) by the maximum throughput per second (Nwf_{clk}), yielding

$$\begin{aligned} E_{\text{bit}} &= \frac{P(\text{switch})}{f_{\text{clk}} \cdot Nw} \\ &= \frac{1}{2} f_d (V_{\text{dd}})^2 \left(\frac{(N-1)}{(m-1)} \cdot (mC_{\text{mx-in}} + C_{\text{mx-int}}) \right. \\ &\quad \left. + H \cdot C_{\mu\text{m}} + \left[\frac{3m^2}{8(m-1)} \cdot HC_{\mu\text{m}} \right] \right). \end{aligned} \tag{25}$$

Equations (23)–(25) offer a conceptually simple interpretation. The energy per bit switched is equal to several terms: the energy dissipated in switching the capacitive load of all multiplexer standard cells (input port and intrinsic), plus the energy dissipated switching the capacitive load on a horizontal bus wire, plus the energy dissipated switching the capacitive load on the vertical wires traversed in a large multiplexer tree. Recognizing from (7) that $H = V$ is $\Omega(N\sqrt{w/m})$, the energy per bit switched grows according to the following:

$$\begin{aligned} E_{\text{bit}} &= \frac{f_d (V_{\text{dd}})^2}{2} \left(\Theta\left(\frac{N}{m}\right) + \Theta\left(N\sqrt{\frac{w}{m}}\right) \right. \\ &\quad \left. + \Theta\left(mN\sqrt{\frac{w}{m}}\right) \right) \\ &\approx \Theta(N\sqrt{mw}). \end{aligned} \tag{26}$$

According to (26), binary multiplexers yield the slowest asymptotic growth in the energy expended per bit switched, however degree-4 cells offer lower energies for the sizes considered here. Interestingly, while the drive strength strongly affects the clock rate and the aggregate capacity of the switch, the energy expended per bit switched is relatively independent of the drive strength of the standard cells, or the clock rate of the switch.

A. Memory and Clock Tree

A more detailed analysis can consider the memory and clock-tree. The memory generally requires negligible power compared to dynamically switching logic, whereas the clock tree can require significant power. The input modules and output models will generally require memory to store the incoming/outgoing data. A designer generally will have parameters on the memory standard cells from their library vendor. In absence of such data, the area of a memory cell can be estimated. A dynamic memory cell is made with a single capacitive load plus two transistors, and can be approximately modeled as equivalent to one inverter of strength 1. Assuming each input and output port pair has 16 kbytes of memory, the area occupied by memory can be approximated by

$$A(\text{memory}) = N(16\,\text{KB})\text{Area}(\text{inv}) \tag{27}$$

For $N = 128$, the memory area will be upper bounded by $\approx 13 \times 13$ mm^2. The memory changes state relatively infrequently. Each I/O port processes data at the rate of 40 Gb/s, so that the power used by all memory in one port can be modeled as equivalent to charging/discharging one basic inverter at the aggregate data rate, given by

$$P(\text{memory}) = \frac{1}{2}(40\ \text{GHz}) f_d C(\text{inv}) \cdot (V_{\text{dd}})^2. \tag{28}$$

Equations (27) and (28) will overestimate the area and power of memory, since static and dynamic memory cells are highly optimized for low area and low power dissipation. They do yield a useful upper bound. Each I/O port will dissipate 0.31 mW, and all 128 input and output ports will dissipate 40 mW. A designer using our analytic methodology should use the real parameters provided by their standard cell library for final results.

Finally, the clock trees can by considered. Clocks are typically distributed in an H-tree layout [23]. Assuming the loads (DFF clock input ports) are evenly distributed, at the top level the H layout distributes the clock to the center of four quadrants over equidistant paths, and requires a wire of length of $3H/2$ μm. Each quadrant is then handled recursively. Given that there are L levels of recursion, it is easily verified that the capacitance of the cumulative wire length is

$$C = \sum_{j=1}^{L} \frac{3(H/2^{j-1})}{2} 4^{j-1} C_{\mu\text{m}} = \frac{3H(2^L - 1)}{2} C_{\mu\text{m}}. \tag{29}$$

The area covered by a leaf in the H-tree is given by (30), and this area should be small enough to keep the clock skew sufficiently small. Clock wires are routed from the center of the leaf to every clocked element within the leaf over a local wires, and these final wires are not generally equidistant and will introduce clock skew

$$\text{Area}(\text{leaf}) = \left(\frac{H}{2^L}\right)^2. \tag{30}$$

To drive the load efficiently requires a clock tree of successively larger inverters, where the drive strength of each inverter is a ratio r times as large as its predecessor [23]. An optimal tree has a ratio $r = e = 2.7138$, and in practice a ratio $r = 4$ can be used. Let there be R DFFs in the design. The load being driven by the clock tree is equal to the clock tree wire capacitance, plus the capacitance of all DFF clock input ports, which according to Table IV equals one standard load. The depth D of inverters in the clock tree is given by

$$D = \log_4 \left(\left(\frac{3HL}{2} C_{\mu\text{m}} + R \cdot C_{\text{std}} \right) \Big/ C_{\text{std}} \right). \tag{31}$$

