Multi-gigabit/second parallel fiber-optic ring network for multimedia applications

by

Bharath Raghavan

A Dissertation Presented to the

FACULTY OF THE GRADUATE SCHOOL

UNIVERSITY OF SOUTHERN CALIFORNIA

In Partial Fulfillment of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY

(ELECTRICAL ENGINEERING -- Electrophysics)

August 2001

Dedication

To my parents

Acknowledgments

I am extremely grateful to my advisor, Professor Anthony Levi, for his invaluable guidance and inspiration throughout the course of this work which was a very fulfilling learning experience. His conception and vision for this project shaped its course and the design targets achieved. This work would not have been possible but for the considerable resources that were provided for the two chips designed.

I was privileged to have on my defense committee Professor Alvin Despain and Professor Richard Kaplan. I am grateful to them for the significant interest they took in participating in my dissertation committee and their useful comments and suggestions. In addition, I also thank my qualifying committee members Professor John Silvester and Professor Peter Beerel for their very useful comments.

I am particularly thankful to Jeff Sondeen for his magnificent layout and layout extraction of the two chips designed for this project. His mastery over PERL and CAD tool maintenance made my design task considerably simpler. In addition, I thank other USC PONI team members Bindu Madhavan, Barton Sano, Youpyo Hong and Yongseon Koh for various technical discussions. I also thank the PONI team at Agilent Laboratories, Palo Alto, for their collaboration.

My graduate school was made that much more of a learning and enjoyable experience through my interaction with the various lab members - Ashok Kanjamala, David Cohen, Sumesh Thiyagarajan, Panduka Wijetunga, Mani Hossein-Zadeh, Young-Gook Kim, Tsu-Yau Chuang, Fernando Harriague and Yanko Todorov. I thank them sincerely for their

assistance over the years in many ways, in both technical as well as non-technical discussions.

A special note of gratitude and appreciation is owed to Kim Reid for her prompt and helpful assistance with all administration related work.

On a more personal note, I am very grateful to my friends, both far and near, for the emotional support they provided as I wended my way through graduate school.

Above all, I am eternally grateful to my family for their patience, unfailing support and encouragement on which foundation I could achieve the goal of obtaining a doctoral degree.

This research was supported partially by the Defense Advanced Research Projects Agency through the POLO and PONI projects with Agilent Technologies through research grants.

Table of Contents

List of Figures

List of Tables

Table of constants and values

c	Speed of Light	3.00×10^{8}	[m/s]
ε_0	Permittivity Constant	8.85×10^{-12}	[F/m]
ε_{si}	Permittivity of Silicon	1.0359×10^{-12}	[F/cm]
ε_{ox}	Permittivity of SiO_2	3.45×10^{-13}	[F/cm]
μ_0	Permeability Constant	1.26×10^{-6}	[H/m]
h	Planck's constant	6.626×10^{-34}	[J.s]
k_B	Boltzmann constant	1.38×10^{-23}	[J/K]
q	electron charge	1.602×10^{-19}	[Coulombs]
m_e	Electron rest mass	9.11×10^{-31}	[Kg]
σ_{Cu}	Conductivity of copper	5.76×10^{7}	[mho/m]

List of Symbols

C	Capacitance	[F]
C_L	Load capacitance	[F]
C_g	Gate capacitance	[F]
C_{gn}	Gate capacitance of an NMOS transistor	[F]
C_{gp}	Gate capacitance of a PMOS transistor	[F]
E	Electric Field	[V/m]
f	Frequency	[Hz]
I	Current	[A]
j	$\sqrt{-1}$	
L	Inductance	[H]
P	Power	[Watt]
q	Electron charge	[Coulombs]
R	Resistance	[Ω]
R_{AC}	AC-Resistance	[Ω]
R_{DC}	DC-Resistance	[Ω]
s	Laplace complex frequency	[rad/s]
t	Time	[s]
t_{ox}	Gate oxide thickness of MOS transistor	[m]
t_r	Rise time	[s]
t_f	Fall time	[s]
t_{dav}	Average delay	[s]
T	Period	[s]
T	Temperature	[°K or °C]
T_d	Delay	[s]
T_o	Period corresponding to frequency f_o	[s]
V	Voltage	[V]
V_{dd}	Power-supply voltage	[V]
V_{in}	Input voltage	[V]
V_{off}	Offset voltage	[V]
V_t	Threshold voltage	[V]
V_{tn}	Threshold voltage of an NMOS transistor	[V]
V_{tp}	Threshold voltage of a PMOS transistor	[V]
W	MOS transistor channel width	[m]
W_n	NMOS transistor channel width	[m]
W_p	NMOS transistor channel width	[m]
Y	Admittance	[S]
Z	Impedance	[Ω]
Z_{in}	Input impedance	[Ω]
Z_o	Output impedance	[Ω]

Z_O	Characteristic impedance	[Ω]
$Z_{O,o}$	Odd mode impedance	[Ω]
$Z_{O,e}$	Even mode impedance	[Ω]
δ	Skin depth	[m]
α	Conductor loss	[dB/m]
ε_{ox}	Permittivity of gate oxide in CMOS process	[F/cm]
$\Delta\phi$	rms frequency deviation	[s]
λ	wavelength	[m]
ρ	Resistivity	[Ωm]
σ	Standard deviation of a distribution	
σ	Conductivity	[Mho/m]
μ	Mobility of carriers in MOS transistor channel	[cm^2/V.s]
μ	Magnetic permeability of material	[Henry/m]
μ_n	Mobility of electrons in NMOS transistor channel	[cm^2/V.s]
μ_p	Mobility of holes in PMOS transistor channel	[cm^2/V.s]

List of Abbreviations and Acronyms

ABR	Available Bit Rate
AC	Alternating Current
AAL	ATM Adaptation Layer
AGP	Accelerated Graphics Protocol
ANSI	American National Standards Institute
APLS	Active Pulldown Level Shift
APLSD	Active Pulldown Level Shifted Diode
ASIC	Application Specific Integrated Circuit
ATM	Asynchronous Transfer Mode
BGA	Ball Grid Array
BER	Bit Error Ratio
BISDN	Broadband Integrated Services Digital Network
BW	BandWidth
CBR	Constant Bit Rate
CFR	Cambridge Fast Ring
CBN	Cambridge Backbone Network
CML	Current Mode Logic
CMOS	Complementary Metal Oxide Silicon
CRC	Cyclic Redundancy Checksum
CSMA/CD	Carrier Sense Multiple Access/Collision Detect
Clk	Clock
CSA	Communications Signal Analyzer
CTI	Coherent Toroidal Interconnect
DARPA	Defence Advanced Research Projects Agency
DC	Direct Current
DCVSL	Differential Cascode Voltage Switch Logic
DFF	Data Flip-Flop
DIBL	Drain Induced Barrier Lowering
DMA	Direct Memory Access
DOS	Disk Operating System
ECL	Emitter Coupled Logic
EOP	End Of Packet
EStore	Elastic Store
FDDI	Fiber Distributed Data Interface
FIFO	First-In First-Out memory
FF	Flip-Flop
FPGA	Field Programmable Gate Array
FTP	File Transfer Protocol
GaAs	Gallium Arsenide
Gb	Gigabit
GB	Gigabyte

GTL	Gunning Transistor Logic
GHz	GigaHertz
HDTV	High-Definition TeleVision
HP	Hewlett-Packard
HSTL	High-Speed Transceiver Logic
HIPPI	High Performance Parallel Interface
IBM	International Business Machines
IC	Integrated Circuit
ID	Identifier
I/O	Input/Output
I/P	Input
Kb	Kilobit
KB	KiloByte
LAC	Link Adapter Chip
LAN	Local Area Network
LVDS	Low Voltage Differential Signals
MAC	Medium Access Control
MAN	Metropolitan Area Network
Mb	Megabit
MB	Megabyte
MHz	MegaHertz
MM	Multi-Mode
MMF	Multi-Mode Fiber
MOSIS	Metal Oxide Semiconductor Implementation Service
MPEG	Moving Picture Experts Group
NSF	National Science Foundation
NIC	Network Interface Card
NMOS	N channel Metal Oxide Semiconductor
P2P	Point to Point
PC	personal computer
PCB	Printed Circuit Board
PCI	Peripheral Component Interconnect
PCM	pulse code modulation
PD	Phase Detector
PECL	Positive Emitter Coupled Logic
PIN	P-type Insulator N-type
PLL	Phase Locked Loop
PLLFS	Phase Locked Loop Frequency Synthesizer
PMOS	P channel Metal Oxide Semiconductor
POLO	Parallel Optical Link Organization
PONI	Parallel Optical Network Interface
PRBS	Pseudo-Random Bit Sequence
QFP	Quad-Flat Pack

QoS	Quality of Service
RAM	Random Access Memory
RF	Radio Frequency
Rx	Receive
RxFIFO	Receive FIFO
RxDATA	Receive data
SAN	System Area Network
SCI	Scalable Coherent Interconnect
SCI	Shielded Controlled Impedance
SCMOS	Scalable CMOS
SONET	Synchronous Optical NETwork
SMF	Single Mode Fibre
SRAM	Static Random Access Memory
TTL	Transistor to Transistor Logic
TSPC	True Single Phase Clocking
TxFIFO	Transmit FIFO
TxDATA	Transmit data
Tx	Transmit
UBR	Unspecified bit rate
VBR	Variable bit rate
VCI	Virtual Channel Identifier
VCO	Voltage Controlled Oscillator
VCSEL	Vertical Cavity Surface Emitting Laser
WAN	Wide Area Network
WDM	Wavelength Division Multiplexing
WSS	Wide-Sense Stationary
XOR	eXclusive-OR
XNOR	exclusive-NOR
demux	demultiplexer
mux	multiplexer
pk2pk	peak-to-peak
rms	root mean square
3-D	three-dimensional
e-mail	Electronic Mail
rt-VBR	Real-Time Variable Bit Rate
nrt-VBR	Non-realtime Variable Bit Rate

Abstract

The focus of this dissertation is an experimental investigation of how the immense bandwidth of emerging parallel fiber-optics and the continuing increase in speed and complexity of CMOS-based electronics impact slotted-ring networks for use in clustering computers spread over a "carpet floor." The motivation for this study is that the simplicity of a shared-medium ring network in comparison with more complex schemes such as mesh networks constructed using crossbar switches makes it a potentially cost-effective high-performance solution for small (less than 64 nodes) carpet cluster groups. Broadcast and multicast features necessary for real-time multimedia applications are easily implemented in a shared medium.

Using emerging low-skew parallel fiber-optic technology, clock may be transmitted in parallel with data. Thus noise-sensitive clock and data recovery circuitry required in traditional serial links can be eliminated resulting in reduced complexity and chip area. Further, the wide interface enables low-cost CMOS-based exploitation of the bandwidth capabilities of optical fiber.

We experimentally demonstrate that high data rates are thus achievable for slotted-ring networks thereby increasing bandwidth available per ring node. Measured network rate of over 16 Gb/s is achieved at 10.5 W power consumption in a 70 mm^2 size die implemented entirely in 0.5 μm CMOS technology. This is, to our best knowledge, the highest shared-medium ring network data rate reported to date. Enhancing the global clock distribution scheme can raise the achievable speed to 18 Gb/s as demonstrated

experimentally using a second chip that implements a point-to-point bidirectional link for connecting two computers.

In future CMOS technologies, ring network rates of over 100 Gb/s may be achievable. Issues relating to threshold voltage mismatch effects in sense-amplifiers, clock jitter at the network system level and printed circuit board interconnect performance will however need to be addressed.

Chapter 1

Introduction

The convergence of computing and communications in recent times has enabled a wholly new class of applications requiring bandwidths of several gigabits per second and having stringent delivery-time constraints. In the case of high-performance small geographical area cluster of computers spread over a "building carpet floor," several bandwidth intensive multimedia applications such as real-time multicast video-conferencing, remote three-dimensional (3-D) graphics rendering and collaborative multimedia creation and editing have emerged. Conventional networks based on serial physical layer links are severely strained to meet these demands at an affordable cost.

1.1 Broadband multimedia applications

The emerging applications in carpet cluster networks are broadband real-time multimedia applications which require a high quality of service from the network. A carpet cluster network refers to a small local area network composed of general purpose machines such as Intel Pentium-based personal computers (PCs) [1][2][3] distributed over an office floor or within a small building. Adjacent computers are separated by less than

100 m of interconnect distance and a typical number of machines in such a network is expected to be less than 64. An example of such a networked environment could be a professional studio involved in collaborative multimedia creation and editing.

One of the fastest growing sectors exhibiting broadband networking requirements is the digital entertainment business. The work flow in a typical digital studio is shown in Figure 1.1. The data that is moved in a day in a digital film studio can be several hundreds of gigabytes a day. Virtually every aspect of film and television post-production - such as color correction, on and off-line editing, special effects, broadcast rendering that aligns various video elements for playback - are handled inside a computer. The application may also be real-time such as collaborative interactive audio editing of high-quality music. The human ear can perceive variations in audio frequencies as low as 20 kHz. Real-time audio editing may require end-to-end propagation delays as low as 50 μs hence necessitating a low-latency network. In addition, collaborative applications may require efficient support of multicasting by the network.

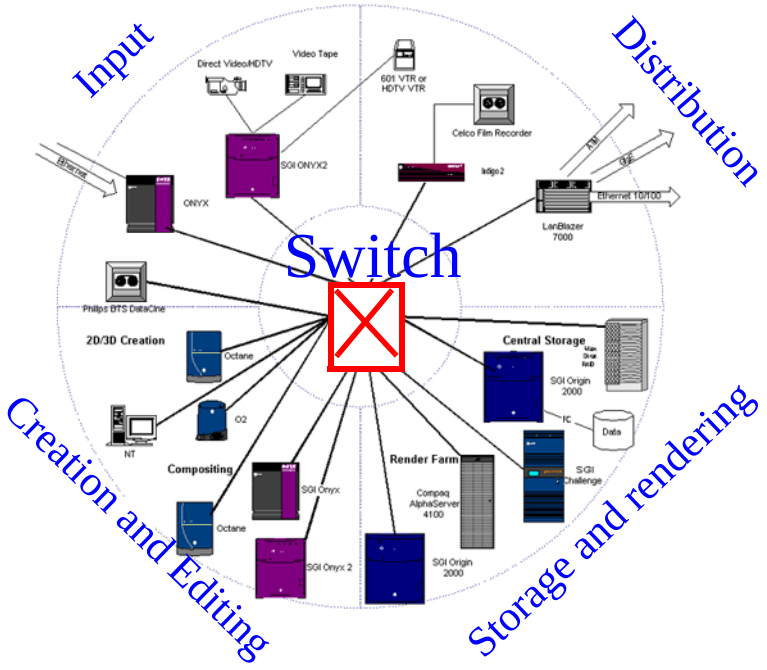


Figure 1.1: Work flow diagram of a typical digital studio. Activities in a typical studio involve input of multimedia data such as from cameras, creation and on-line real-time editing of multimedia data such as compositing, broadcast rendering that aligns various video elements for playback, storage in disks and distribution to external output points.

A second example of a broadband network environment is in the broadcast industry involving transmission of high definition television (HDTV) [4]. The work flow diagram in a HDTV studio is shown in Figure 1.2. Here, transmitting high-resolution uncompressed video of 640 x 480 pixels of 24-bit color at a frame rate of 30 frames per second results in results in network traffic of close to 200 MB/s. A half-hour sitcom with about 22 minutes of length excluding commercials can hence generate disk data of over 250 GB.

A more general description of network traffic types involves classification into synchronous and asynchronous categories.

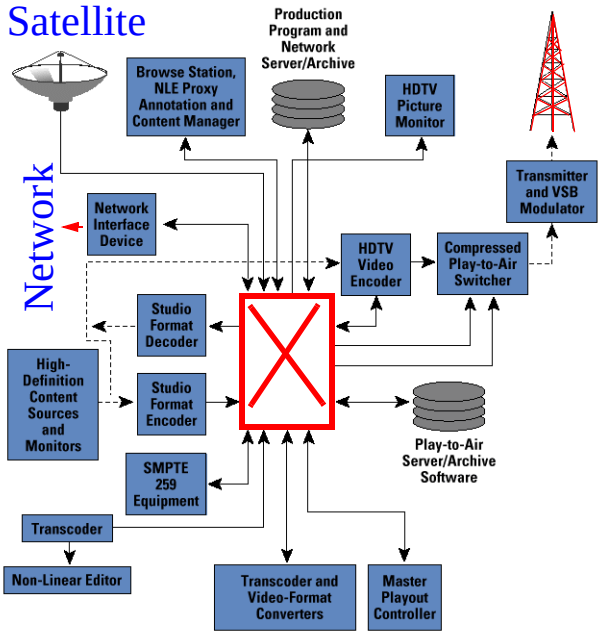


Figure 1.2: Constituents of a typical HDTV studio. Full-resolution uncompressed real-time HDTV signals at 30 frames per second generate as much as 200 MB of data per second. A half-hour sitcom (with about eight minutes of commercials) generates data in excess of 250 GB.

1.1.1 Synchronous traffic

This is a delay-sensitive traffic type with tight quality of service requirements. It could be further divided into constant bit rate and variable bit rate traffic types.

1.1.1.1 Constant bit rate (CBR)

This is also called isochronous traffic. Servicing this type of traffic needs the emulation of a circuit switched network. Here, data is generated in fixed amounts at fixed intervals of time. Some examples of CBR traffic are uncompressed video at 30 frames/second (one frame of 640 x 480 pixels using 24-bit color generates 220 Mb/s of

data) and uncompressed voice (e.g., pulse code modulation, PCM, encoder at 64 Kb/s). This kind of traffic requires low cell delay and low cell delay variation (i. e., jitter) support by means of a guaranteed bandwidth from the network. These applications are however tolerant to network losses and delays, in that they can tolerate occasional loss of quality. For instance, in an uncompressed video stream, the occasional loss of a frame may not hurt its quality as scenes usually do not change by much from one frame to another.

1.1.1.2 Variable bit rate (VBR)

This is also a form of synchronous traffic, but the traffic output is unpredictable and bursty in nature. Some examples are compressed video (output of MPEG-2 encoders) and voice (compressed voice with silence suppression). This kind of traffic again needs low delay and low jitter support from the network. For this kind of traffic, however, the occasional loss of quality can be severely detrimental to performance. For instance, in compressed video, there is not much repetition from one frame to another. Hence, the loss of a frame can have a severe impact on the quality of the received image.

1.1.2 Asynchronous traffic

This kind of traffic represents data that is generated at non-uniform rates. The delay and jitter requirements are not as stringent. However, the fairness property of the network must be observed. The asynchronous traffic could be further subdivided into several priority classes, such as higher priority interactive traffic (bursty traffic such as telnet and bulk transfers such as ftp) and lower priority non-interactive traffic (bulk transfers such as email).

A third classification is that described within the ATM community [5]. Here, the various traffic types are categorized under five service categories in relation to traffic management: constant bit rate (CBR), real-time variable bit rate (rt-VBR), non-real-time variable bit rate (nrt-VBR), unspecified bit rate (UBR), available bit rate (ABR). Practical networks provide up to four quality-of-service (QoS) classes to support the different traffic requirements. each of which has service parameters associated with it.

With CMOS fine-line dimensions projected to be as small as 0.05 μm, microprocessor clock speeds could be as high as several GHz [7]. Networked application performance will depend critically on the ability to move data rapidly. Emerging host I/O bus standards such as Rapid IO [8] illustrated in Figure 1.3 and Infiniband [9] illustrated in Figure 1.4 will enable I/O bandwidths on the order of gigabytes per second.

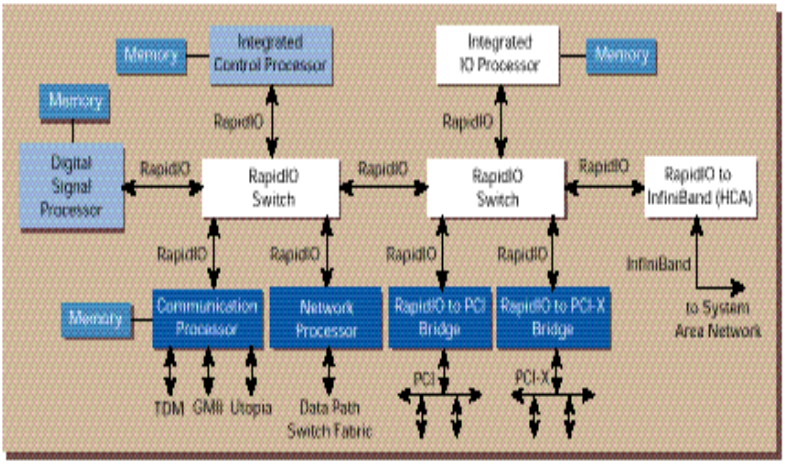


Figure 1.3: Schematic of switching architecture in the RapidIO interconnect scheme. This is primarily intended as a multiprocessor interconnect connecting chips and boards within a chassis. Network data rate could vary from 250 GB/s to 2 GB/s depending on bus width and clock rate.

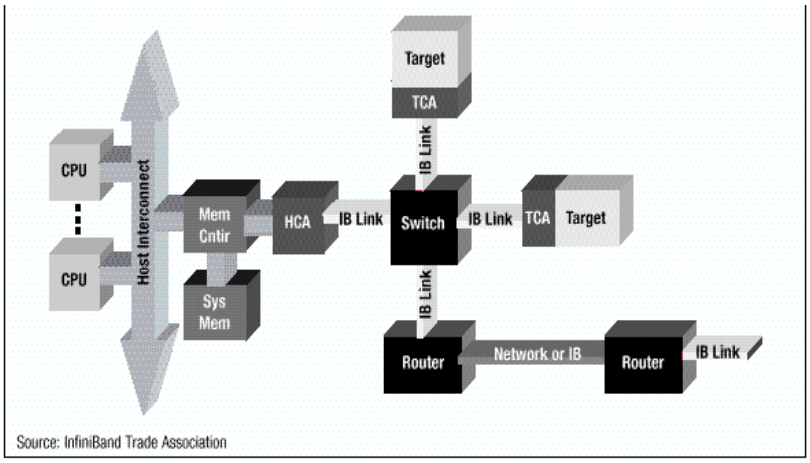


Figure 1.4: Schematic of the Infiniband I/O architecture. This is intended to be a multicomputer interconnect serving as an I/O interconnection standard for servers. It is a channel-based switch fabric I/O interconnect with data rates of 2.5 Gb/s per wire depending on a bus of one, four or twelve wires for different server types.

There is hence a need for high-bandwidth networks that are capable of handling the emerging traffic requirements cost-effectively.

1.2 Potential network solutions

Popular local area networks such as the 10 Mb/s CSMA/CD based Ethernet [10] and 100 Mb/s FDDI [20] are incapable of providing the performance needed for broadband real-time applications such as in the professional studio application described in previous sections. Many earlier gigabit commercial and research network implementations and proposals were serial link shared-medium bus and ring networks [11]-[24]. A shared medium network must have total bandwidth of several Gb/s to adequately support these emerging applications. Currently, the construction of a 10 Gb/s serial link would require

the use of expensive bipolar electronics and single-mode optical fiber technologies. Copper-based media cannot sustain these data rates for distances beyond a few meters [25].

Recent focus has shifted towards networks constructed using crossbar switches which have multiple input and output ports. Crossbar switches are attractive for features such as aggregate bandwidth and scalability. Examples of current gigabit packet networks are crossbar switch-based Gigabit Ethernet [25] and ATM [5] constructed using centralized crossbar switches as shown in Figure 1.5. Gigabit Ethernet can also be operated in the standard shared-bus configuration of Ethernet. Contention for the bus however degrades the available bandwidth. ATM switches provide sophisticated quality of service features. Admission control and congestion control in switches are however more complex to implement. The complexity may be unnecessary for networks of a smaller dimension. There is an $O(N^2)$ hardware complexity in crosspoint elements in an N-port crossbar switch. For example, a 64 port crosspoint switch features 4096 crosspoint elements while a 36 bit-wide 64-port crossbar switch features 147,456 crosspoint switching elements. This increases complexity of implementation and makes multicast and broadcast implementation particularly hard.

(a) Compaq Gigaswitch/Ethernet

(b) IBM 8265 Nways ATM switch

Figure 1.5: (a) Compaq Gigaswitch/Ethernet and (b) IBM 8265 Nways ATM switch. The Compaq switch has 60 100 Mb/s ports, or 24 1 Gb/s ports and provides multiprotocol/multilayer switching capabilities. The IBM switch provides for 56 155 Mb/s ports or 14 622 Mb/s ports with integrated ATM/Ethernet switching capabilities.

The advantages of shared-medium ring networks include the simplicity of the topology which leads to less complex hardware (*O(N)* complexity increase for each additional ring node) and allows the possibility of performance optimization at low cost. Importantly, deterministic bandwidth and delay bounds may be achieved without penalty of a significant drop in overall network performance. In addition, broadcast and multicast modes needed for multimedia applications are natural implementations. Congestion

control is implemented by traffic sources monitoring the current network traffic of the shared medium before initiating transmission, avoiding the need for backward propagation of congestion information to the sender.

Three media access control protocols for a multi-Gb/s ring network are token, slotted and buffer insertion. The increases in system I/O bus bandwidth have not kept pace with those in the network bandwidth. A single host cannot constantly transmit data to utilize the peak network bandwidth or receive at the same rate unless there are large buffers on the network interface chip. However, on-chip area for buffers is limited. A token ring protocol is hence inappropriate for a multi-Gb/s ring network since it leads to an under-utilization of network bandwidth. Thus some form of multiplexing is needed for medium access to the ring to ensure efficient bandwidth utilization. A slotted ring, where the available network bandwidth is sectioned into slots, is the simplest alternative and has the potential for lowest node latency. This is because the alternative buffer insertion ring requires considerable hardware to arbitrate access to the ring and buffer incoming cells while a host transmits. The simpler slotted ring protocol does not have to handle this complexity and simply waits for a free transmission slot. However, the bandwidth limitations of ring networks thus far implemented using serial links needs to be assessed in light of development in emerging technologies as will be outlined in the following section.

1.3 Broadband network enabling technologies

The continuing advances in Complementary Metal Oxide Semiconductor (CMOS) technology and shrinking of minimum feature sizes in accordance with “Moore’s Law” have enabled the development of high-performance processors and highly integrated

systems on a chip. Inexpensive Personal Computers (PC) have been consistently improving in performance with clock rates for the Intel Pentium III and Pentium IV processors [1] and the AMD Athlon processors [2] in excess of 1 GHz. They thus offer a cost-effective alternative to conventional workstations. By using a number of such machines in a high-bandwidth networked environment significant reconfigurable computing resources can be achieved at reasonable cost.

High-performance CMOS-based integrated circuits (ICs) can be designed to interface between a host computer and network [27] providing bandwidth capabilities that were only available previously using expensive technologies such as bipolar Emitter-Coupled Logic (ECL) [29]. CMOS-based IC design is attractive because it leverages the cost benefits of an inherently simpler process compared to silicon bipolar and leverages the infrastructure of high-volume commodity ICs for high-performance circuit applications. It also offers a higher level of integration and the potential to replace older multi-chip circuitry with single-chip CMOS-based solutions.

A continued shrinking of minimum feature sizes in CMOS has enabled higher switching speeds and more functionality on a single chip. With each technology generation, the semiconductor industry roadmap [7] calls for a 30% reduction in minimum feature size which hence results in a 30% reduction in gate delay, double the transistor density and a 30% to 65% reduction in the energy per transition. At 0.1 μm CMOS technology, microprocessor clocks could be as high as 3.5 GHz as seen from Figure 1.6.

This enables the implementation of fast and complex digital logic, high-speed analog circuitry and their integration onto a single-chip. Thus there is the potential for cost-effective CMOS-based exploitation of bandwidth capabilities of optical fiber.

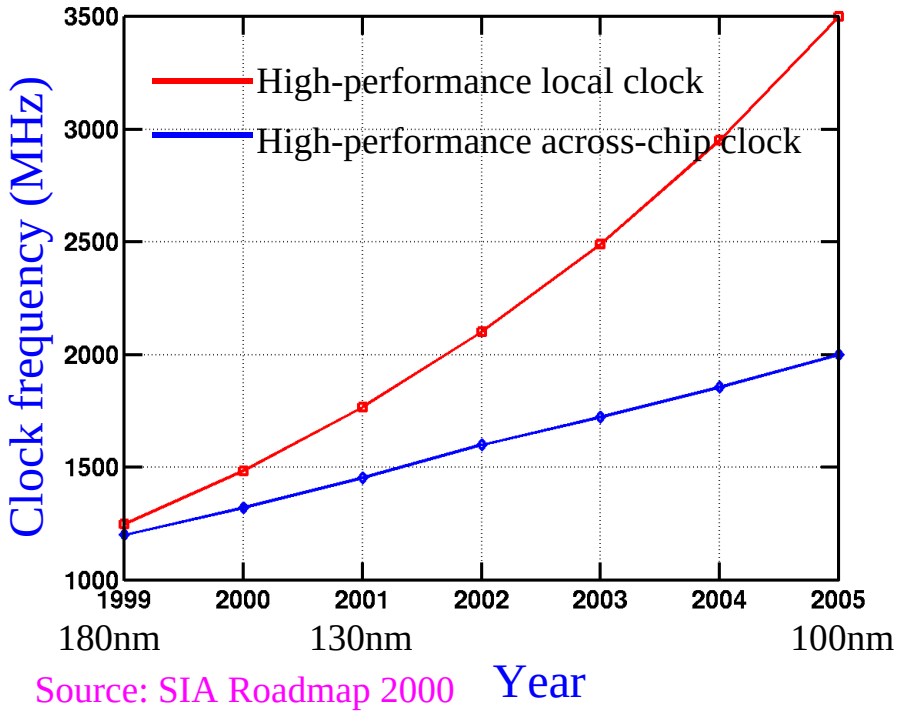


Figure 1.6: Semiconductor industry association roadmap [7] for the year 2000 showing expected scaling of high-performance local clock and high-performance across-chip clock until the year 2005. Microprocessor clock frequencies could be as high as 3.5 GHz using 0.1 μm CMOS processes.

The network physical medium needed to meet the requirements of a professional high-performance carpet cluster environment could be either electrical or optical [28]. Current small area networks are constructed using electrical links. Copper cables are however bulky and restricted in density, bandwidth and interconnect distance capabilities. For

example, the electrical link described in [25] provides a 10 Gb/s serial link over distances less than 20 m, using thick coaxial cable. In contrast, optical fiber-based links offer immense benefits including essentially unlimited bandwidth, immunity from electromagnetic interference, and a significantly higher edge connection density (form-factor). The fiber medium used in optical links could be either single-mode or multimode. While single-mode fiber is capable of achieving a higher bandwidth-distance product than multimode fiber, the cost of the transceiver module, packaging and fiber connectors is higher rendering it inappropriate for carpet cluster environments. A parallel fiber-ribbon constructed using multimode fibers is a natural and low-cost solution to provide the needed total interconnect-bandwidth while interfacing with the wide data buses of CMOS-based systems in a less complex, and power-efficient way compared to conventional serial link approaches. Multimode fiber-ribbon based links can be used at Gb/s/line signal rates over distances up to a kilometer [30]. Fiber Distributed Data Interface (FDDI) [20] and the Scalable Coherent Interface standard (SCI) [31] are examples of existing link standards that use serial fiber-optic technology. The High Performance Parallel Interface HIPPI-6400 is an emerging standard for parallel fiber-optic links [36].

The actual data bits are transmitted on fiber using light from laser diodes which is detected using photodiodes. Advances in Vertical Cavity Surface Emitting Lasers (VCSEL) [37][38] and new optoelectronic packaging technologies make possible the construction of low-cost optoelectronic components so that there is now the potential for constructing inexpensive high-performance network interfaces.

Recent advances in parallel fiber-optics technology such as the POLO and PONI parallel fiber-optic optoelectronic modules [39][40][41], work done on Jitney [45] and Paroli [42][43][44] links, have enabled a low-cost high-bandwidth physical medium. The PONI modules are designed for 12-wide fiber-ribbon. Separate transmitter modules based on VCSEL array laser technology and receiver modules based on photodiode arrays have been designed to support data rates of 2.5 Gb/s per multimode signal line at 850 nm wavelength. Conventional serial optical links require encoding of clock with the transmitted data stream. In contrast, parallel fiber-optics enables transmission of clock in parallel with data. Thus the use of costly and performance-limiting clock encoding or clock extraction circuitry is avoided. Furthermore, by transmitting data across multiple signal lines of a multimode fiber-ribbon its distance-bandwidth product limitations relative to single-mode fiber can be mitigated.

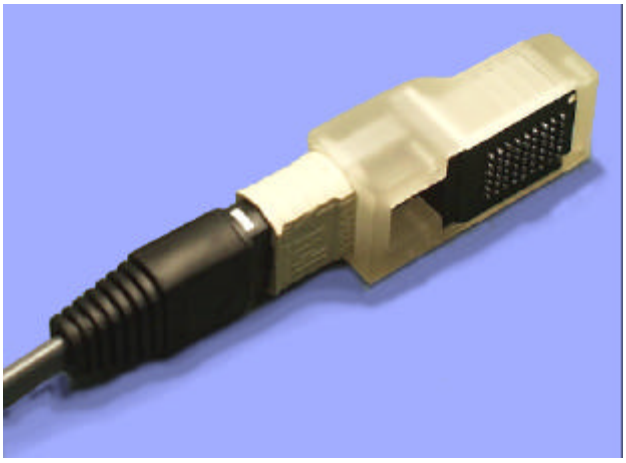


Figure 1.7: PONI module, now part numbers HFBR-712BP Tx VCSEL array or HFBR-722BP Rx GaAs PIN detector receiver array. The figure shows a single 12-wide transmitter module that can support data rates of 2.5 Gb/s per multimode fiber at 850 nm λ with low bit error ratio (BER<10^{-14}).

The module is shown on its side to expose the Ball Grid Array (BGA) on the underside of the package. The BGA provides electrical I/O and is surface mounted onto a printed circuit board occupying 1.5” x 0.5” of board space. The optical I/O to the module is via an MT push/pull fiber-ribbon connector seen on the left in the figure.

The use of parallel fiber-optic and advanced CMOS technologies to construct high-performance ring networks can thus potentially mitigate bandwidth limitations of previous serial-link ring networks and justifies a reassessment of rings for their effectiveness.

1.4 Conventional serial link-based approach

In a conventional serial link-based approach, clock information is embedded onto the data stream. This necessitates the use of clock encoding and clock extraction circuitry typically realized using phase-locked loops. A schematic of a simple data recovery scheme based on a phase-locked loop is shown in Figure 1.8. Embedded clock information in the received network data is extracted using a phase-locked loop. The extracted clock is then used to retime the incoming data using latches. To achieve high link phase margin a low-jitter phase-locked loop is necessary.

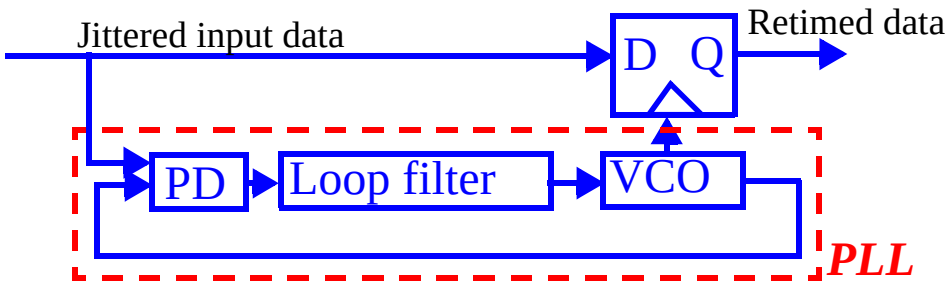


Figure 1.8: Schematic of a simple data recovery scheme using a phase-locked loop (PLL). Embedded clock information in the received network data is extracted using a phase-locked loop. The extracted clock is then used to retime the incoming data using latches. To achieve high link phase margin a low-jitter phase-locked loop is necessary.