The tree is not necessarily fully populated at the bottom level, so we let D assume a fractional value. The number of inverters of strength 4 in the clock tree is given by $(4^D - 1)/(4 - 1)$. The power dissipated in the clock tree, including the input port and

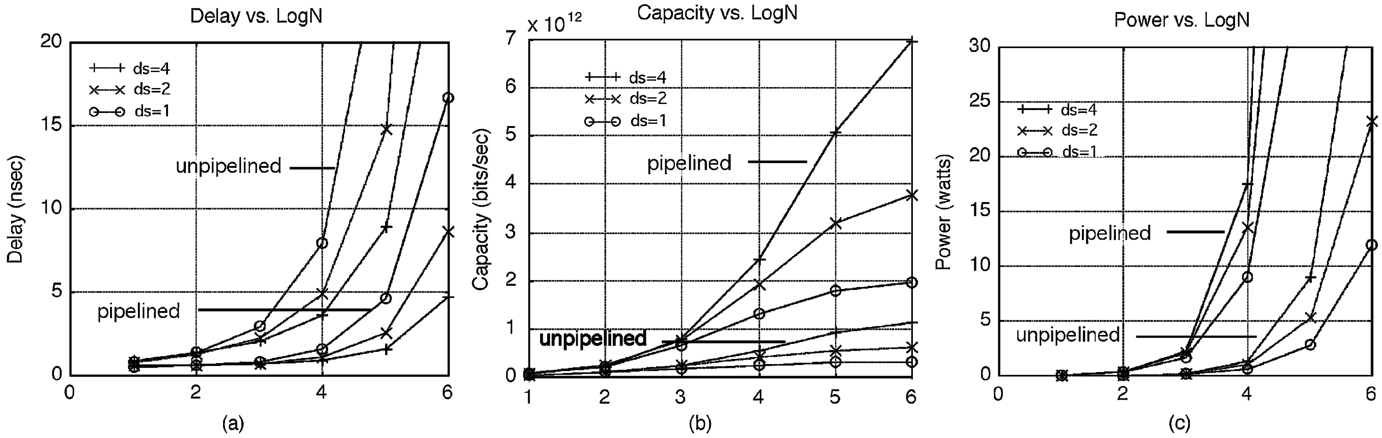


Fig. 7. Crossbar switch. (a) Delay versus $\mathrm{Log}_4\mathrm{N}$. (b) Capacity versus $\mathrm{Log}_4\mathrm{N}$. (c) Total power dissipation versus $\mathrm{Log}_4\mathrm{N}$.

intrinsic capacitance of the inverters, the cumulative wire length, and the capacitance of the clock input ports on all R DFF cells, is given by

$$P = \frac{f_{\mathrm{clk}}(V_{\mathrm{dd}})^2}{2C_{\mathrm{std}}} \cdot \left[\frac{(4^D-1)}{(4-1)}C_{\mathrm{INV4}} + \frac{3H(2^L-1)}{2}C_{\mu\mathrm{m}} + R\cdot C_{\mathrm{std}}\right]. \qquad (32)$$

B. Design Optimizations

1) Strong Inverters: We have observed that the performance is slightly better when all DFF and multiplexer cells have unity strength, and inverters with drive strengths$=1, 2,$ and 4 are used to provide drive strength. The analysis can be easily adapted for design variation.

2) Power Reduction: A significant amount of power is dissipated unnecessarily in the conventional multiplexer-based crossbar, as follows. Referring to Fig. 6, the data input ports to each multiplexer at each level of the tree are actively switched, even though each multiplexer tree eventually forward only one VC to the output port. This phenomena is reflected in (9) and (24), where the input port and intrinsic capacitances of every multiplexer standard cell are repeatedly switched. A plot of the power distribution in an unpipelined switch indicates that the multiplexer trees consume a significant percentage of the power, as do the horizontal busses. Unfortunately, standard cell libraries do not make three-state enabled multiplexer standard cells, so there is no convenient way to disable this effect without adding extra cells, which increase the area and the parasitic capacitance of the switch. Our analysis indicates that it is preferable to add an array of N^2w gates (typically two-input NAND gates not shown in Table IV, or three-state buffers/inverters) at the interface between the horizontal busses and the multiplexer tree, as shown in the dotted box in Fig. 6(b). These gates can be used to remove the data flowing into significant fractions of the multiplexers trees. Fortunately, it is relatively easy to control these gates. The most significant (ms) control bit of each multiplexer tree (and its inverted value) can control all N gates at the input to that tree. If the ms bit$=0$, the lower half of the NAND gates in enabled, otherwise the upper half of the gates in enabled. This change will immediately disable one half of the data signals from propagating into every multiplexer tree, and reduce the power consumption of each multiplexer tree in (24) by one half. However, the area occupied by these cells must be reflected in (6), the delay while small should be added into (23), and the power required by the NAND gates must be added into (24). Most NAND gates will not switch, and their intrinsic capacitance will not contribute to their power dissipation.

This power-savings technique can be extended with negligible hardware cost, by including a small number of most-significant control bits. By considering the four most-significant bits, 16 "enable" control wires can be generated, each controlling the data flow into 1/16th of the multiplexer tree. The generation of each signal requires a four-input AND gate, a negligible overhead. This technique will reduce the total power consumed by the multiplexer trees (standard cells and wires) by a factor of 16 or $\approx 93\%$. One could decode all $\log_2 N$ bits and generate N enable signals, but full decoding will increase the cost of the decoding tree to a nonnegligible value.