The complexity in such a serial-link approach involves designing a low-jitter phase-locked loop [46][47] as well as the additional chip area and power consumed by the circuitry. An example of a 1.25 GHz phase-locked loop designed in 0.5 μm CMOS [48][49] is shown in Figure 1.9. The PLL exhibits peak-peak jitter below 40 ps. It measures 3.3 x 1.63 mm^2 with a loop filter capacitor of size 1.7 x 1.2 mm^2. The power consumed is 1.2 W. Using such a phase-locked loop for each of eight data channels results in an additional chip area of nearly 40 mm^2 and power consumption of 9.6 W for each of transmit and receive directions. This represents a substantial cost of chip area and power

consumption. From a link phase margin perspective, there is the complexity of designing an oscillator that is tolerant of supply and substrate noise [50][51][52] and a large loop filter capacitor to achieve low loop bandwidth and hence low jitter [53].

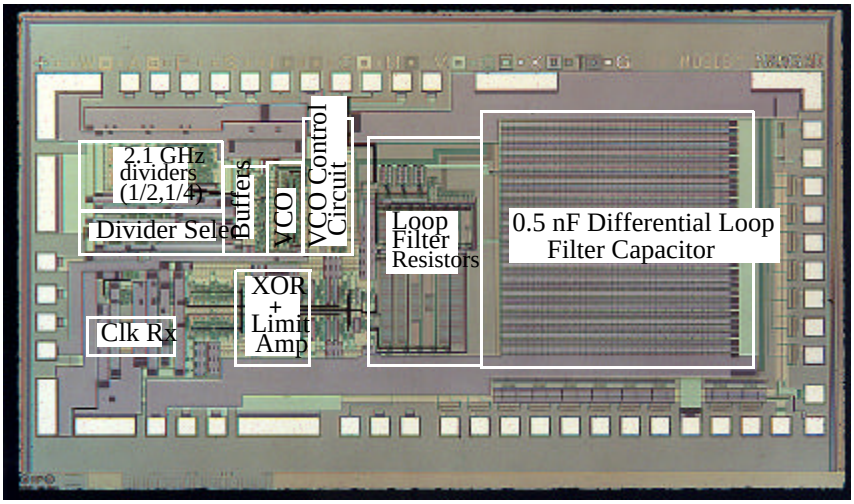


Figure 1.9: A 1.25 GHz phase-locked loop designed in 0.5 µm CMOS technology. The PLL exhibits peak-peak jitter below 40 ps. It measures 3.3 x 1.63 mm^2 with a loop filter capacitor of size 1.7 x 1.2 mm^2. The power consumed is 1.2 W. Using such a phase-locked loop for each of eight data channels results in an additional chip area of nearly 40 mm^2 and power consumption of 9.6 W for each of transmit and receive directions.

1.5 Thesis question

The technological advances described in Section 1.3 coupled with the inherent advantages of a shared medium ring network as outlined in Section 1.2 hence lead to the central question investigated in this dissertation described as follows:

How do the continuing advances in CMOS electronics and emerging parallel fiber-optics technologies impact the bandwidth capabilities of slotted-ring networks for carpet cluster applications?

Our hypothesis is that the simplicity of a shared-medium ring topology in contrast with crossbar switches allows the realization of very high data rates over parallel fiber using wholly CMOS-based electrical interfaces. Parallel low-skew fiber enables direct transmission of clock in parallel with data which mitigates the need for the complexity and expense of clock and data recovery circuitry and active deskewing circuitry. The simplicity of a slotted-ring network interface chip which avoids the quadratic increase in switching element complexity with number of ports in an N-port crossbar switch enables optimization for high digital logic data rates. Further, by eliminating noise-sensitive circuitry such as phase-locked loops (PLL) performance-reducing effects of digital switching noise on sensitive analog circuitry is reduced along with a reduction in total chip area. Coupled with broadcast and multicast advantages of shared-medium ring networks, this enables low-cost high-performance ring networks for carpet cluster applications which can potentially achieve 10 GB/s network rates in 0.1 μm CMOS technology.

The remaining sections describe how the hypothesis is experimentally verified. We experimentally benchmark the attainable ring network rate of 16 Gb/s achieved in a network interface chip implemented in 0.5 μm CMOS technology as representative of the optimally designed achievable rates for a ring network system (the network interface chip comprises of both the physical layer interface components as well as digital logic blocks implementing higher-level system tasks) in a current technology. Improving the clock distribution scheme will further raise this achievable rate to 18 Gb/s as demonstrated

experimentally using a second chip that implements a point-to-point link. From scaling considerations an attempt is made to predict future achievable rates in ring networks which could potentially be in excess of 10 GB/s. The results are useful to a network architect faced with system level choice regarding the use of a shared-medium ring network or a crossbar switch-based network. Issues relating to threshold voltage mismatch effects on memory speed in CMOS chips, clock jitter at the network system level and transmission line losses and reflections in printed circuit board interconnect performance will however have to be addressed to enable these higher rates.

1.6 Thesis contributions

The main contributions of this dissertation are:

1. Validation of the high data rates achievable in topologically simple shared-medium ring networks using single-chip CMOS network interfaces to parallel fiber-optic link technology, potentially enabling as much as 10 GB/s network data rates in future CMOS technology. Key differences of this network compared to previous networks include:

- Use of low-cost parallel multimode fiber-link technology to achieve 16 Gb/s network data rates in contrast with existing and previous Gb/s serial links which use expensive single-mode optical fiber-link technology. The data rate is to our best knowledge the highest achieved to date in a shared medium ring network.
- Direct transmission of clock in parallel with data thereby avoiding the need for clock and data recovery circuitry which is necessary in serial links. The need for area-consuming and noise-sensitive phase-locked loops on chip is thus eliminated, resulting in significant savings in chip complexity.

2. Components required to achieve over 500 MHz digital logic operation for the CMOS network interface chip, such as counters, FIFOs, RAMs and finite state machines.

3. A lightweight link layer protocol used for achieving multi-Gb/s network rates.

4. A study of the impact of scaling of electrical technology on ring network performance at the chip level, PCB board level and network system level. An evaluation of scaling of electrical CMOS technology is performed using simulations for four sub-micron CMOS processes indicating that over 100 Gb/s ring network bandwidths may be achievable in CMOS processes below 0.1 μm fine-line dimensions. Frequency variations between nodes can have a noticeable impact on network bandwidth utilization for even small variations at high frequencies. Transmission line losses due to skin and dielectric losses with scaling of PCB trace dimensions and measured link budget of 20 dB in 0.5 μm CMOS technology indicates that copper PCB interconnects may not be competitive with optics for distances exceeding a meter.

The organization of the rest of this dissertation is as follows. In Chapter 2, current and previous work on networks appropriate for local area networks is presented. In Chapter 3, the design of a network interface chip for a measured 18 Gb/s point-to-point link is described. In Chapter 4, the architecture and design of a slotted-ring network known as the PONI network is described. In Chapter 5, the network interface chip known as the LAC implementing the PONI network for a measured data rate of 16 Gb/s is described. In

Chapter 6, an investigation into impact of scaling CMOS process dimensions, clock jitter impact on network utilization and PCB interconnect scaling on electrical signal transmission is performed. In Chapter 7, some ideas for future work are presented.

Chapter 2

Related work

Up to the early 1990s, networks for LANs have predominantly been serial link shared-medium bus-based networks such as Ethernet [10] and ring networks such as FDDI [20] and token ring. Later networks have dominantly been crossbar or crosspoint switch-based networks such as Gigabit Ethernet [24] and ATM [6]. In this chapter, we briefly outline some of the characteristic features of previous serial link-based networks and emerging parallel link-based networks.

2.1 Medium access control (MAC) protocols

Networks differ in the topology, the physical media, the mechanism of sharing bandwidth and the technology used for implementation. Network bandwidth usage is arbitrated using medium access control (MAC) protocols. A network may support traffic that could be either synchronous with tight timing requirements or asynchronous with relaxed timing constraints. The goal of a MAC is to balance the twin goals of satisfying a node's bandwidth in a fair manner needs as well as maximizing the total network bandwidth usage.

A medium access protocol has the following characteristics:

- Immediate access under light load and a fairness control mechanisms for overload conditions.
- High utilization of the available network bandwidth.
- Low and bounded access delays for transmitting nodes and low packet delay variations at the destination end for delay sensitive traffic such as interactive video.
- Support for various priority levels and traffic classes; integrated services for a variety of data types such as image, video, audio, and traditional data such as file transfer.
- Simple, robust and easy to implement.

Fairness could be addressed globally where the usage of the network is equally shared between all competing nodes. It could also be addressed locally where fairness mechanisms operate only in heavily loaded segments where it is really needed. Thus, a node located in a lightly loaded segmented is not penalized or throttled by the fairness control mechanism as long as the needs of nodes positioned in heavily loaded segments are satisfied. This could also be addressed by giving nodes with heavier requirements such as schedulers more transmission opportunities. The following sections discuss how previous networks, which were configured as buses, rings or switches address these issues. The physical medium used could be serial link copper, serial link optical fiber or parallel link optical fiber.

2.2 Serial link networks

This section describes some of the serial link networks currently existing. They are: bus-based networks such as Gigabit Ethernet [24][26], DQDB [16], token ring networks such as FDDI [21][20], slotted ring networks such as Cambridge Fast Ring [13][23], Orwell Ring [14], MD^3Q [12] and ATMR [11], and buffer-insertion rings such as MetaRing [19] and CRMA-II [15].

2.2.1 DQDB (IEEE 902.6-1990)

DQDB [16] is a 150 Mb/s dual bus topology operating on the distributed queueing protocol and has been adopted as the subnetwork for the IEEE 802.6 metropolitan area network standard. The network can accommodate up to 512 stations over a geographical distance of up to 160 km. Access to the shared transmission medium is slotted. There are two slot generators to provide the slots in either direction. The slots are either queued arbitrated (QA) or pre-arbitrated (PA). Access to the QA slots is as follows. A node that wants to transmit in one direction of the bus places a request for a slot in the opposite direction. Each node maintains a request counter and a countdown counter. The request counter is incremented by each passing request and decremented by each passing empty slot counter. The counter cannot fall below zero. When a node generates a packet for transmission in a certain direction, in the basic protocol, it places a request in the opposite direction and transfers the contents of its request counter to the countdown counter (in some alternative protocols, a node need not place a request for its transmission if a slot is immediately available). As soon as the node has let an appropriate number of empty slots pass, it transmits in the next empty slot that it encounters. Multiple outstanding requests

for QA slots are not permitted. Hence, under ideal conditions of zero propagation delay and infinite slot reservation bandwidth, the protocol tries to achieve a performance as close to first-come-first-serve (FCFS) as possible. However, due to the finite propagation delay, the DQDB protocol is unfair under heavy load conditions. Hence, a concept called "bandwidth balancing" has been proposed [17] and has been incorporated as part of the standard. In this method, a node wastes a fraction of the bandwidth, as given by a parameter called BWB_MOD, that it is entitled to use in the standard DQDB protocol. Thus, the system settles into a balanced state where all active nodes share the available bandwidth equally. The rate of convergence to this stable equilibrium state is directly proportional to the amount of bandwidth that each node is willing to waste (i. e., its BWB_MOD parameter). A node can also be allocated some pre-arbitrated slots to support the transmission of real-time traffic with tight delay and delay variation constraints. The slot size is 53 bytes with a 5-byte header, thus making it compatible with the ATM networks [6].

There are position dependent fairness problems [17] in DQDB due its being a bus-based network. In DQDB, message size is short and each segment has to be individually scheduled, thus introducing a lot of overheads in bandwidth usage as well and increasing the segmentation/reassembly burden. DQDB performs better than FDDI at low to moderate loads, as well as for short message sizes. Its performance is relatively unaffected by network size as compared to FDDI. FDDI-II provides better isochronous service than DQDB.

2.2.2 MD^{3}Q

This network [12] extends the distributed queueing concept to be applied to dual ring architectures. The protocol employs an early cancellation of requests. It introduces the concepts of a minimum window size (MinWS) to avoid the need for bandwidth balancing (BWB) mechanism as well as to provide a faster convergence time to a balanced sharing of the bandwidth. By means of the MinWS parameter, a station is allowed to have multiple outstanding requests, thereby compensating for the positional discrepancy in DQDB. A guaranteed bandwidth mechanism is implemented by a constant rate of request generation.

2.2.3 FDDI (ANSI 1988) and FDDI-II

FDDI [21], [20] is a packet switched token ring network. The link speed in FDDI is 100 Mb/s (125 MBaud with 4B/5B encoding). The medium access protocol supports the transmission of both synchronous traffic with hard real-time requirements as well as asynchronous traffic with moderate delay constraints. Each station in the ring maintains a token rotation timer (TRT). A node holds the token as long as it has synchronous packets to transmit, or it has asynchronous packets and the timer has not expired (reached a maximum of T_{OPR}). The timer is enabled only for the transmission of asynchronous traffic. Thus, a guaranteed bandwidth is provided for synchronous traffic while the remaining bandwidth is shared in a fair and dynamic network load dependent manner among the asynchronous traffic using the timed token protocol.

FDDI-II is a hybrid mode protocol with both circuit switched as well as packet switched capabilities. It is capable of handling isochronous traffic, which provides a circuit switched emulation service for applications with even tighter real-time constraints than those of synchronous traffic. Here, a node which is designated as the cycle master creates a 125 μs frame which is used for the transmission of 16 isochronous channels, each capable of carrying 6.144 Mb/s for an aggregate circuit-switched bandwidth of 98.304 Mb/s. This service also guarantees 0.768 Mb/s for asynchronous traffic for a total channel bandwidth of 99.072 Mb/s. Any of the unused isochronous channels can be used for the transmission of asynchronous traffic according to the timed token protocol. Synchronous traffic is integrated along with the asynchronous traffic by assigning it the highest priority among the different service classes. The asynchronous traffic can be operated in the restricted token mode or in the normal token mode, of which the latter is further subdivided into eight priority classes. The remaining 0.928 Mb/s of the link are used for carrying the header overheads (0.16 Mb/s is used for transmission of a preamble and 0.768 Mb/s for the transmission of a cycle header which contains the programming template indicating which of the 16 channels are used for providing isochronous bandwidth).

FDDI was originally designed [20][21][22] with optical fiber as the intended transmission medium. The FDDI standard specifies a wavelength of 1300 nm and recommends the use of 62.5 μm/ 125 μm multimode fiber (for distances up to 2 km). However, it can be operated over single mode fiber over long distances (up to 60 km) as well as unshielded or shielded twisted pair over short distances (a maximum of 100 m).

The maximum slot size is 4500 bytes based on the maximum specified clock tolerance (0.001) between successive stations and the minimum size of the elasticity buffer. The network protocol can accommodate up to 500 stations on each ring, with a maximum ring length of 100 km (for a fiber ring length of up to 200 km on the dual ring) as determined by the maximum allowable slot size of 4500 bytes. A larger network would require a larger default token timer for operation. However, an increasing ring latency results in throughput degradation as well as increased packet delays. The bit error rate of the medium for a station to station link should not exceed 2.5E-10. FDDI employs a distributed clocking scheme and specifies the use of an elasticity buffer to compensate for frequency differences between ring stations, and a smoother to maintain the ring diameter constant, accounting for clocking jitter. It has two counter-rotating rings where normally, only one of the rings is used while the second ring is provided for reliability purposes. However, the second ring may be used for transmission as well. The FDDI protocol guarantees a maximum packet delay time of $2*T_{OPR}$ for the transmission of synchronous traffic. For heavily loaded networks (for instance, up to 60% of link capacity), the choice of T_{OPR} heavily influences the packet delay times (waiting time in transmit buffer + transmission time + propagation time).

2.2.4 Cambridge Fast Ring (1989; 100 Mb/s)

The Cambridge Fast Ring [13], implemented in 1989, was a slotted ring with two modes of operation - a normal mode and a slotted mode. The normal mode slots are released at the source and cannot be reused by the station. The channel mode slots may be reused by the source. A node is allowed to use only one slot at a time. A later

implementation, the Cambridge Backbone Network [23] relaxed this restriction. The CFR link speed was targeted for 100 Mb/s. The first implementation was realized at 60 Mb/s. The repeater was constructed using ECL logic while the network controller was constructed using CMOS 3 μm technology. The node latency is greater than 4 μs. The CBN, also using similar technology, achieved a link speed of 500 Mb/s.

2.2.5 ATMR (1991; 622 Mb/s)

In this network [11], there are different priority classes to handle multimedia traffic with different QOS requirements. For QOS support within a priority class, the ATMR protocol uses two concepts for its functioning: window size and reset period. The protocol working within one priority class is as follows. Each of the ring nodes is assigned a transmission quota indicated by the window size. A node can transmit as long as it has a valid quota. When all nodes have exhausted their transmission quotas, a reset is issued which refreshes every node's quota.

The ATMR ring network chip was implemented in 0.8 μm CMOS technology to obtain a link speed of 622 Mb/s. ATMR is designed to be operated in an ATM-LAN. The slot size is hence the same as that used for an ATM ring, i. e. 53 bytes with a 5 byte header.

ATMR provides a sophisticated cycle reset mechanism. There is a drop in the network throughput however while going across from one reservation cycle to another, as nodes that have exhausted their quotas have to wait until all ring nodes have exhausted their quotas before a new cycle is initiated.