C. Results—Unpipelined Crossbar

In the following graphs, the width of the crossbar datapaths is fixed at 8 bits. The preceding equations have been modified to reflect the previous two design optimizations. The memory area is not included in these graphs since it is constant, and the memory power is not included. The clock tree power is not included since this unpipelined crossbar switch has no internal clocked elements on the datapaths.

For an unpipelined switch with $N=256$, $w=8$, the area is 15.4 mm^2, relatively independent of the drive strength. Fig. 7(a) illustrates the critical path delay of the switch versus port size N. The delay also grows according to the port size N. As the drive strength increases, the delay decreases as expected. Fig. 7(b) illustrates the aggregate throughput of the switch versus the number of ports N. Fig. 7(c) illustrates the power dissipation of the switch, which grows as the switch size increases, or as the drive strength increases, as expected. A crossbar switch with 256 ports and a drive strength of 4 has an area of 15.4 mm^2, a delay of 3.66 nsec corresponding to a clock rate of 273 MHz, a throughput of 560 Gb/s, and a power dissipation of 1.4 W, corresponding to a bandwidth-power efficiency of 2.5 mW per Gb/s.

IV. Pipelining

In a crossbar switch, pipelining allows a smaller amount of VLSI area to switch a larger amount of bits. However, in the VLSI domain the pipeline latches introduce additional area and intrinsic capacitances, which may counteract the improvements due to pipelining. Our analytic methodology will be used to explore the effect of pipelining.

A. Broadcast Busses

To pipeline the multiplexer trees, a latch (DFF) can be added after each multiplexer standard cell in the tree. The latch stores the data, and provides the drive strength to drive the data through the wire and multiplexer to reach the latch at the next level of the tree. One may be tempted to place DFFs before the multiplexer tree input ports. However, this design option will require N^2w DFFs which would dramatically increase the area and power of the switch, and it is not pursued.

To balance the pipeline delays, pipeline latches should be added to reduce the delay along the horizontal busses. Typically, the horizontal wires are considerably more heavily loaded than the vertical wires, so more pipeline latches should be used on the horizontal wires to balance the delays per stage. Since the vertical wires use $\log_m N$ stages of latches, the horizontal busses can use $\mathrm{K}\log_m N$ pipeline stages, for some $K \geq 2$, to achieve similar latencies. Our analysis indicates that $K = 3$ reasonably balances the delays between the broadcast bus stages and the multiplexer tree stages. The area then becomes

$$\begin{aligned} A = {} & NwK\log_m N(A(\mathrm{dff}) + A(\mathrm{buff})) \\ & + Nw \cdot \frac{(N-1)}{(m-1)} \cdot (A(\mathrm{mux}) + A(\mathrm{dff})) \\ & + N \cdot \log_2(N) \cdot A(\mathrm{dff})) \end{aligned} \tag{33}$$

As a result of the area change, the horizontal and vertical dimensions H and V in (7) will also change. The maximum clock frequency is determined by the critical path. In the pipelined design above, the critical path is either (a) the time required to drive a horizontal bus segment and the first level of the multiplexer tree, or (b) the time to traverse the longest wires in the last stage of the average-case multiplexer tree, or (c) the time to traverse the wire leading from the root of an average-case multiplexer tree to the edge of the layout, similar to that shown in Fig. 6(b).

Assuming the latches are placed evenly along the horizontal bus, the delay of one pipelined horizontal bus stage becomes

$$\mathrm{D(H\ stage)} = 0.168 + 0.024\frac{(N \cdot C_{\mathrm{mx-in}} + H \cdot C_{\mu\mathrm{m}})}{\mathrm{K}\log_{\mathrm{m}}\mathrm{N} \cdot \mathrm{C_{std}(ds)}}. \tag{34}$$

A signal must propagate through a horizontal bus stage and the first level of multiplexers in one clock period, and the delay becomes

$$\begin{aligned} \mathrm{D(H\ stage)} = {} & 0.168 + 0.024\frac{(N \cdot C_{\mathrm{mx-in}} + H \cdot C_{\mu\mathrm{m}})}{\mathrm{K}\log_{\mathrm{m}}\mathrm{N} \cdot \mathrm{C_{std}(ds)}} \\ & + \left(0.240 + 0.031\frac{(\mathrm{H}/N)\mathrm{C}_\mu}{\mathrm{C_{std}(ds)}}\right). \end{aligned} \tag{35}$$

The last stage of the average case multiplexer tree has the longest wires into the multiplexer cells. Referring to (13) and (14), the best-case and worst-case trees each have a longest wire of length $\approx H(m-1)/m$ and $H(m-1)/(2m)$ respectively, for an average length *of* $(3/4)H(m-1)/m$. The delay through the wires leading into the root multiplexer is, therefore

$$\mathrm{D(mux\ stage)} = \left(0.240 + 0.31\frac{\mathrm{3H(m-1)C_\mu}}{\mathrm{4m} \cdot C_{\mathrm{std}}\mathrm{ds}}\right). \tag{36}$$

For a more accurate model, the capacitive load of the destination DFF can be added. The delay to traverse the wire to the edge of the layout, of length $\approx H/4$, is also given by (36), with the appropriate substitution. The clock period is determined by the largest delay and, therefore, the clock rate is

$$f_{\mathrm{clk}} = \frac{1}{\max(D(\mathrm{H\ stage}), D(\mathrm{mux\ stage}))}. \tag{37}$$

The power dissipated must be modified to include the power of the pipeline latches (DFFs) in the horizontal busses and the multiplexer trees, which is easily determined from (33). The power is given by

$$\begin{aligned} P(\mathrm{DFFs}) = {} & \frac{(Nw)f_d \cdot f_{\mathrm{clk}}(V_{\mathrm{dd}})^2}{2}\left(\frac{(N-1)}{(m-1)} + K\log_{mN}\right) \\ & \cdot (C_{\mathrm{dff-in}} + C_{\mathrm{dff-int}}). \end{aligned} \tag{38}$$

The latter two terms represent the input port and intrinsic capacitances of the DFF cell. The energy per bit switched (excluding the clock tree and memory) then becomes

$$\begin{aligned} E_{\mathrm{bit}} = {} & \frac{P(\mathrm{switch})}{f_{\mathrm{clk}} \cdot Nw} \\ = {} & \frac{1}{2}f_d(V_{\mathrm{dd}})^2\left(\frac{(N-1)}{(m-1)} \cdot (mC_{\mathrm{mx-in}} + C_{\mathrm{mx-int}})\right. \\ & + [H \cdot C_{\mu\mathrm{m}}] + \left[\frac{3m^2}{8(m-1)} \cdot HC_{\mu\mathrm{m}}\right] \\ & \left. + \left(\frac{(N-1)}{(m-1)} + K\log_{mN}\right) \cdot (C_{\mathrm{dff-in}} + C_{\mathrm{dff-int}})\right). \end{aligned} \tag{39}$$

The last term in the large bracket in (39) reflect the additional power consumed by the pipeline latches.

B. Design Optimizations

1) Power Reduction Technique: The same power reduction technique discussed earlier can be applied. We assume that 16 enable signals are generated from the 4 most significant control bits, which will reduce the total power consumption of the multiplexer trees.

2) Disabling DFFs in the Multiplexer Tree: A similar technique can be used to remove data signals from fractions of the multiplexer tree, as follows. The DFF standard cells have an asynchronous reset, which can be held asserted by an enable signal to hold the DFF output constant, effectively eliminating the data signal passing through that DFF. We assume the previous technique is used.

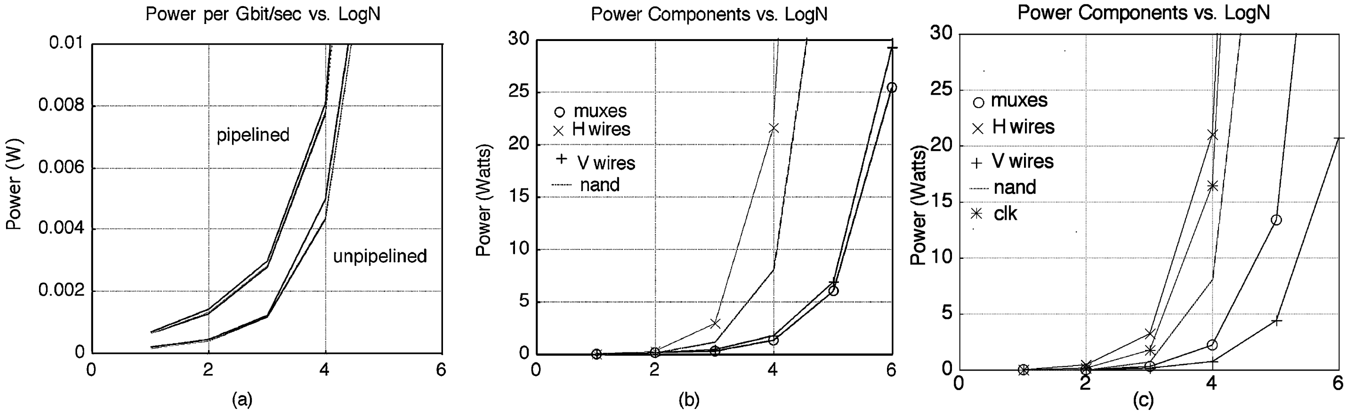


Fig. 8. Power analysis. (a) Power per Gb/s versus $\mathrm{Log}_4\mathrm{N}$. (b) Power profile versus $\mathrm{Log}_4\mathrm{N}$, unpipelined switch. (c) Power profile versus $\mathrm{Log}_4\mathrm{N}$, pipelined switch.

C. Results

In Fig. 7, the *datapath* width is fixed at 8 bits, and the analysis is adjusted so that all multiplexer and DFF cells have unity strength and inverters (INVX) cells are used to provide current drive. The design optimization to disable datapaths discussed previously using 16 enable signals has been incorporated. The memory area is not included in these graphs since it is constant, and the memory power is not included since it is relatively small and constant. The clock tree power is included since the pipelined crossbar switch has numerous internal clocked elements on the datapaths. We have adjusted the number of levels in the clock tree so that each leaf spans an area of $\leq 5000\mu\mathrm{m}^2$, equivalent to the area of 500 standard gates.