2.2.6 CRMA-II (1991; 2.4 Gb/s)

This network [15] employs a dual slotted ring with spatial reuse. The buffer insertion mode is enabled for the transmission of contiguous slots. In this protocol, there is a cycle based reservation scheme where fairness is maintained across consecutive cycles. A scheduler collects the transmission requests from each node. It then uses a scheduling algorithm to set the fairness threshold and nodes are assigned quotas based on the history of their transmissions in the previous cycle as well as their current transmission needs. The scheduler then informs each node of its transmission quota. Then, the scheduler marks an appropriate number of slots as being reserved. Only nodes with a valid reservation quota can access these reserved slots. A node that accesses a reserved slot changes the slot mode to gratis after loading its message. After being released at the destination, this slot can be used by any node, which maintains a count of all such gratis slot accesses. The gratis counter contents are communicated to the scheduler to be used for establishing fairness in the next scheduling cycle, so that nodes that used a lot of gratis slots in the current cycle are throttled in the next cycle.

An experimental testbed was setup at a link speed of 2.4 Gb/s [86]. An earlier version of the CRMA protocol for a folded, slotted bus was implemented [87] at a link speed of 1.13 Gb/s using a combination of GaAs for the serializing/deserializing high speed functions, ECL for demultiplexing, TTL technology for 8B/10B encoding/decoding and CMOS technology for buffering. The transmission medium used was single-mode optical

fibers driven by a distributed feedback laser with a carrier wavelength of 1310 nm. The network was prototyped using commercial off-the-shelf components. The MAC logic was performed at a clock speed of 47 MHz.

There is a problem of node starvation while going across from one cycle to another while the scheduler recomputes the transmission quotas for each node. There could also be unnecessary throttling of a node that needs a larger share of the bandwidth, such as a scheduler. There is a means of avoiding this situation by introducing the concept of a throughput class whereby, each node is allocated different amounts of bandwidth based on its throughput class.

2.2.7 MetaRing (1989)

This is a buffer insertion dual ring [19] which can also be operated in the slotted mode to minimize delay. The protocol utilizes a single bit control signal known as SAT, which acts like a global token and quota-refresh mechanism. In this scheme, each of the ring nodes is assigned a minimum *l* and maximum *k* quota of transmissions that it can perform in every round-trip of the SAT signal. A node holds the SAT signal as long until it is able to transmit at least *l* messages. By holding the SAT signal, it signals upstream nodes to stop the transmission of asynchronous traffic. When a node is satisfied according to the SAT-algorithm, it passes on the SAT signal. A SAT signal can be propagated in the direction of data flow, or in the opposite direction. The latter yields better network performance. After a node passes on the SAT signal, it can transmit up to *k* more messages. The arrival of a SAT signal refreshes each node's quota. Isochronous traffic is transmitted using an integration mechanism similar to FDDI by using three ASYNC-EN

signals - ASYNC-EN(GR), ASYNC-EN(YL), ASYNC-EN(RD) in combination with the SAT signal. The SAT signal is used to ensure the fairness of asynchronous traffic, while the ASYNC-EN signals are used to enable and disable the transmission of asynchronous traffic. Synchronous traffic is given higher priority over asynchronous traffic for access to the transmission medium. Its transmission is initiated by a call setup procedure. Any node that has a backlog of synchronous traffic utilizes the ASYNC-EN signals to stop other nodes from transmitting and thus obtain transmission rights.

This network was prototyped at the IBM T. J. Watson Research Center in 1989 at a link speed of 100 Mb/s. A gigabit implementation was constructed in 1990 for IBM's participation in the NSF/DARPA Aurora gigabit testbed. The MetaRing's SAT fairness algorithm has now been incorporated in the SSA (serial storage architecture) standard ANSI X3T10. This is a full-duplex ring used to interconnect computer peripherals such as disk drives, tape drives, and optical drives to computer workstations, servers and storage subsystems.

The performance degrades at low to moderate loads for large rings, since there is a latency period equal to the round-trip pass of the SAT signal (or token). There is no literature currently available on the performance analysis of the new ASYNC-EN signal based protocol.

2.2.8 Gigabit Ethernet (1998)

Gigabit Ethernet [24][26] is specified to run at a maximum data rate of 1 Gb/s over 550 m when using multimode fiber. The network can operate in half-duplex or full-duplex modes. Flow control in full-duplex communications is implemented by receivers

monitoring the receive buffer space. Senders are signalled to stop transmitting when receiver buffer runs out of space as indicated by a maximum allowable occupancy parameter while transmission can restart when buffer space falls below a minimum allowable occupancy parameter. Gigabit Ethernet is specified to run over many cable types including coaxial cable, twisted pair and multimode fiber.

Collisions due to contention in the half-duplex mode which uses the traditional CSMA/CD Ethernet MAC protocol degrade the available 1 Gb/s bandwidth. Gigabit Ethernet is more efficient in the full-duplex mode where collisions are avoided. A multiported Gigabit Ethernet switch is however necessary to fully avoid collisions in addition to the network interface cards (NICs) in each host. Buffering requirements in switches are higher due to the possibility of output port contention. Gigabit Ethernet switches commercially available provide for sophisticated quality of service features and multiprotocol capabilities and they are likely to be more expensive than individual line cards that are sufficient for a ring network. Broadcast and multicast modes are particularly hard tasks for crossbar switches.

2.2.9 10 Gigabit Ethernet (standardization in progress)

Gigabit Ethernet is further extended to the next generation of Ten Gigabit Ethernet [54] which raises the link data rate to 10 Gb/s. This is specified to run over multimode fiber and single-mode fiber. The multimode fiber links may be serial or use parallel link technology in the form of fiber-ribbon or WDM technology. There are two changes specified for Ten Gigabit Ethernet to extend its applicability to the metropolitan area

network (MAN) and wide area network (WAN) domains. The first is a specification for single mode fiber for distances greater than 40 km. Secondly, it is also intended to run 10 Gigabit Ethernet over OC-192 Sonet links.

2.2.10 Fibre Channel (ANSI Std. 1993)

Fibre Channel [125] was initially created as a system area network to interconnect computers, storage devices, displays and other peripherals, but can also be used as in local area network applications. It can be connected in switch, point-to-point or ring configurations. The serial link fibre channel was specified to run at a maximum of 1 Gb/s scalable to a maximum of 4 Gb/s. Due to the provisioning of a switch architecture, a credit based flow control is implemented. Fibre Channel is also currently being defined to operate using the HIPPI standard.

2.3 Parallel link networks

In this section, there is a description of networks based on parallel link physical media which have been proposed or adopted as industry standards. The parallel link may be implemented using fiber-ribbon technology or using WDM technology. Some of these networks such as SCI and 10 Gigabit Ethernet may also be realized using serial link implementations while HIPPI-6400 may also be realized using parallel electrical links.

2.3.1 HIPPI-6400 (1998)

HIPPI-6400 [36][35] is a high-performance parallel interconnect with a link data rate of 6.4 Gb/s. It is a networking technology targeted for local area networks (LAN) and system area networks (SAN). It employs wormhole routing to utilize the available bandwidth with very low overhead. It also features four virtual channels as a multiplexing

mechanism to minimize blocking of message transmissions by large messages. Thus, also network bandwidth can be better utilized. Reliability functions such as error detection (through the use of a link-level CRC) and retransmission (using a go-back-N scheme) as well as flow control functions (based on a link-level hop-by-hop credit-based scheme) are incorporated into the hardware. This relieves burden on the network protocol stack, thus enabling higher application throughputs.

The cell (basic packet transfer unit) size in HIPPI-6400 is 32-data bytes and eight control bytes. HIPPI-6400 defines a parallel copper cable interface that uses 16 data lines, 4 control lines, 1 framing line and two clocks, for a total of twenty-three lines in each direction. Data transmitted onto the link is encoded using 4B/5B encoding. The standard allows for up to 10 ns of differential skew between signal lines at the receiver. Deskew circuits in the receiver based on tapped delay lines compensate for skew across signal lines.

2.3.2 SCI (IEEE standard 1596 - 1992)

SCI (Scalable Coherent Interface) [32] is a high-speed interconnection system optimized specifically for a distributed shared memory system. But though it was originally intended to interconnect commodity computers with high bandwidth and low software overhead to realize supercomputer-class computation it is also applicable for LAN environments. The design goal was to minimize the role of software protocols in exchanging data by reducing data copies and minimizing the role of the operating system in transferring data across processors. Instead, data is exchanged directly between memory physically distributed across multiple processors but configured in a single flat

address space by direct memory access (DMA) operations. SCI nodes could be interconnected to construct buses, rings or mesh-based networks. A buffer-insertion ring can be constructed using SCI nodes and larger networks can be constructed using rings interconnected by SCI switches. Data is transferred over an SCI network using read and write transactions. Each of these transactions is composed of a request from the sender and end-to-end responses. Where reliable transfer is not required, there is also provision for move operations which are unconfirmed writes. There are two allocation protocols for sharing link bandwidth called the bandwidth allocation protocol and queue allocation protocol. The bandwidth allocation protocol can be used when there is contention for the link bandwidth along a packet's path, wherein link bandwidth can be assigned to a specific node. The queue allocation protocol can be used when there is contention for queue buffers along the path of a packet, wherein buffer space can be reserved for senders that have been previously unsuccessful in transmitting due to inadequate buffer space.

Fiber-ribbon is a natural choice for SCI [33] to realize GB/s implementation. A prototype was implemented to demonstrate a 1 GB/s SCI link [34] in 0.8 μm BiCMOS.

2.3.3 POLAR

A chip-set for parallel optical links called PAROLI [42][43] was developed by SIEMENS for high-speed data transmission. The system consists of a transmit module and a receiver module for 12-wide fiber-ribbon for 2.5 Gb/s per signal line to give an aggregate throughput capability of 30 Gb/s over 300 m link distances. An FPGA-based approach was used to design a full duplex parallel fiber-optic link using the PAROLI chip-

set called POLAR (Parallel Optical Link Architecture) [55]. A link controller was designed to implement a transmission protocol using Xilinx FPGAs [56] to realize a sustained data rate of 32 bits at 40 MHz, effectively 1.28 Gb/s.

2.3.4 Control Channel Fiber-ribbon pipeline ring network (1999)

This is a unidirectional ring network [57] based on fiber-ribbon technology developed by Motorola called OPTOBUS [58]. It is designed for parallel processing and distributed real-time systems. The fiber-ribbon is composed of ten parallel fibers, eight of which are used for data, one for clock and one for the control signal. The control channel enables low level support for barrier synchronization, global reduction, and reliable transmission. Access to the network bandwidth is using a time division multiplexing access, similar to a slotted ring network with slots being released at the destination and being available for immediate reuse. There is one slot per node on the ring. In each slot a node passes or transmits one control packet and one data packet. The control packet is used for controlling access permission for the next data slot. Slots can also be reserved for real-time applications. A prototype was built using commercial DSP board for protocol processing, an I/O board, memory and control logic implemented in an FPGA. Test results are not available. In an alternate version of the fiber-ribbon pipeline ring network proposal, nine of the ten fibers are proposed to be used for circuit-switched traffic (eight data lines and one clock line) while the tenth is proposed to be used for packet-switched traffic using, for example, a token ring protocol. The tenth fiber also carries the control packets described above.

2.3.5 HORNET (2000)

This is an experimental ring network based on WDM technology [59][60] intended for MAN networks. This network has a multiple access architecture, in which nodes access any WDM channel using a medium access control protocol called CSMA/CA and fast tunable laser transmitters. Data packets are directly transported over the WDM ring, eliminating the SONET transport.

The medium access control (MAC) protocol is a distributed scheme. The network is designed to scale to 100 access points (or ring nodes) with a circumference of around 100 km. The data rate is specified at 2.5 Gb/s per wavelength. It has a tunable laser transmitter with a fixed receiver design. The MAC protocol for sharing bandwidth per wavelength uses a carrier sense multiple access scheme with collision avoidance based on subcarrier signaling. Each network wavelength has an associated unique subcarrier frequency which carries header information and can be monitored in the RF domain. The subcarrier provides information relating to wavelength occupancy and destination address. The wavelength availability information is used in arbitrating transmission onto a wavelength, thereby avoiding collision.

The prototype implementation [60] operates with two channels using two wavelengths operating at 2.5 Gb/s data rates each.

2.4 Summary of network protocols

The performance of all networks involves some trade-offs and depends strongly on the network parameters chosen - the transmission quotas in case of the quota-based networks and the control cycle duration in case of the reservation schemes such as CRMA-II - and

the current network load. Early serial link ring networks such as MetaRing, CRMA-II, ATMR and MD^3Q and serial link bus networks such as DQDB were overshadowed by crossbar switches due to scalability reasons and because the available bandwidth for a node falls inversely in proportion to the total number of nodes connected to the ring network. In recent times, the focus has shifted towards crossbar switch based networks such as Gigabit Ethernet. However, the crossbar switches suffer from a quadratic increase in switching element complexity with number of ports. Problems relating to output port contention in a mesh network or multistage switching networks necessitate extensive buffering and complex flow control protocols. Besides, multicast and broadcast are hard problems for crossbar switch-based networks unlike in a shared medium ring network where they are natural implementations.

While the front-end analog circuitry in serial links were earlier realized using analog bipolar circuitry, they are increasingly being replaced by CMOS solutions. Parallel link technology based on fiber-ribbon and WDM is a relatively new technology that can yield high data rates through CMOS implementation. In subsequent chapters, we shall describe our attempt at constructing a broadband ring network that will mitigate the bandwidth concerns of traditional ring networks. As a first step towards this end, we constructed an 18 Gb/s chip for a point-to-point link which is described in the following chapter.

Chapter 3

Point-to-point link IC

3.1 Introduction

In this chapter, we present details of an implemented broadband point-to-point data communication link using a mixed-signal link interface chip known as the P2P. The P2P was constructed to validate components key to constructing a full-fledged ring network interface chip which will be described in Chapter 5. It also attempts to highlight issues in designing direct CMOS interfaces to parallel fiber-optics that effectively exploit the bandwidth advantages offered by optical interconnects. The P2P was fabricated in a standard 0.5 μm three-layer metal HP CMOS process. Measured data rates of over 18 Gb/s were obtained on its independent transmit and receive physical layer ports at peak power consumption of less than 5.5 W.

The P2P chip is designed to operate over a 10-wide multimode glass fiber (MMF) ribbon physical layer that provides eight data lines, and separate clock and frame control signal lines. The key innovation afforded by parallel fiber-optics is thus two-fold. Firstly, since clock is transmitted in parallel with data the use of complex clock encoding and recovery circuitry necessary in a conventional serial link is avoided. Secondly, high

aggregate data rate is achieved cost-effectively by transmitting on multiple multimode fiber signal lines in parallel. This enables the use of conventional low-cost CMOS processes to achieve the desired signalling rate as opposed to using expensive bipolar technology.

Techniques used to achieve the desired speed in the digital logic circuitry which interfaces a host computer to the high-speed fiber-optic physical layer ports include dynamic TSPC logic style, deep pipelining with single gate delay per stage for the digital logic controllers and a high-speed FIFO architecture. They are discussed in greater detail in section 5.5 on page 135. The physical layer interface analog circuitry produces LVDS signal output format [74] at half-speed clocking over the external interface and uses reduced swing differential clocked circuitry for its internal multiplexing and demultiplexing circuitry.

The chapter is organized as follows. In Section 3.2, details of the P2P link interface chip architecture are presented. In Section 3.3, salient features of the high-speed interface circuitry design are discussed. In Section 3.4, the design of the digital logic which includes the FIFO memory, FIFO controllers, the aligner and clock circuitry details are discussed. In Section 3.5, details of the layout of the chip are presented. In Section 3.6, packaging of the P2P and PCB test board are summarized. In Section 3.7 measurement results for the P2P chip are presented. In Section 3.8, details of an implemented link testbed interconnecting two PCs using the P2P chip are presented. In Section 3.9, throughput measurements over the implemented testbed are given. In Section 3.10, lessons learned and directions for future work are outlined.

3.2 Point-to-point chip (P2P) architecture

The P2P is designed to validate components key to constructing a CMOS-based link interface chip for multi-Gb/s data rates. It is important because it shows that CMOS can be used to bridge a conventional CMOS electrical interface to a high-performance parallel fiber-optic module. It can also be cascaded with commercial buffers to enable integration into a system.

The P2P has independent transmit and receive ports. The physical layer interface is designed for a 10-wide parallel fiber ribbon, one each for clock, control and eight data lines. The P2P CMOS die, constructed in 0.5 μm CMOS technology via the MOSIS service, measures 10.2 mm x 4.1 mm. A summary of key parameters of the P2P is shown in Table 3.1. The architecture of the P2P is shown in Figure 3.1. The P2P interfaces with a host computer over a TTL interface. Digital logic consisting of FIFO blocks and FIFO controllers bridge the host computer interface with a high-speed analog circuitry which implements the physical layer interface. The host interface transmit and receive buffers, TxFIFO and RxFIFO, are two 1 KB on-chip buffers constructed using dual-ported static random access memory (SRAM). Cells of size up to the maximum of 1 KB are written into the TxFIFO and a special signal line (TxEOP) is used by the host interface to indicate the end of a cell. Cells are transmitted as soon as a complete cell has been loaded onto the on-chip buffer.

An initial version of the P2P [27] was implemented in 0.8 μm HP C26 CMOS technology to achieve 8 Gb/s data rates at operating voltage of 6.2V consuming 3.2A (maximum power consumption of nearly 20 W). The high-speed interface with the

Parameter	Value
Process technology	0.5 micron CMOS (HP-AMOS14TB)
Tox	9.7 nm
Interconnect	3-level Al
Substrate	P-epi with n-well
Supply voltage	3.6V
Max data rate	2.3 Gb/s per signal line
P2P transistor count	200,000
Peak power	5.3 W
P2P die size	10.2 mm x 4.1 mm

Table 3.1: Summary of P2P chip parameters

physical layer uses full-speed clocking where data is clocked once every clock cycle conforming to the PECL signalling format. Following this initial version of the P2P a new chip designed to achieve over 2 Gb/s data rates per signal line in 0.5 μm CMOS technology was implemented. The FIFO memory of this chip exploits the predictable sequential nature of access of a FIFO to achieve the desired speed in a power-efficient manner as will be described in section 3.4 on page 53. Data that is 32 bits wide is read out of the TxFIFO and multiplexed onto eight parallel lines using low-voltage differential signal (LVDS) output signal format by the serializer. The high-speed physical layer interface uses half-speed clocking where data is clocked on both clock edges. The eight data lines, clock and frame can be transmitted onto the network either on optical fiber or connected directly with to coaxial cables with a 50 Ω characteristic impedance. An adjustable clock-delay element for the transmitted clock assists in providing adequate phase margin to latch data at the receiver.

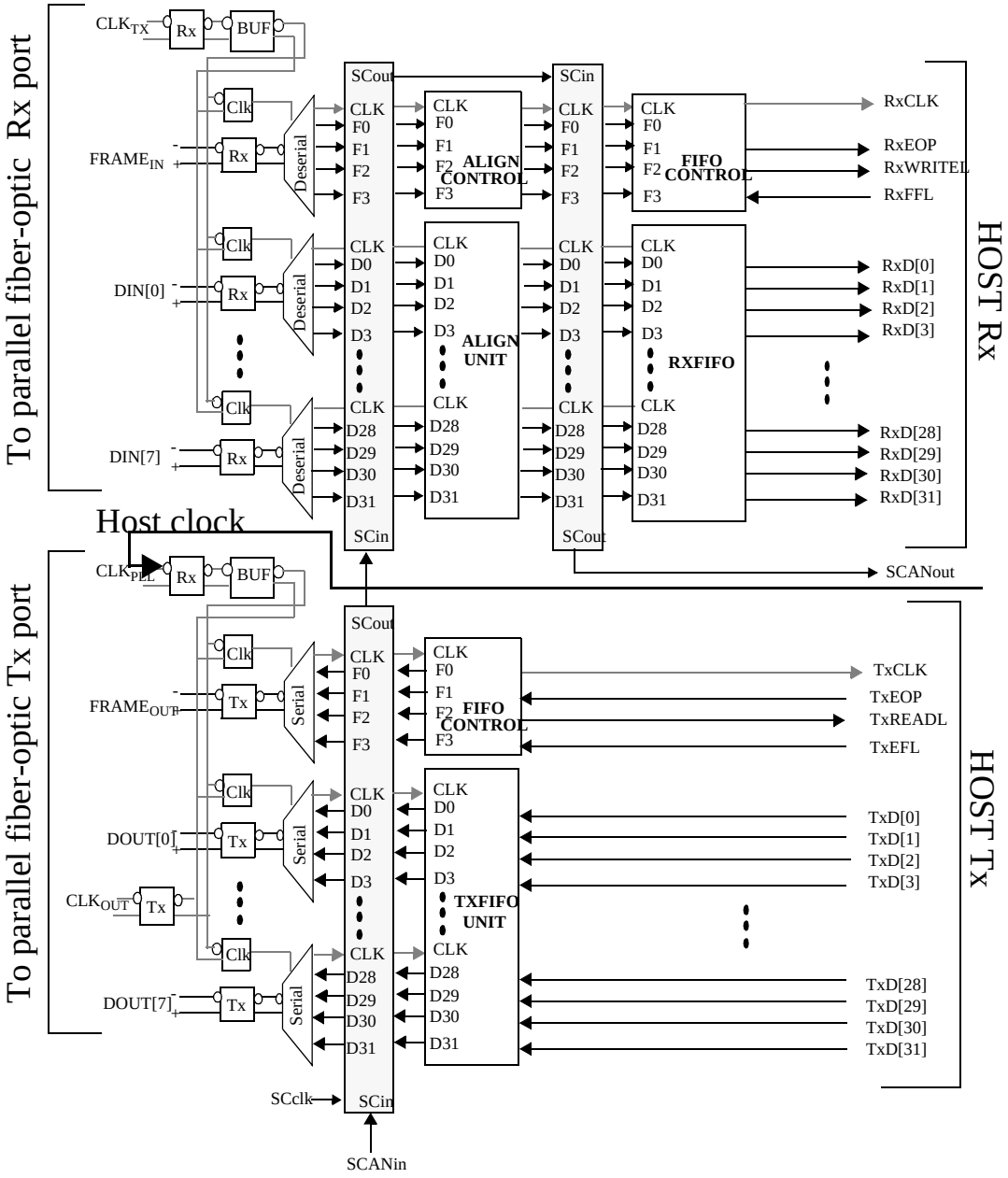


Figure 3.1: Architecture of the P2P link interface chip for a point-to-point link with independent transmit and receive ports. The 32-bit wide transmit FIFO output, received from the host over a TTL interface, is serialized onto an 8-wide LVDS signal format data stream. Clock and control are also additionally transmitted. Received data is deserialized, aligned and passed onto the host from the RxFIFO over a TTL interface.

At the receiver, the eight high-speed LVDS signal parallel lines are demultiplexed to again produce full swing 32-bit wide data. This data is word aligned in hardware by the aligner module. Alignment is required when the demultiplexed signals at the receive port are not phase aligned with the multiplexed signals of the transmit port. Valid data is indicated when the frame control line shows a logic high value. Scan cells are provided on the P2P at the output of the TxFIFO memory block and at the inputs to the aligner and the RxFIFO memory block for testing purposes.

The P2P has a TTL host computer interface with unidirectional independent 32-bit wide data buses (TxDATA and RxDATA) for transmission and reception and some control lines. For the TxFIFO interface, the P2P provides a clock (TxCLK), an internal buffer empty flag line (TxEFL), and a data read enable line (TxREADL) to the external buffers and receives an end-of-packet signal (TxEOP). For the RxFIFO interface, the P2P provides a clock (RxCLK), a data write enable line (RxWRITEL), an end-of-packet signal (RxEOP) and receives an external buffer full flag line (RxFFL). The TTL host interface has a programmable clock speed of divide by 2, 4, 8 or 16 times the digital logic frequency. It is designed to run at a maximum of 50 MHz which will provide sufficient timing margin for its reverse-clocked interface with the external buffers. The slew rate of the TTL pads is adjustable using an external control line. This feature is useful when interfacing to commercial TTL circuits of differing performance.

3.3 High-speed physical layer interface

The high-speed analog transceiver circuitry comprises serializer and deserializer units which are designed to perform the necessary transmit and receive functions at the physical layer interface. The serializer multiplexes 32-bit wide data received from the digital logic onto eight data channels and additionally transmits a clock and control signal line. The deserializer demultiplexes the received data and control stream using the received clock. The entire high-speed design uses a reduced swing differential signal style architecture. The data on the physical layer is of low-swing differential signal (LVDS) format with 500 mV peak-to-peak swing about a common mode voltage of 1.8V for a supply voltage of 3.6V.

The key elements used in constructing the high-speed circuitry are high-speed differential flip-flops, toggle flip-flop dividers, multiplexers, buffers, LVDS receivers and transmitters and an adjustable clock-delay chain. Complete details of the high-speed circuitry may be obtained from [48]. All the high-speed differential circuitry elements use active pulldown level-shifted diode-connected (APLSD) load devices as shown in Figure 3.2. The APLSD cells enable the differential circuitry to operate at higher speeds due to lower parasitics at the output for the same bias currents as conventional load devices in CMOS technology.

A basic differential master-slave flip-flop structure that uses the APLSD load device format is shown in Figure 3.3. Multiplexers and logic functions are created by merging the multiplexing or desired logic function into the master latch of the flip-flop. A toggle flip-flop is created by connecting the negative output of the flip-flop to the positive input terminal of the flip-flop.

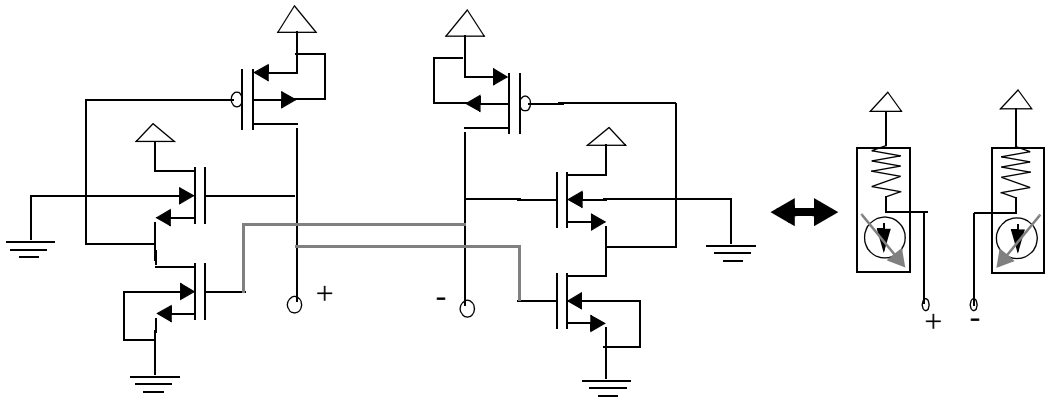


Figure 3.2: Schematic and symbol of Active Pull-down Level-Shift Diode-connected (APLSD) PMOS transistor loads used in differential circuits of high-speed interface circuitry

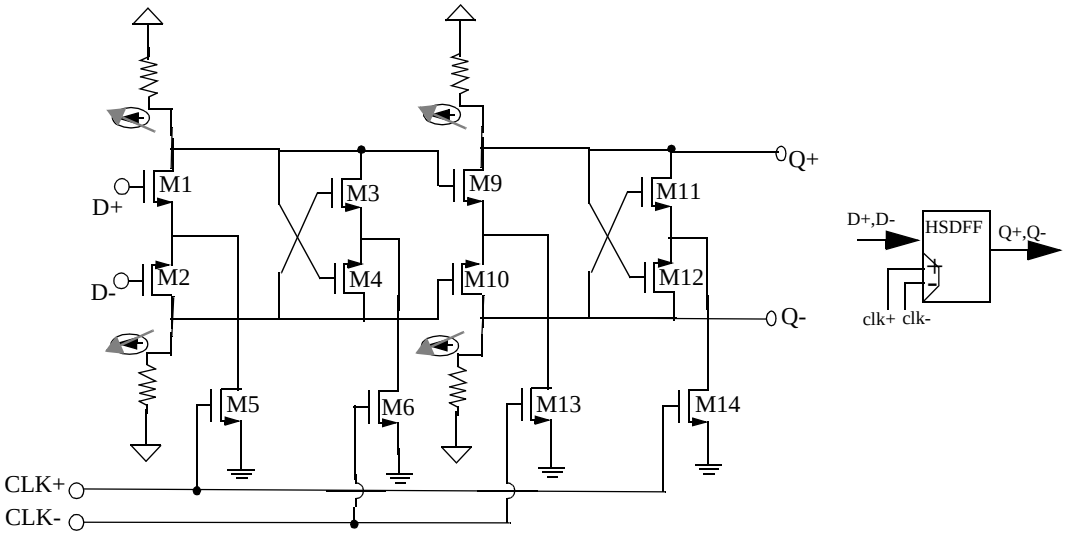


Figure 3.3: Schematic and symbol of high-performance CMOS differential master-slave flip-flop used in high-speed interface circuitry with APLSD load devices shown in Figure 3.2, representative of latching elements, logic gates and clock dividers. Elements such as multiplexers are realized by merging logic functions with the master stage.

3.3.1 High-speed circuitry clocking

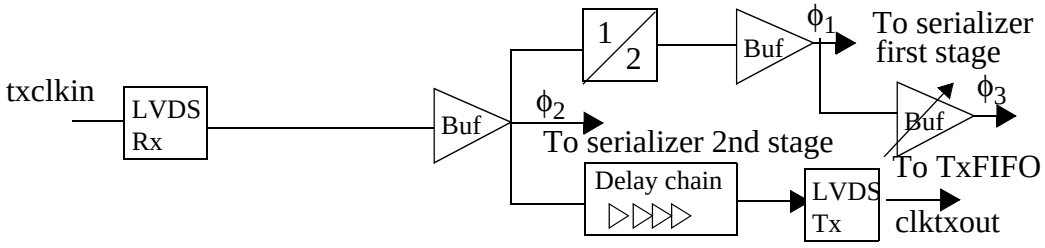
The entire high-speed serializer/deserializer circuitry employs a reduced swing differential clocking style architecture. In the serializer unit, an external clock input is used to generate reduced swing clocks for its internal multiplexing circuitry, the LVDS clock to be transmitted onto the physical layer and the full rail voltage swing clock signals needed by the transmitter digital logic circuitry. In the deserializer, the clock received from the physical layer is used to generate reduced swing clocks for the internal demultiplexing circuitry and the full rail voltage swing clock signals needed by the receiver digital logic circuitry.

A block diagram of the clock distribution circuitry for the transmit high-speed circuitry is shown in Figure 3.4. A high-speed differential clock is received from an external clock source by an LVDS receiver amplifier with the clock input terminated to a termination voltage of VTT (~2 V) that sinks current through a passive 50 Ω polysilicon sheet resistor. The input clock amplitude is assumed to be 400 mV peak-to-peak for maximum bandwidth and performance. The serializer output signal format is chosen to be LVDS in preference to the PECL signalling standard [29] used in the previous P2P [27] because LVDS is better suited for realizing high speeds in CMOS. The reason is that CMOS transistors perform better when the common-mode voltage is half of the supply voltage so that circuit transistors can remain in saturation for the duration of the signal swing. The LVDS receiver amplifier is a cascade of two stages, the first of which is a regulated gain cascode amplifier with the active pulldown level-shifted diode-connected (APLSD) load devices shown in Figure 3.2. The second stage is a differential amplifier stage with controlled peaking to increase the bandwidth of current mirrors by resistive compensation as proposed in [61]. The LVDS receiver produces an output signal ranging in amplitude from 1.3V to 2.5V.

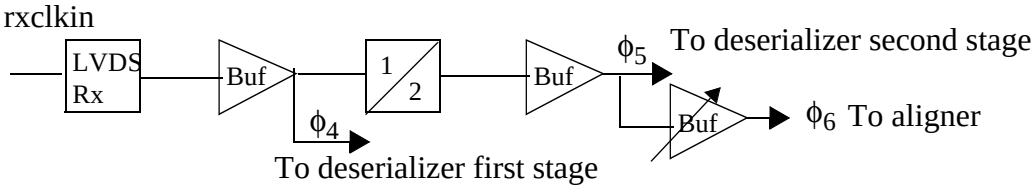
The clocked circuitry in the data channels is designed so that it can operate with low-voltage swings. The LVDS receiver output clock is buffered and a reduced swing differential clock ϕ_2 clocks the second level multiplexers of a serializer unit shown in Figure 3.5. The clock driver is designed to have a low output impedance so as to tolerate large variation in clock loads. The output of the clock driver can then be modeled as a voltage source. The low output impedance clock driver is implemented as a three stage

amplifier. The first stage is a broadband regulated gain cascode amplifier with APLSD load devices. The second stage is a large-swing differential buffer amplifier with a final level-shift stage. The level-shift driver outputs drive the capacitive load of the clock line as a push-pull driver and are insensitive to load capacitances from 0.5 pF to 2.0 pF. The clock driver output swings between 0.8 V to 2.1 V, which was found to be the optimal swing suitable to drive the clocked transistors in differential flip-flops of the serializer-deserializer circuitry. The clock driver drives the clock distribution line from one end. The clock distribution channel was assumed to be a total of 6 pF in gate and load capacitance with a length of 4.3 mm.

The clock ϕ_2 is divided using a toggle flip-flop divider and buffered to produce a half-speed clock ϕ_1 which is used to clock the first stage of the serializer as shown in Figure 3.5. The clock ϕ_1 is converted to full supply rail voltage of 3.6 V and supplied to the TxFIFO digital circuitry. The clock ϕ_2 is also fed to an adjustable clock-delay chain, the output of which is transmitted in LVDS format onto the physical layer.



Tx high-speed clock distribution



Rx high-speed clock distribution

Figure 3.4: Schematic diagram of clock distribution in high-speed transmit and receive circuitry. External PLL clock input is received by the transmitter and the buffered output and its divided version are used to clock the serializer. The LVDS clock input at the receiver is buffered. The buffered output and a divided version clock the deserializer. The divided versions also clock the digital logic in the transmit and receive circuitry.

Each of the LVDS output lines of the transmit circuitry is parallel load terminated at the receiver to a termination voltage of VTT (~2 V) through a passive 50 Ω resistor using polysilicon. Optionally, it could be source terminated to VTT through 50 Ω as well. This would reduce impact of impedance mismatch at the cost of increased drive current for a given signal level.

The adjustable clock-delay element is another key block in the clock path of a high-speed data link. Delay elements control the phase spacing of clock and data at the receive input to the chip. The delay element is composed of differential delay cells with local positive feedback as outlined in [53]. It implements an analog delay line with infinite resolution.

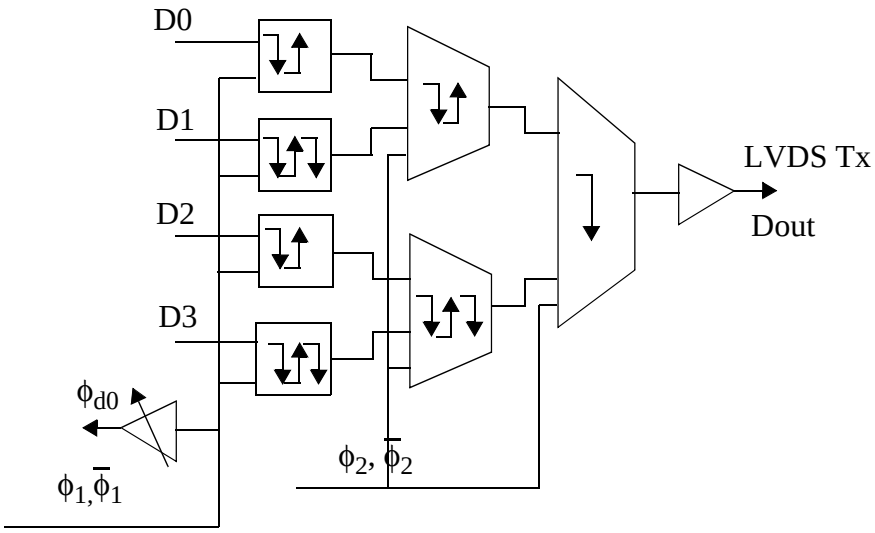


Figure 3.5: Block diagram of one of ten output channels on the transmit circuitry of the serializer. Each serializer channel performs a 4:1 multiplexing operation on the four input bits it receives from the TxFIFO memory output. Output signal format is LVDS with 400 mV peak-peak swing about 1.8 V for supply voltage of 3.6V with maximum skew of 100 ps across ten channels.