For a pipelined switch with $N = 256$, $w = 8$, the area is 26.8 mm^2, relatively indepdendent of the drive strength. Fig. 7(a) illustrates the delay of the switch versus port size (N). The delay of the pipelined switch has dropped dramatically. Fig. 7(b) illustrates the aggregate throughput versus the number of ports (N). The throughput of the pipelined switch has increased dramatically: Fig. 7(c) illustrates the power dissipation in the pipelined switch (nonoptimized). The power dissipation grows as the switch size increases, or as the drive strength increases. Compared to the unpipelined switch with $N = 256$ ports and drive strength$= 4$, the area has increased from 15.4 to 26.8 mm^2, the clock rate has increased from 273 MHz to $\approx$ 1.2 GHz, the throughput has increased from 560 Gb/s to 2.43 Tb/s, and the power dissipation has increased from 1.4 to 17.4 W. The bandwidth-power efficiency has dropped, from 2.5 to 7.2 mW per Gb/s.

V. Some Design Examples for Nonpipelined and Pipelined 5.12 Tb/s Switches

In this section, we explore the design of a 5.12 Tb/s single-chip optoelectronic switch built using 0.18-μm technology. Referring to Fig. 1, to build a switch with $N = 128$ fiber ribbons, a total capacity of 5.12 Tb/s will be required. We will search for a switch with 256 ports using degree 4 multiplexer cells with $\approx$ 5.12 Tb/s capacity.

Referring to Fig. 7, for $N = 256$ the maximum throughput of the unpipelined switch is $\approx$ 560 Gb/s. To achieve an aggregate throughput of 5.12 Tb/s the datapath width w must be increased, by linear extrapolation from 8 to 73 bits. After two iterations of mathematical analysis, the w is determined to be 127 bits. This large value of w will violate the vertical wire routing constraint in (16), and the area has been increased appropriately. The chip area will be 343 mm^2., the side length will be 18.5 mm, the aggregate throughput will be 5.13 Tb/s, and the power dissipation will be 26 W. This figure is for the crossbar switch alone, without the memory, arbitration logic, and optical I/O. When the power for the optical I/O and CDR is added from Section II the total power is $\approx$ 62 W. When the memory area is added, the power will be distributed over a larger silicon area, lowering the power density. According to the International Technology Roadmap for Semiconductors [5], in 2004 the allowable maximum power for chips with heat-sinks is 160 W, so the proposed switch is well within the upper bound on heat dissipation. Clearly, a single-chip nonpipelined switch with throughput 5.12 Tb/s will be large but feasible using 0.18-μm technology.

We now explore the pipelined switch design. Referring to Fig. 7, for $N = 256$ the maximum throughput is $\approx$ 2.43 Tb/s (for $w = 8$ and drive strength $= 4$). To achieve an aggregate throughput of 5.12 Tb/s, the datapath width w must be increased. After two iterations, the parameter w is determined to be 20. With drive strength $= 4$, the chip area will be 67 mm^2, the side length will be 8.1 mm, the clock rate will be 1 GHz, the aggregate throughput will be 5.14 Tb/s, and the power dissipation will be 42 W. Pipelining has reduced the VLSI area from 343 to 67 mm^2, although the power for switching has increased from 25.6 to 42 W. When the power for optical I/O is added, the pipelined switch consumes 80 W. When the area and power for memory is considered, the switch occupies < 2.5 cm^2 of area and consumes $<$ 85 Ws.

Fig. 8(a) illustrates the energy dissipation per bit for the unpipelined and pipelined switches with $w = 127$ and 20, respectively, expressed in mJ/s = mW per Gb/s. The energy dissipation grows essentially linearly with the port size N, as determined by (25). The energy dissipated per gigabit increases in the pipelined design. This result is expected; each bit still causes the same amount of charge displacement along the horizontal and vertical busses, and in the multiplexer trees. In the nonpipelined case the charge displacement occurs in one clock cycle, whereas in the pipelined case the charge displacement occurs over several clock cycles. However, the pipelined switch must also incur the charge displacement in all the pipeline latches, which can be a large component of the total power dissipation.

Figs. 8(b) and (c) illustrates the major components of the power dissipation in the unpipelined and pipelined designs. For the $N = 256$ pipelined switch, these include the multiplexer tree cells at 2.3 W (all data input port capacitances and intrinsic capacitances, plus the associated INV drivers), the wires in the multiplexer trees at 0.77 W, the horizontal broadcast busses at 21 W (which includes the NAND gate input ports), the NAND gates at 0.65 W (due to the intrinsic capacitance of switching gates only), and the clock tree at 17 W. The proposed power-savings technique reduces the multiplexer tree power by 88%, after the power of the NAND gates has been considered. The pipeline latches consume 2.9 W, although this power is also accounted for in the horizontal bus and multiplexer tree powers. Each curve represents the power dissipation of one component only.

Scalability: The limit to achieving higher bandwidths in the future will be the VLSI area and the power dissipation for larger OEIC switches. However, each of these issues will decrease with decreasing CMOS technology size. Our analytic methodology indicates that the same 5.12-Tb/s crossbar switch implemented in 0.09-μm technology will dissipate significantly less power. To achieve higher bandwidths, several single-chip optoelectronic switches can be arranged in a three-stage self-routing CLOS-like switch architecture, with optical interconnections between chips [15]. Using a self-routing 3 stage switch with sixteen 5.12-Tb/s optoelectronic switches per stage, the switch will support an peak capacity of $\approx$ 82 Tb/s, likely sufficient to meet the bandwidth needs of the next decade.