At the receive side of the P2P chip, LVDS clock and data inputs are received using LVDS receivers and a demultiplexing operation is performed on the received data as shown in Figure 3.6. The received clock is buffered and the resultant reduced swing (0.8V to 2.1V) differential clock ϕ_4 originating from the center of the deserializer block is used to clock the first stage of the deserializer as shown in Figure 3.6. A half-speed clock is generated from the received clock using a toggle flip-flop divider and the buffered output

clock ϕ_5 drives the second stage of the deserializer as shown in Figure 3.6. The clock ϕ_5 is converted to full rail voltage swing and the resultant output clock is supplied to the receiver digital logic.

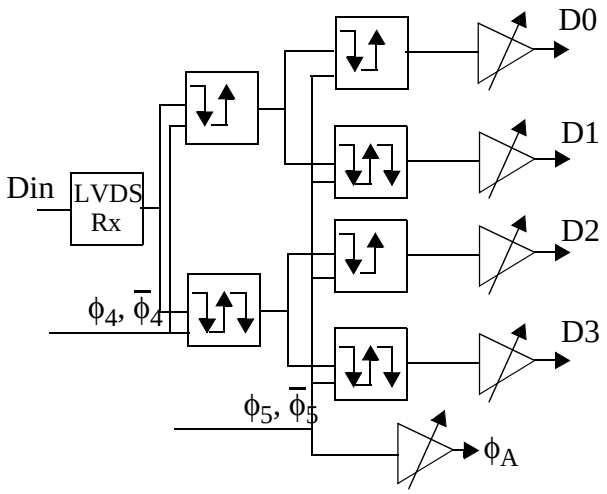


Figure 3.6: Block diagram of one of ten deserializer channels. The deserializer performs a 1:4 demultiplexing operation on the received LVDS signal line in each channel. Clock ϕ_5 is derived from ϕ_4 using a toggle flip-flop divider. Deserializer output feeds the aligner unit.

3.4 Digital logic circuitry

The digital logic comprises of two FIFO blocks - the TxFIFO on the transmit side and the RxFIFO on the receive side, in addition to an aligner block on the receive side. The memory used in either FIFO has a size of 1 KB (to accommodate 256 33-bit words) and is constructed using dual-ported SRAM-based memory organized as 64 rows of four 33-bit wide interleaved banks. The architecture of the memory block is shown in Figure 3.7. The memory allows for independent and separately clocked write and read operations. The TxFIFO has a slow-speed write port running at the TTL interface frequency while its read port communicates with the serializer and runs at half the clock speed of the serializer

clock transmitted onto the network. The RxFIFO has a slow-speed read output port running at the TTL interface frequency while its write input port communicating with the aligner runs at half the speed of the high-speed clock received from the physical layer.

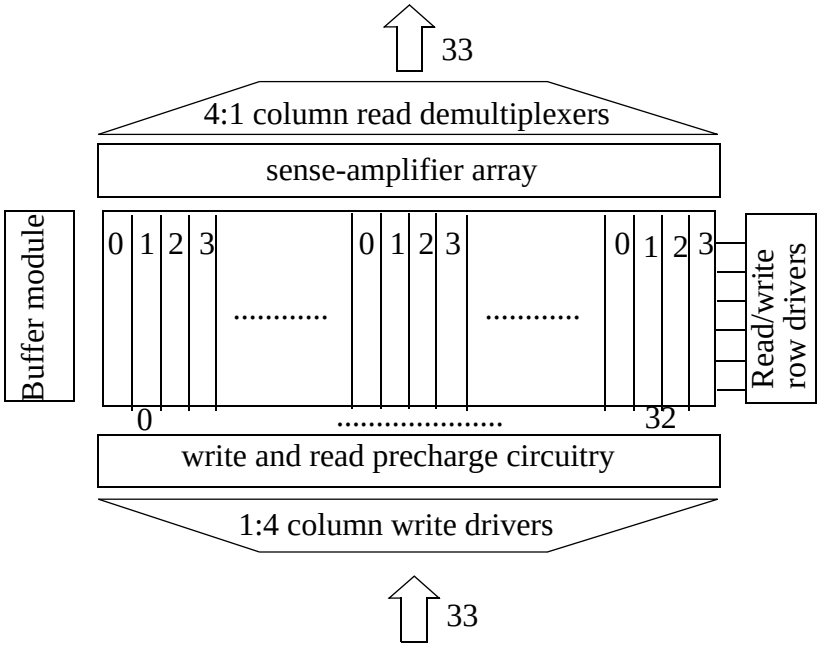


Figure 3.7: Block diagram of the FIFO memory block. The FIFO is a 1056 byte SRAM-based dual-ported memory organized as four 33-bit wide banks of 64 rows. Signals controlling the write and read pointers generated by a FIFO controller are latched in the buffer module in the memory which also generates the clocks used for the digital logic circuitry. A shift-register array generates the row driver lines. Column read and write bitlines are precharged prior to data assertion for speed.

3.4.1 FIFO memory design

The memory cell is a standard eight transistor, cross-coupled inverter circuit topology as shown in Figure 3.8. Bit access to the memory cells is through dynamic write and read ports constructed using complementary bitlines (columns) that are precharged prior to data

assertion for high-speed operation using precharge circuitry shown in Figure 3.9. Bitline access to memory cells for write and read operations is through pass transistors that are enabled by wordlines (row pointers) generated by a shift register array.

Higher operational speeds can be achieved in a FIFO as compared to a RAM because of the predetermined sequential nature of access to the memory cells as opposed to the random access used in a RAM. While a decoder is necessary in a RAM to address a specific memory cell, a simpler shift register chain will suffice in a FIFO. Secondly, a whole clock cycle can be made available for reading from a memory cell in a FIFO as compared to a RAM resulting in larger read bitline swings and better sense-amplifier performance. Thirdly, FIFOs may be designed more power-efficiently by using smaller precharge transistor sizes than would be needed to achieve the same speed in a RAM since by using multiple banks (or column arrays), precharging can take place over multiple clock cycles as opposed to a single clock phase available for a RAM. By precharging write and read bitlines only when the particular bank is not being read from (for the read bitlines) or written into (for the write bitlines) and disabling precharge when the bank is being operated upon, multiple cycles are available for the precharge operation. In the FIFO designed with four banks, there are hence at least three clock cycles for precharging a column bitline as opposed to the single phase in a RAM. Hence, the precharge transistor sizes are substantially smaller resulting in simpler circuitry and lower power consumption.

For a 500 MHz target clock frequency, a FIFO can be implemented with substantially simpler precharge circuit requirements than a RAM as seen from Table 3.2. It also obviates the need for complex precharge enable circuitry such as matched impedance

drivers that would be needed if transmission line effects need to be taken into account. For example, for a 33-bit wide bank, to achieve 500 MHz clock speed operation in a RAM, the precharge transistor sizes are as shown in Table 3.2 resulting in a precharge enable load of 6.9 pF (for a gate oxide capacitance of 3.9 fF per square μm). For a precharge metal interconnect length of 2500 μm, at metal width of 1.2 μm and metal resistance of 0.07 Ω per square, the resistance is nearly 150 Ω resulting in a wire RC delay-product alone of more than 1 ns which is the phase time of a 500 MHz clock signal. Hence, transmission line effects have to be taken into account in designing a precharge enable driver to overcome frequency-limiting reflections from impedance mismatches.

Further, resistive drops on the line would reduce the signal amplitude and prevent proper functioning of the enable lines (approximately 0.6V for a 0.6 ns transition time and 3.6V swing). To reduce resistive drops, precharge enable lines would have to be excessively wide to lower the metal resistance and the drivers would need to be large to provide low output resistance. Besides the added complication of a higher capacitive load, power consumption will still be prohibitively high. Hence this is not a practical solution for the target digital clock frequency of 500 MHz. On the other hand, the precharge load for a precharge enable line for a 33-bit wide bank in a FIFO, as seen from Table 3.2, is only 650 fF which at a tenth of the load of that in a RAM design can be implemented using simple buffered drivers.

The FIFO controllers generate the signals that determine the wordline and bank being accessed for a write or read operation in the FIFO memory as well as the precharge controls. Control signals are latched within the memory block instead of being directly driven from the controller to provide the necessary timing margin for 500 MHz operation.

Write operation: Write operations are performed to the four different banks in succession before the row pointer is advanced. Precharging is enabled for a bank that is not being written into and disabled for the bank being written into. Hence, each bank is precharged for at least three clock cycles. The precharge signals and the shift register control signals from the FIFO controller are retimed in the memory block using latches to provide the delay margin for the targeted speed.

Read operation: Read operations are performed to the four different banks in succession before the row pointer is advanced. Precharging is enabled for a bank that is not being read from and disabled for the bank being read from. Hence, each bank is precharged for at least three clock cycles. The complementary read bitline values are evaluated using a sense-amplifier. The sense-amplifier clock is positioned to maximize the bitline voltage amplitude at the instance of evaluation. Latches are used at the output of the memory to resynchronize data with the FIFO controller clock.

The circuit schematic for the sense-amplifier is shown in Figure 3.10. The operation of the sense-amplifier is as follows. When ϕ_1 is held low, the tail current source is enabled, the bitline values din and dinbar assert the values of the arms of the cross-coupled pair, bit and bitbar. While ϕ_1 could be enabled only prior to evaluation of the sense-amplifier for power-consideration reasons, in our implementation it is always held low.

When ϕ_2 is held high, the voltages on the cross-coupled pair arms bit and bitbar are equalized. If din and dinbar are simultaneously asserting their values, there is a net imbalance as dictated by the complementary values of din and dinbar. When ϕ_2 goes low and simultaneously ϕ_3 turns high, bit and bitbar evaluate to either rail using the cross-coupled inverter pair. The resultant amplitude is latched and made available at outputs Q and Qbar. The sense-amplifier sensitivity is degraded by any process-induced variations in the two arms of the cross-coupled pair such as those due to threshold voltage mismatches. In simulations, the sense-amplifier is sensitive to 20 mV input bitline swing under perfectly matched conditions at 85^0C junction temperature using slow libraries.

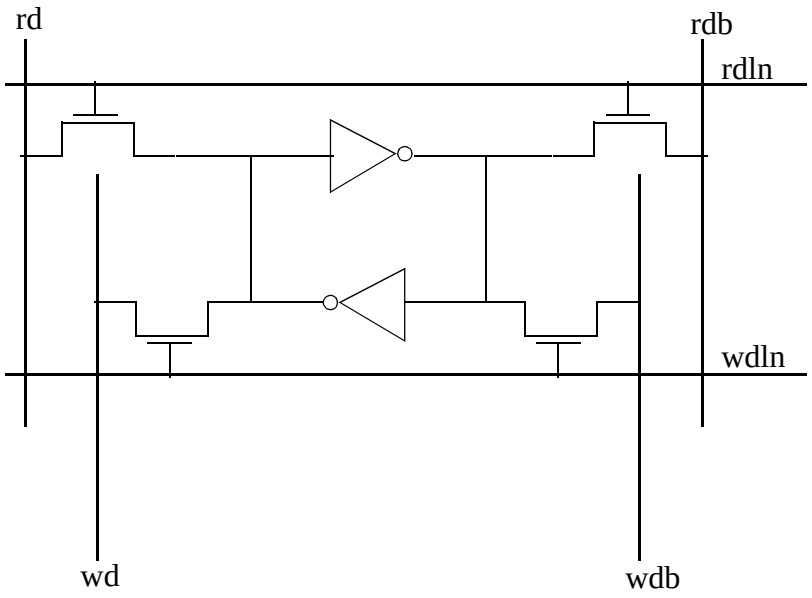


Figure 3.8: Circuit diagram of the eight-transistor cross-coupled inverter SRAM memory cell. Pass transistors provide access to complementary bitline columns (rd, rdb and wd, wdb) and access is controlled by row enable lines wdln and rdln.

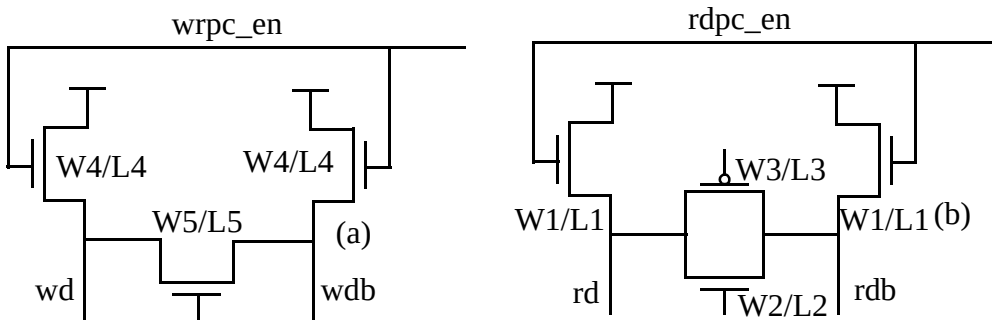


Figure 3.9: Circuit diagram of precharge circuitry for (a) write bitlines and (b) read bitlines

Transistor	RAM style	FIFO style
W1/L1	38.4/0.6	2.4/0.6
W2/L2	12.0/0.6	3.6/0.6
W3/L3	18.0/0.6	12.0/0.6
W4/L4	9.6/0.6	2.4/0.6
W5/L5	3.0/0.6	0.9/0.6

Table 3.2: Precharge transistor sizes for 500 MHz memory operation

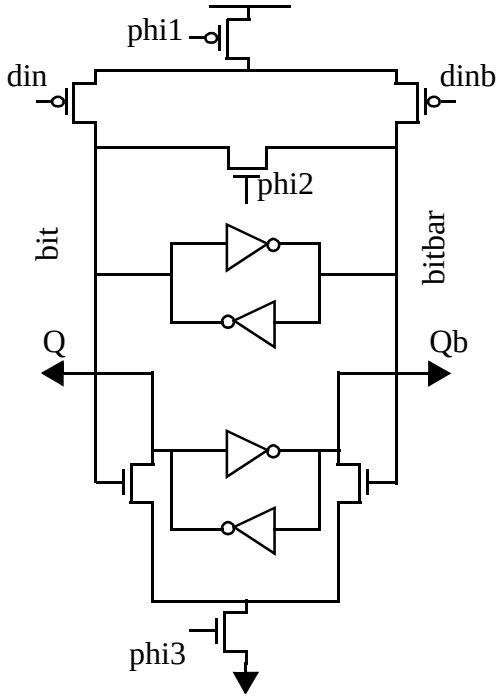


Figure 3.10: Circuit diagram of a cross-coupled inverter pair sense amplifier. Bitline voltages din and dinb are the inputs to the sense-amplifier. The differential input at the falling transition of clock phi2 is latched and available at outputs Q and Qb. phi3 is 180^0 out of phase with phi2.

3.4.2 FIFO Controller design

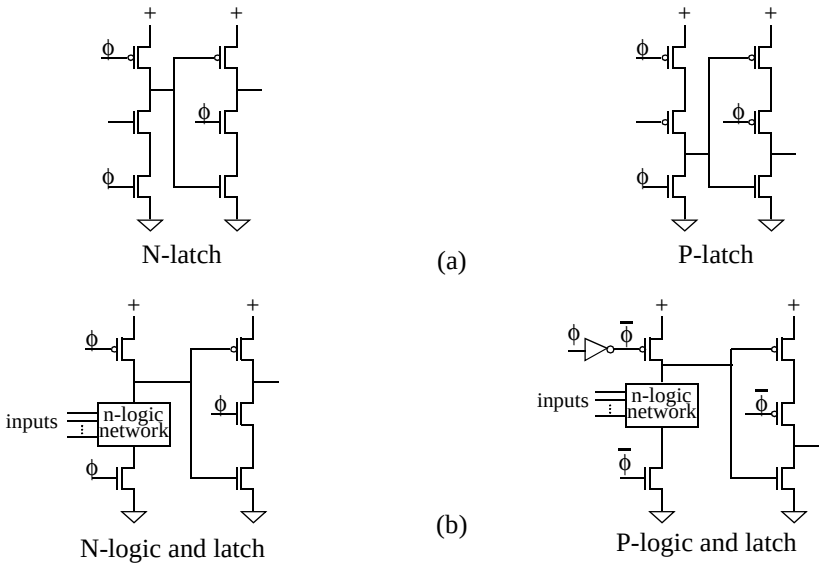


Figure 3.11: Circuit diagram of dynamic TSPC style latches and logic used for digital logic controller. P-logic cells are realized using n-logic cell with clock input locally inverted for power-efficient speed optimization. Data to logic gates has to be set up before the active phase and stay stable during evaluation. Data to latches can however change during the active phase so long as it meets setup constraints of the latch. This property is useful for retiming across clock interfaces.

The FIFO controllers generate signals that control the memory cell being accessed for a write or read operation and also generate the precharge control circuitry signals. They also implement logic necessary for handshaking over a TTL interface with an external host computer. The FIFO controllers are implemented using dynamic true single phase clocking (TSPC) logic style [64] to achieve a 500 MHz digital logic frequency. Figure 3.11 shows the latches and logic cells used to implement the controller logic. A p-logic

cell is implemented from an n-logic cell by locally inverting the clock input signal. The logic gates are used to implement fast counters, comparators and other random logic. The controller is deeply pipelined with single gate delay per pipeline stage.

A standard cell-based approach is used for the implementation of the controller. Clock delay on each of the rows of the controller is matched and a reverse clocking interface is used between the controller and the FIFO memory. Place-and-route of the controller block and all top-level routing was performed using the Cadence [63] Cell Ensemble tool and Cadence Dracula tool was used for layout verification. Extracted parasitics from layout were used in simulation using HSpice [62].