VI. Conclusion

An analytic methodology to determine the area, delay, throughput, and power complexity of a digital circuit, given a specific CMOS technology and standard cell library, has been proposed. Using the methodology, a closed-form analysis for the area, delay, throughput, and power dissipation of a conventional broadcast-and-select crossbar switch has been presented. In 0.18-μm CMOS technology, the optical I/O (VCSELs, VCSEL drivers, PD, PD receivers, and CDR circuitry) for a 5.12-Tb/s switch will require 36 W. It was shown that in large crossbars, significant power is dissipated in the horizontal broadcast busses and the multiplexer trees which process the broadcasts. To minimize power dissipation, the inclusion of circuitry to disable significant fractions of the multiplexer trees from switching is proposed. It was shown that pipelining can increase the aggregate capacity of the switch and reduce the VLSI area substantially, at a cost of a larger energy per bit switched. The power scalability analysis indicates that single-chip integrated optoelectronic switches with 5 Tb/s of capacity are feasible using current 0.18-μm technology, and single-chip switches should be scalable to 10–20 Tb/s capacity, with reasonable power dissipations, given smaller and faster CMOS technologies. However, the true power of this analytic methodology is the freedom the designer has in exploring numerous architectural tradeoffs, using different CMOS technologies (i.e., 0.13-, 0.09-μm CMOS) from different standard cell library vendors (i.e., American Semiconductor, TSMC). To perform the same architectural evaluations using existing CAD tools would require several hundred hours of VHDL coding time and CAD tool computation time, and would not yield much insight as to where the power is being consumed. In conclusion, optoelectronic crossbar switches may enable a new generation of powerful computing and communications systems in the forseeable future.

Acknowledgment

The authors would like to thank the anonymous reviewers for their thorough questions and suggestions. The use of CAD software and silicon CMOS VLSI foundry services from the Canadian Microelectronics Corporation are acknowledged.

References

[1] A. V. Krishnamoorthy and D. A. B. Miller, "Scaling optoelectronic-VLSI circuits into the 21st century: A technology roadmap," *IEEE J. Sel. Topics Quantum Electron.*, vol. 2, no. 1, pp. 55–76, Apr. 1996.

[2] MEL-ARI Technology Roadmap Optoelectronic Interconnects for Integrated Circuits (1999, Sep.). [Online]. Available: http://www.cordis.lu/esprit/src/melop-rm.htm

[3] T. Mudge, "Power: A first class architectural design constraint," *IEEE Computer*, pp. 52–58, May 2001.

[4] A. Alan *et al.*, "2001 technology roadmap for semiconductors," *IEEE Computer*, pp. 42–53, Jan. 2002.

[5] TQ8033 Data Sheet [Online]. Available: www.triquint.com

[6] 3.6 Gb/s 144 × 144 Crosspoint Switch Datasheet, VSC3140 [Online]. Available: www.vitesse.com

[7] K. Yun, "A terabit multiservice switch," *IEEE Micro*, vol. 21, no. 1, pp. 58–70, Jan.–Feb. 2001.

[8] T. Wu, C.-Y. Ying, and M. Hamdi, "A 2 Gb/s 256 × 256 CMOS crossbar switch fabric core design using pipelined MUX," in *Proc. IEEE Int. Symp. Circuits and Syst.*, vol. 2, May 2002, pp. 568–571.

[9] K.-Y. K. Chang, S.-T. Chuang, N. McKeown, and M. Horowitz, "A 50 Gb/s 32 × 32 CMOS crossbar chip using asymmetric serial links," in *Proc. IEEE 1999 Symp. VLSI Circuits*, Jun. 1999, pp. 17–19.

[10] J. Chang, S. Ravi, and A. Raghurathan, "Flexbar: A crossbar switching fabric with improved performance and utilization," in *Proc. IEEE Custom Integrated Circuits Conf.*, May 2002, pp. 405–408.

[11] J. G. Delgado-Frias *et al.*, "A VLSI crossbar switch with wrapped wave front arbitration," *IEEE Trans. Circuits Syst. I, Fundam. Theory Applicat.*, vol. 50, no. 1, pp. 135–141, Jan. 2003.

[12] F. M. Chiussi and A. Francini, "Scalable electronic packet switches," *IEEE J. Sel. Areas Comm.*, vol. 21, no. 4, pp. 486–500, May 2003.

[13] A. V. Krishnamoorthy *et al.*, "The AMOEBA switch: An optoelectronic switch for multiprocessor networking using dense-WDM," *IEEE J. Sel. Topics Quantum Electron.*, vol. 5, no. 2, pp. 261–275, Mar. 1999.

[14] T. H. Szymanski, M. Saint-Laurent, V. Tyan, A. Au, and B. Supmonchai, "Field programmable logic devices with optical I/O," in *OSA Appl. Opt.—Inform. Process.*, Feb. 10, 2000, pp. 721–732.