3.4.3 Aligner design

At the receiver, the eight high-speed parallel lines are demultiplexed to again produce 32-bit wide data. This data is word aligned in hardware by the aligner module shown in Figure 3.12. Alignment is required when the demultiplexed signals at the receive port are not phase aligned with the multiplexed signals of the transmit port. The serializer of the transmitting node serializes the data supplied to it in groups of 4 bits. The data then travels down the medium and is deserialized at the receiving end which again samples the received data in groups of 4 bits. The received sample group may be shifted by 0, 1, 2 or 3 bit periods with respect to the actual starting bit of a sample group that was transmitted by the serializer. Hence, the received bits have to be shuffled to reproduce the original group at the transmitting end. The aligner performs this function using the received frame valid signal as a control line for the grouped received bits to be aligned with. The alignment is performed using nine 4-to-1 multiplexers, one for each data channel and an additional

frame channel. The four inputs to the multiplexers are four adjacent bits of the input stream to the deserializer. By locating the start of the frame control line, the aligner decoder generates the mux-select line that is used to select the appropriate bit and produce groups of four bits that are aligned to the frame control line and hence correspond to the original group of the transmitted data.

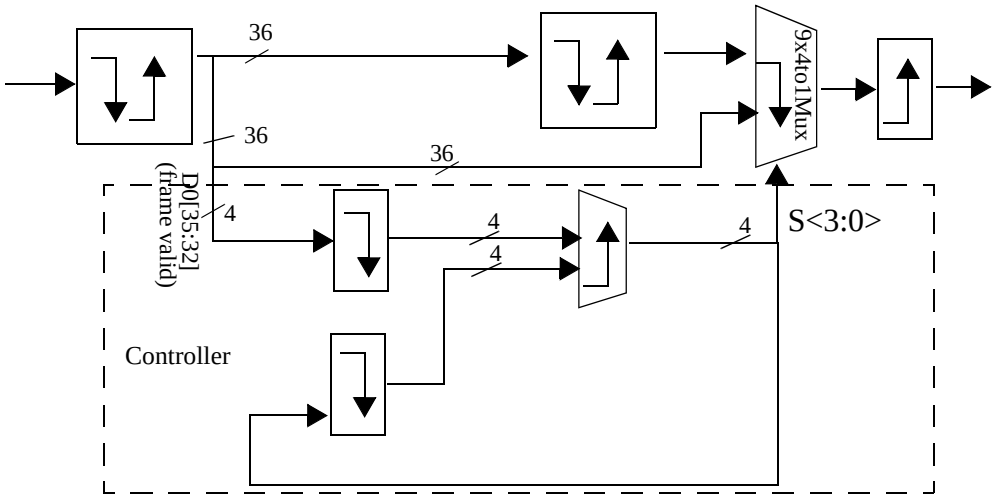


Figure 3.12: Block diagram of aligner composed of nine 4:1 multiplexers and a decoder that generates the multiplexer select lines to perform the desired bit reshuffling for alignment. Aligner receives input from deserializer and produces 32-bit wide data output, an end-of-packet bit (EOP) and a frame control bit. Output of aligner is connected to the RxFIFO. The aligner controller generates the 4-bit select code for shuffling by 0, 1, 2 or 3 bits.

3.4.4 Digital logic circuitry clocking

For the digital logic portion, there are three clocking modes - forward clock distribution, reverse clock distribution and matched-delay clocking [64]. In a forward clock distribution strategy, clock and data flow in the same direction and data that is clocked out using an n-latch (p-latch) is retimed using a delayed clock using a second n-latch (p-latch) where delay on clock is more than delay on the data line. In a reverse clock

distribution strategy, clock and data flow in opposite directions and data that is clocked out using an n-latch (p-latch) is retimed by a p-latch (n-latch) using an earlier clock. Under matched-delay clocking, data that is clocked out using an n-latch (p-latch) is retimed by a p-latch (n-latch) using a matched-delay clock

A forward clocking mechanism is used to retime the 36-wide data coming out of the nine channel deserializer into the aligner and from the aligner into the RxFIFO memory and RxFIFO controller and from the RxFIFO to external buffers over the TTL interface. A reverse clocking mechanism is used for the control signals from the RxFIFO controller to the RxFIFO memory. The control signals that control the read and write pointers as well as the precharge enable lines for the RxFIFO are retimed in the memory block to achieve the 500 MHz target frequency. Clock delays for each of the rows in the controller block are matched.

A reverse clocking mechanism is used to retime the data flowing into the transmit high-speed blocks from the memory, for the interface from the TxFIFO controller to memory or from memory output to the controller and for the TTL interface to external buffers. As in the RxFIFO, control signals from the TxFIFO controller are retimed in the memory block to maximize performance. Clock delays for each of the rows in the controller block are matched.

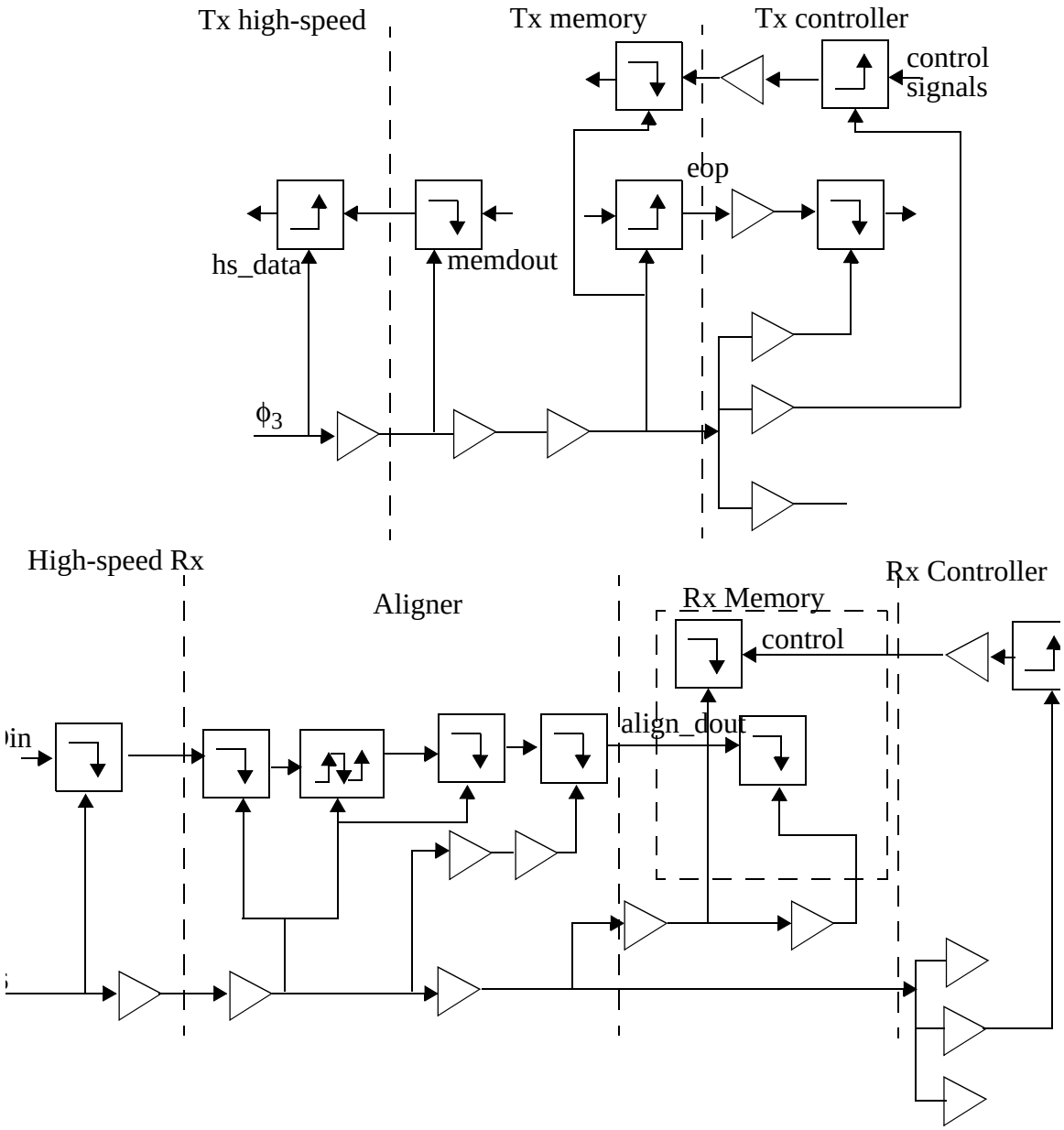


Figure 3.13: Clock distribution for the read circuitry of transmit and write circuitry of receive digital logic. On the transmit side, clock and data move in opposite directions (reverse clocking) at the high-speed to FIFO interface, and between the FIFO memory and controller. On the receive side, clock and data move in the same direction (forward clocking) at the high-speed to aligner interface, the aligner to RxFIFO interface. They move in opposite directions (reverse clocking) between the FIFO memory and controller. Row clock delays are matched in either controller.

3.5 Layout of Chip

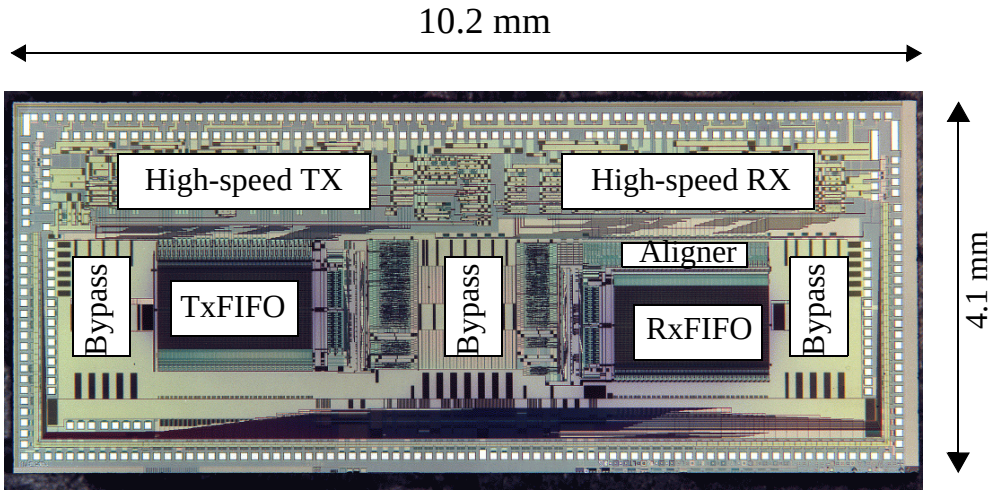


Figure 3.14: Photograph of the P2P die with size 10.2 mm x 4.1 mm fabricated in 0.5 µm 3-layer metal HP CMOS technology. The chip was submitted on 5/31/1999 and the fabricated die was received on 8/26/1999. LVDS signals at the top of the chip are transmitted onto the physical layer while TTL pads at the bottom provide the host computer interface. Power and ground for analog and digital circuitry are isolated and bypass capacitors are provided on chip. Substrate contacts and guard rings provide noise isolation for sensitive circuitry.

The total die size of the P2P shown in Figure 3.14 is 10.2 mm x 4.1 mm implemented in 0.5 µm CMOS process provided by MOSIS using the HP-AMOS14TB process. This is an nwell process with 3 metal and one poly layer and operates with a nominal 3.6 volt power supply. The size of the transmit and receive digital blocks are roughly 3 mm x 1.7 mm. The size of the transmit high-speed block is roughly 5 mm x 0.8 mm while that of the receive block is roughly 4.3 mm x 0.8 mm. The remaining chip area is occupied by bypass capacitors, metal routing and pads. The number of transistors on the chip is close to 200,000. Separate power connections are provided for the high-speed circuitry, the

digital circuitry clock buffers and the rest of the digital circuitry. The ground connections for the digital circuitry and the high-speed circuitry are also separated. The substrate connection for the digital circuitry clock buffers is isolated from the ground connection for the rest of the digital circuitry. Substrate contacts and guard rings are placed to protect circuitry from substrate-coupled noise. In the high-speed circuitry, each of the ten channels has a distinct power and ground connection to avoid power supply degradation due to resistive drops. A standard cell-based approach is used for the digital logic circuitry. A custom layout of the high-speed interface circuitry and the FIFO memory blocks was performed. Place-and-route of the transmit and receive controllers and all top-level routing was performed using the Cadence Cell Ensemble [63] tool and the extracted layout was simulated using HSpice [62] at a junction temperature of 80^0 C.

Parameter	Value
Process technology	0.5 μm CMOS (HP-AMOS14TB)
Tox	9.7 nm
Interconnect	3-level Aluminum
Substrate	P-epi with n-well
Supply voltage	3.6 V (nominal)
Max data rate	2.3 Gb/s per signal line
Transistor count	~ 270,000
Peak power	5.5 watts
Die size	10.2 mm x 4.1 mm

Table 3.3: Summary of P2P chip features

3.6 Packaging

For testing purposes, the die is housed in a ceramic quad flat package (QFP) with 244 pins and differential 50 Ω controlled signal impedance lines for the high-speed physical layer interface signals. The package provides separate power planes needed for isolation of digital VDD, analog VDD and TTL VDD. The QFP was designed at USC and manufactured by Kyocera, Inc. [65]. The package is mounted on an FR-4 printed circuit board (PCB) specifically designed for high-speed electrical testing. High-speed signals are deskewed on the PCB to within ±20 ps inclusive of corrections in trace lengths to account for the measured QFP skew of ±40 ps.

3.7 Measurement of P2P IC

Measurements of the P2P verified operation up to data rates of 2.3 Gb/s per data line as shown in Figure 3.15. At higher speeds, inadequate bitline swing in memory prevents proper operation. Maximum power consumed is less than 5.5 W. Power consumption of analog circuitry is essentially unchanged with clock frequency because of the large contribution from biasing current drawn by the differential circuitry. However digital circuitry consumes less power at lower frequencies as shown in Figure 3.17. The skew on the high-speed data output lines is less than 125 ps inclusive of a 20 ps skew on the board.

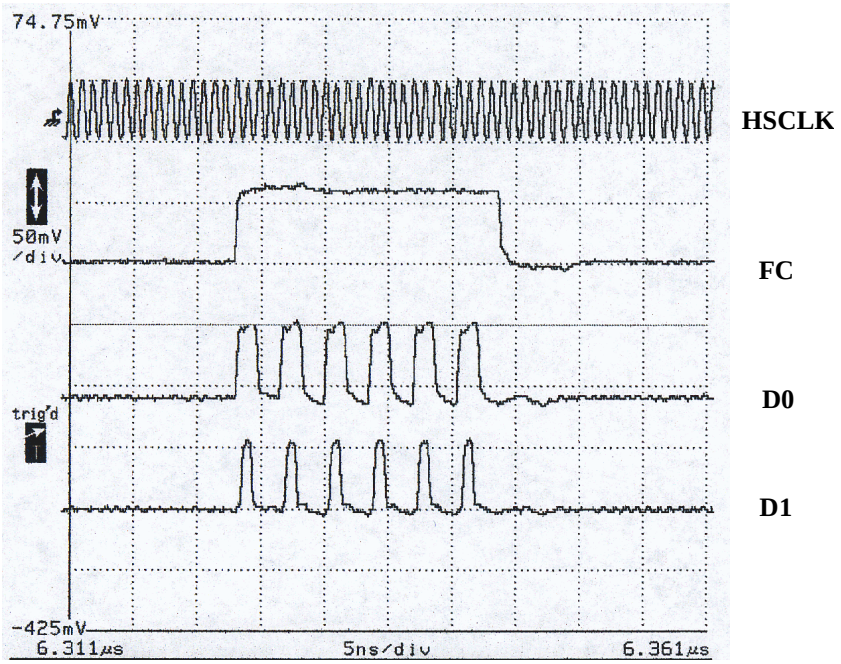


Figure 3.15: Measured transmit port output at a data rate of 2.3 Gb/s per data channel for an aggregate data rate of 18.4 Gb/s. The figure shows high-speed clock (HSCLK), frame control (FC), data channel 0 (D0), and data channel 1 (D1). The X-axis shows time at 5 ns/division while the y-axis shows output peak-peak signal amplitude at 500 mV/division (with power attenuation of 20 dB). Total power consumption is 5.3W at 575 MHz digital logic frequency.

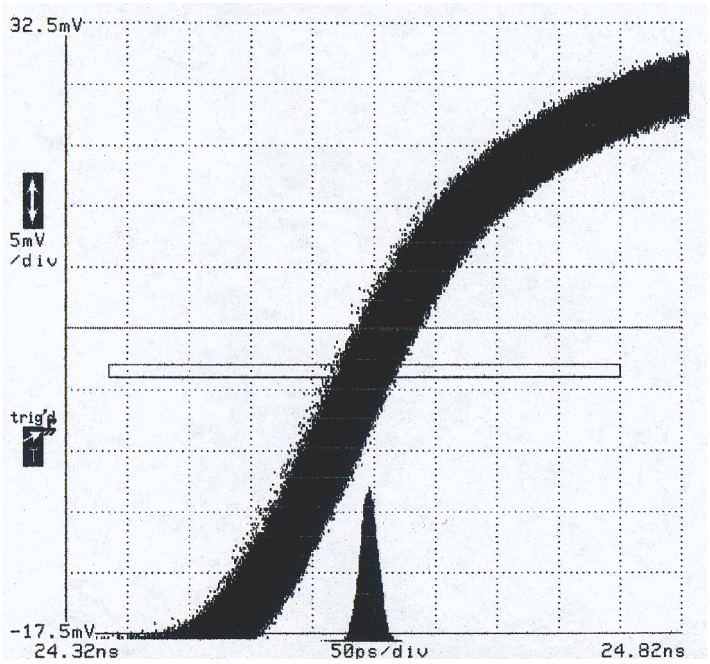


Figure 3.16: Frame-referenced jitter on high-speed clock output at clock frequency of 1.15 GHz. RMS jitter is less than 8 ps, while peak-peak jitter is 65 ps. Peak-to-peak jitter over long time cycles is still well within the data phase time of not more than 400 ps. Hence, clock jitter will not impact reliable data transmission at these data rates.