[15] T. H. Szymanski, "Design principles for practical self-routing nonblocking switching networks with $\mathrm{O(N \log N)}$ bit complexity," *IEEE Trans. Comput.*, vol. 46, no. 10, pp. 1057–1069, Oct. 1997.

[16] A. Kirk, D. Plant, T. H. Szymanski, Z. Vranesic, J. Trezza, F. Tooley, D. Rolston, M. Ayliffe, F. Lacroix, B. Robertson, E. Bernier, D. Brosseau, F. Michael, and E. Chuah, "A modulator-based multistage free-space optical interconnect system," *OSA Appl. Opt.—Inform. Process.*, vol. 42, no. 14, pp. 2465–2481, May 2003.

[17] T. H. Szymanski and H. S. Hinton, "Optoelectronic smart pixel array for a reconfigurable intelligent optical interconnect," U.S. Patent 6 016 211, Jan. 18, 2000.

[18] T. H. Szymanski, A. Au, M. LafreniereRoula, V. Tyan, B. Supmonchai, J. Wong, B. Zerrouck, and T. Obenaus, "Terabit optical LANs for multiprocessing systems," *OSA Appl. Opt.—Inform. Process.*, pp. 264–275, Jan. 1998.

[19] O. Kibar, P. J. Marchand, and S. C. Esener, "High-speed CMOS switch designs for free-space optoelectronic MINs," in *IEEE Trans. VLSI Syst.*, vol. 6, Sep. 1998, pp. 372–386.

[20] T. T. Ye, L. Benini, and G. De Micheli, "Analysis of power consumption on switch fabrics in network routers," in *Proc. 39th IEEE Design Automation Conf.*, Jun. 2002, pp. 524–529.

[21] E. Geethanali, V. Narayanan, and M. J. Irwin, "An analytical power estimation model for crossbar interconnects," in *Proc. 15th Annu. IEEE Int. ASIC/SOC Conf.*, Sep. 2002, pp. 119–123.

[22] D. Langen, A. Brinkmann, and U. Ruckert, "High level estimation of the area and power consumption of on-chip interconnects," in *Proc. 13th Annu. IEEE Int. Conf. ASIC/SOC*, Sep. 2000, pp. 297–301.

[23] M. J. Smith, *Application-Specific Integrated Circuits*. Boston, MA: Addison-Wesley, 1997.

[24] Paroli Optical Link Datasheet (V23814, V23815) (2002, July). [Online]. Available: www.Infineon.com

[25] Z. Mao and T. H. Szymanski, "A 4 Gb/s fully differential analog dual delay locked loop clock/data recovery circuit," in *IEEE Int. Conf. Electronics, Circuits, and Systems*, Nov. 2003, pp. 559–562.

[26] E. Yeung and M. A. Horowitz, "A 2.4 Gb/s/pin simultaneous bidirectional parallel link with per-pin skew compensation," *IEEE J. Solid-State Circuits*, vol. 35, no. 11, pp. 1619–1628, Nov. 2000.

[27] K. M. Geib, K. D. Choquette, D. K. Serkland, A. A. Allerman, and T. W. Hargett, "Fabrication and performance of two-dimensional matrix addressable arrays of integrated vertical-cavity lasers and resonant cavity photodetectors," *IEEE J. Sel. Topics Quantum Electron.*, vol. 8, no. 4, pp. 943–947, Jul. 2002.

[28] M. Ochiai, K. L. Lear, V. M. Hietala, H. Q. Hou, and H. Temkin, "Modulation properties of high-speed vertical cavity surface emitting lasers," in *Proc. 5th Device Research Conf. Dig.*, Jun. 1997, pp. 102–103.

[29] T. Knodl, H. Choy, J. Pan, R. King, R. Jager, G. Lullo, J. Ahadian, R. Ram, C. Fonstad Jr., and K. Ebeling, "RCE photodetectors based on VCSEL structures," *IEEE Photon. Technol. Lett.*, vol. 11, no. 10, pp. 1289–1291, Oct. 1999.

[30] D. V. Plant, M. B. Venditti, E. Laprise, J. Faucher, K. Razavi, M. Chateauneuf, A. G. Kirk, and J. S. Ahearn, "256 channel bi-directional optical interconnect using VCSELs and photodiodes on CMOS," *J. Lightw. Technol.*, vol. 19, no. 11, pp. 1093–1103, Nov. 2001.

[31] F. E. Kiamilev and A. V. Krishnamoorthy, "A high-speed 32-channel CMOS VCSEL driver with built-in self-test and clock generation circuitry," *IEEE J. Sel. Topics Quantum Electron.*, vol. 5, no. 2, pp. 287–295, Mar.–Apr. 1999.

[32] F. E. Kiamilev, "A 500 Mb/s 10/32 channel, 0.5 μm CMOS VCSEL driver with built-in self-test and clock generation circuitry," in *Proc. 1998 IEEE Lasers and Electro-Optics Society Annu. Meeting* , vol. 1, Dec. 1998, pp. 168–169.

[33] G. J. Simonis *et al.*, "1 Gb/s VCSEL/CMOS flip-chip 2-D-array interconnects and associated diffractive optics," in *Proc. 6th Int. Conf. Parallel Interconnects*, Oct. 1999, pp. 43–51.

[34] T. K. Woodward and A. V. Krishnamoorthy, "1-Gb/s integrated optical detectors and receivers in commercial CMOS technologies," *IEEE J. Sel. Topics Quantum Electron.*, vol. 5, no. 2, pp. 146–156, Mar.–Apr. 1999.

[35] ——, "1 Gb/s CMOS photoreceiver with integrated detector driving at 850 nm," *Electron. Lett.*, vol. 34, no. 12, pp. 1252–1253, 1998.

[36] T. K. Woodward, A. V. Krishnamoorthy, K. W. Goossen, J. A. Walker, J. E. Cunningham, R. E. Leibenguth, and W. Y. Jan, "1 Gb s single beam smart-pixel receiver transmitter realized in hybrid MQW-CMOS OE-VLSI technology," in *Proc. IEEE/LEOS Summer Topical Meeting Smart Pixels*, Monterey, CA, Jul. 1998, pp. 23–24.

[37] G. P. Agrawal, *Fiber-Optic Communication Systems*. New York: Wiley, 2002.

[38] T. H. Szymanski and V. Tyan, "Error and flow control for a terabit free-space optical backplane," *IEEE J. Sel. Topics Quantum Electron.*, pp. 339–352, Mar.–Apr. 1999.

[39] T. H. Szymanski, "Optical link optimization using embedded forward error correction," *IEEE J. Sel. Topics Quantum Electron.*, no. 6, pp. 647–656, Nov.–Dec. 2003.

[40] ——, "Bandwidth optimization of optical datalinks using error control codes," in *OSA Appl. Opt.—Inform. Process.*, Apr. 10, 2000, pp. 1761–1775.

[41] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches," *IEEE/ACM Trans. Networking*, vol. 7, no. 2, pp. 188–201, Apr. 1999.

[42] X. Zhang and L. N. Bhuyan, "Deficit round robin scheduling for input-queued switches," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 4, pp. 584–594, May–Apr. 2003.

[43] VST 0.18 μm CMOS Standard Cell Library (2002). [Online]. Available: www.tsmc.com

[44] 0.35-μm CMOS Standard Cell Library [Online]. Available: www.amis.com

[45] MOSIS Scalable CMOS Technology Design Rules, MOSIS Integrated Circuit Fabrication Service [Online]. Available: www.mosis.com

[46] N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design*, 2nd ed. Boston, MA: Addison-Wesley, 2001.

[47] A. Boni, A. Pierazzi, and D. Vecchi, "LVDS I/O interface for Gb/s-per-pin operation in 0.35 um CMOS," *IEEE J. Solid-State Circuits*, vol. 36, pp. 706–711, Apr. 2001.

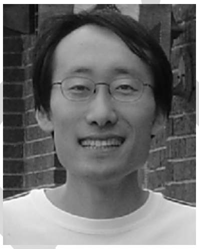
[48] DS90LV011A/048A Data Sheet [Online]. Available: http://www.national.com

[49] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2003.

Ted H. Szymanski (M'88) received the Ph.D. degree from the University of Toronto, Toronto, ON, Canada, in 1988.

He has held professorial positions at Columbia University, New York, and McGill University, Montreal, Montreal, QC, Canada. Currently, he is the L. R. Wilson–Bell Canada Enterprises Chair in Data Communications and the Associate Chairman (Undergraduate) within the Electrical and Computer Engineering Department, McMaster University, Hamilton, ON, Canada. He was the Principal Architect of a ten-year research program on Photonic Systems, funded by the Networks of Centers of Excellence program of Canada. The program brought together significant industrial and academic collaborators, including Nortel Networks, Newbridge Networks, Lucent Technologies, Lockheed-Martin/Sanders, McGill University, McMaster University, the University of Toronto, and Heriot–Watt University, to develop a free-space "intelligent optical backplane" exploiting emerging technologies. He holds a U.S. patent on "intelligent optical interconnects" using integrated optoelectronic technologies, along with Prof. S. Hinton, currently Dean of Engineering, Utah State University, Logan. He has presented numerous invited talks at international conferences and research institutes, has several invited book chapters on the topic of intelligent optical systems, and several of his papers have been reprinted in IEEE textbooks. He has also served on the technical program committees of numerous international conferences on optical systems. He is currently on sabbatical leave at the University of Victoria, BC.

Dr. Szymanski is a member of the Prairre Inn Harriers.



Honglin Wu received the B.Sc. degree in electrical engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 1995, and the M.A.Sc. degree in electrical engineering from McMaster University, Hamilton, ON, Canada.

He worked in the telecommunication industry from 1995 to 2001. He is currently working as a Research Associate at McMaster University. His research interests include deep submicron VLSI design for high performance, VLSI Interconnect modeling, synthesis, & optimization, system RTL performance modeling, and VLSI design & analysis of algorithms.

Mr. Wu received the OGSST Nortel Networking Scholarship from McMaster University in 2003.

Amir Gourgy (S'97) received the B.Eng. degree in computer engineering from McMaster University, Hamilton, ON, Canada, in 1998, and the M.A.Sc. degree in computer enigneering from the University of Waterloo, Waterloo, ON, Canada. He is currently working toward the Ph.D. degree at McMaster University.

He worked in the telecommunication industry from 2000 to 2002.