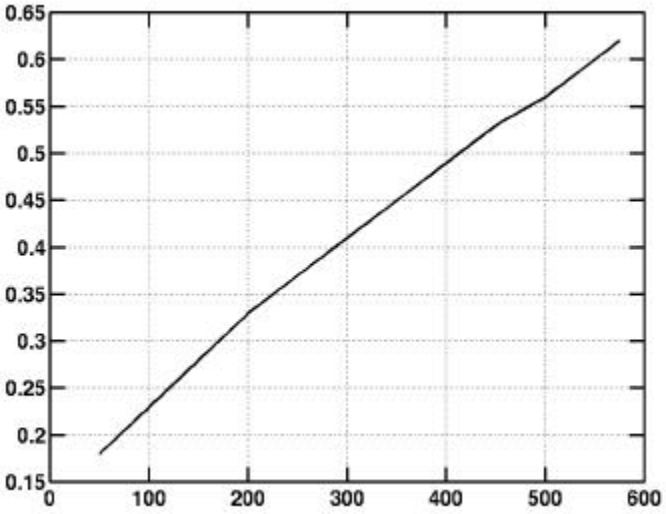


Figure 3.17: Digital logic current consumed at 3.6V operation with variation in frequency for a TTL clock interface running at a divide-by-16 of the digital logic frequency. The figure shows that digital logic power scales linearly with operating frequency. Maximum power consumed in the digital logic is 2.25 W at 575 MHz. Analog power is nearly constant with operating frequency with a power consumption of nearly 3 W.

3.8 Experimental link testbed

In this section, we describe a prototype experimental testbed that implements a point-to-point link connecting two PCs. We demonstrate the potential for constructing CMOS-based interfaces with parallel fiber-optics. The testbed also enables us to develop multimedia applications and obtain initial system performance measurements. The architecture and photograph of the link testbed are shown in Figure 3.18.

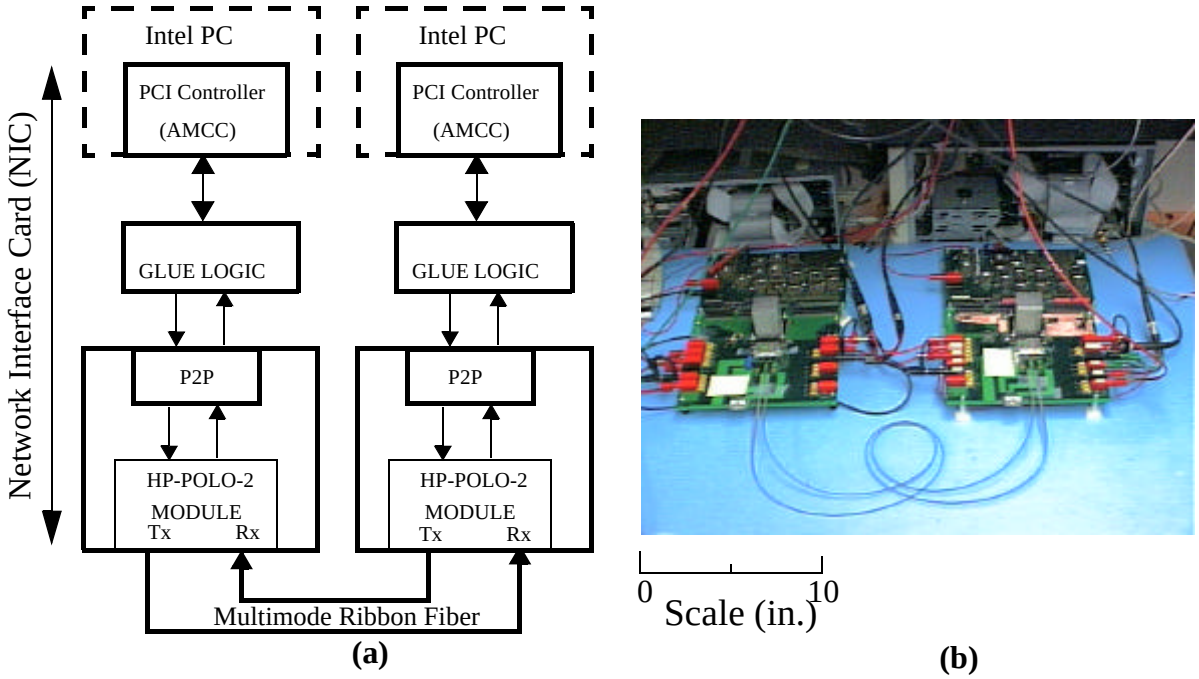


Figure 3.18: (a) Illustrates the experimental arrangement for a point-to-point PCI-based link. (b) Photograph of the experimental arrangement. In the foreground is the looped multimode fiber-ribbon which connects two HP-POLO-2 transceivers. Intel Pentium-based PCs in the background constitute the host computers. A commercially available PCI controller (AMCC) is installed inside the PC and connected to an external glue logic board using an electrical ribbon connector.

The initial link testbed interconnects two PCs over a point-to-point link using a network interface card (NIC) that was constructed at USC for this project. The schematic of the NIC is shown in Figure 3.18(a). The NIC was implemented using a combination of novel specialized hardware solutions, commercially available components, and experimental components. Intel Pentium-based PCs running the Windows NT 4.0 operating system constitute the host. A commercially available AMCC Matchmaker

developer's board [66] installed in the host provides a 32-bit 33 MHz half-duplex interface to the internal PCI [67] bus of the PC. A commercially available device driver [68] was modified at USC to perform data transfers over the AMCC board.

A photograph of the custom-designed portion of the NIC is shown in Figure 3.19. The board on the left is called the glue logic board. The glue logic board performs signaling functions necessary to bridge the AMCC board with the P2P. The glue logic and P2P are located on separate boards for ease of testing in our experimental implementation. By handling the glue logic functions external to the P2P, the use of the P2P is not restricted to any particular I/O bus interface. By constructing a new glue logic board, the same P2P can be connected with any other existing or emerging bus standards such as Infiniband [9]. The glue logic board consists of independent 4 KB buffers on its transmit and receive directions. These buffers temporarily store and synchronize data transfers between the PCI controller and the P2P. The glue logic interface with P2P is a full-duplex 32-bit interface designed to operate at a maximum frequency of 50 MHz providing a peak bisection bandwidth of 3.2 Gb/s. The glue logic interface with the PCI controller is a 32-bit half-duplex interface at a clock frequency of 33 MHz.

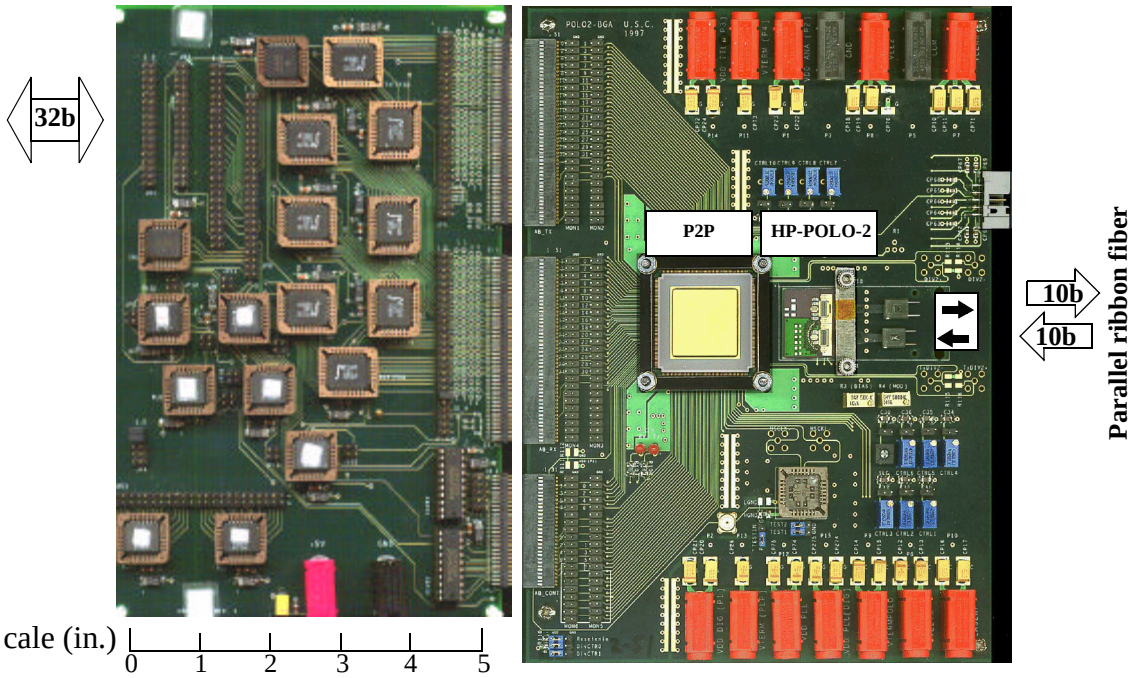


Figure 3.19: Glue logic board (left) with FIFOs for data storage and PLDs for control logic. The P2P chip that was designed for point-to-point interconnections (which will later be replaced by the LAC in a ring network currently being designed) and HP-POLO-2 parts are mounted on the board shown on the right. The fiber connector I/Os are indicated by the two solid black arrows on the right and demonstrate the high edge-connection density offered by optics.

The board on the right in Figure 3.19 has two main components - the P2P and a HP-POLO-2 optoelectronic module. The HP-POLO-2 module is an 850 nm VCSEL/PIN detector array-based fiber transmit/receive interface [39] designed by Hewlett-Packard (HP) Research Laboratories for parallel fiber-optic links. Each POLO-2 module consists of 10 transmitters and 10 receivers in a compact Ball Grid Array (BGA) package. A newer version of the HP optoelectronic module called the PONI module (shown in Figure

1.7 in Chapter 1) has separate 12-wide transmitters and receivers. The modules can handle data rates of 2.5 Gb/s per multimode fiber. More details on the PONI modules can be found in [41].

3.9 PCI throughput measurements

We performed some measurements to determine the maximum sustainable send throughput over the AMCC PCI bus interface board using a file-based I/O transfer over Windows NT 4.0. Theoretically, the highest possible send throughput achievable through a PCI bus operating at 33 MHz is 1.06 Gb/s. However, in a real system such as a typical Pentium-based PC, this throughput value cannot be sustained. We performed some throughput measurements on the PCI bus of a Triton (82430 VX PCI chipset) motherboard in a 166 MHz Intel Pentium-based PC using Windows NT 4.0. The sustained send throughput as shown in Figure 3.20 saturates at a value of 163 Mb/s, while that obtained using a file-based I/O transfer over DOS resulted in a sustained throughput of 300 Mb/s. Throughput degradation is due to the scatter/gather operations on physically discontiguous memory data being performed in software, since the current AMCC card does not have hardware scatter/gather capabilities. The send throughput obtained using actual applications will be further limited by protocol and operating system overheads [71]-[73] and is beyond the scope of discussion here.

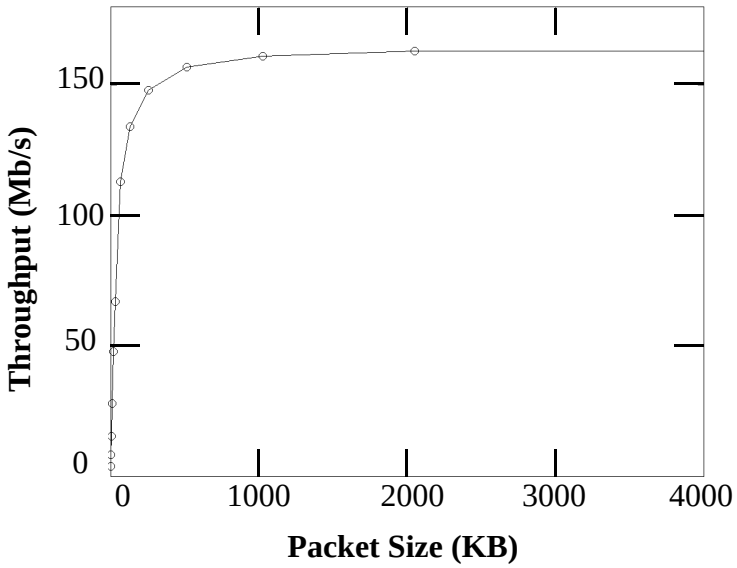


Figure 3.20: Measured sustained send throughput using file-based I/O for a 166 MHz Intel Pentium-based Triton motherboard with 82430 VX PCI chipset (33 MHz) motherboard using Windows NT 4.0 operating system. Due to file I/O transfer overheads, the throughput saturates at 163 Mb/s.

3.10 Summary

In this chapter, the implementation of a CMOS chip with measured data rates of 2.3 Gb/s per data channel in 0.5 μm CMOS for a point-to-point link was discussed. This chip experimentally demonstrates the feasibility of direct CMOS interfaces to fiber-optics. It was designed as a precursor the link adapter chip for the ring network discussed in Chapter 5. All the components used in the P2P have been used in a network interface chip designed to implement a broadband ring network using a parallel fiber-optics physical medium. However, as will be explained in that chapter, it also shows that a zero-skew global clock distribution scheme in the ring network will yield higher data rates as the controller standard cell library cells and the memory are both functional at up to 575 MHz

digital logic clock speed. Limitations in achievable clock speed arise from the sense-amplifier in the memory. Host interface limitations arise from a full-rail swing TTL signal level-compatible interface.

The scaling to higher data rates and smaller transistor sizes is discussed in Chapter 6. In the following chapter, the network protocol used to implement a parallel fiber-optic slotted-ring network known as the PONI network will be described.

3.11 Acknowledgments

We gratefully acknowledge Bindu Madhavan for providing the serializer, deserializer, standard cell and sense-amplifier circuits, Young-Gook Kim and Tsu-Yau Chuang for the design of the printed circuit test board and measurements of the PCI link implementation.

Chapter 4

PONI ring network

4.1 Introduction

The point-to-point link described in Chapter 3 was a testbed implemented to verify the feasibility of constructing high-performance CMOS-based interfaces to parallel fiber-optics. The experience gained from designing this link was leveraged to construct an NIC for a network of PCs based on the slotted ring architecture. As outlined in Chapter 1, the slotted-ring network architecture was chosen due to its potential for broadband implementation as well as cost-effectiveness arising from its simplicity relative to alternative schemes such as those based on centralized switch-based and shared bus-based networks.

In subsequent sections of this chapter, we describe the architecture and implementation of an experimental low-cost GB/s slotted ring network called the PONI network. The network designed is to our knowledge the fastest ring network reported to date with a network aggregate data rate in excess of 16 Gb/s achieved entirely using conventional CMOS technology. Comparisons with other ring networks existing currently are described in the following chapter.

Some of the goals guiding the network design are:

- Construction of the entire medium access control (MAC), support logic and high-speed circuitry using a single CMOS link adapter chip (LAC) to provide a link data rate in excess of a Gb/s/signal line
- Provide fair access to hosts along with bandwidth guarantees for multimedia applications
- Low node latency and low delay variation (jitter)
- Support for broadcast and multicast modes
- Interoperability with ATM networks in the B-ISDN domain
- Interface to PCs through a high-performance PCI bus interface

4.2 PONI Network Architecture

The PONI network is a low latency, ultra-high bandwidth, unidirectional slotted ring for use in a scalable cluster of host computers which could be PCs or workstations (see Figure 4.1). The physical medium is a 10-wide multimode fiber-ribbon. The physical layer portion of the PONI network is optimized for parallel fiber-optic link technologies with an 8 Gb/s or greater data transfer rate. The end-systems in a typical network are envisioned to be, though not restricted to, low-cost PCs.

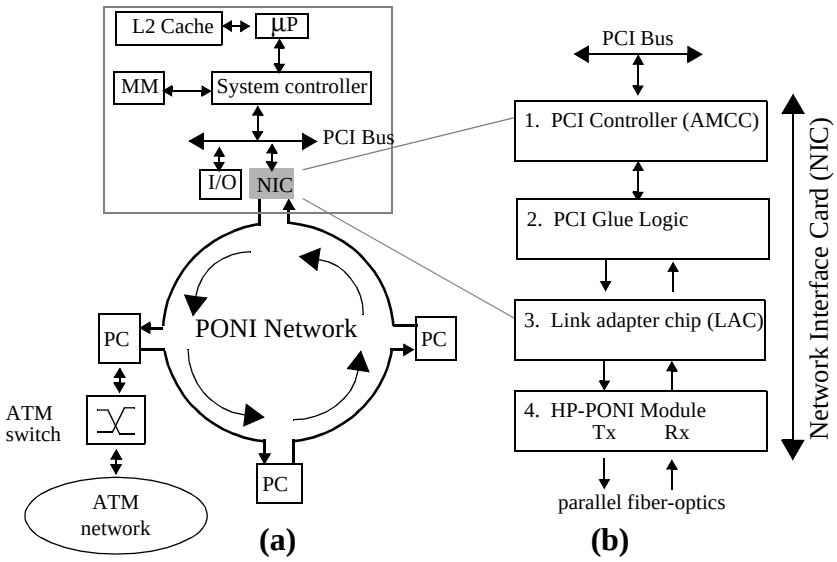


Figure 4.1: (a) Schematic of unidirectional PONI ring network showing interconnected host PCs (b) Components of the NIC in a host PC that interfaces between the PCI bus and the parallel multimode fiber-ribbon network medium. A commercially available PCI controller interfaces with the host PCI bus. The link adapter chip (LAC) will implement the medium access control (MAC) and our glue logic design bridges PCI and LAC. The HP-PONI module is an experimental fiber transceiver designed by Hewlett-Packard Research Laboratories.

The PONI network is cost-scalable (same incremental cost with each additional node) as it will be formed from identical nodes. The terms “host” and “node” are used interchangeably in this chapter and may be considered equivalent. The address field in the prototype PONI network packet for transmitting sources is 5 bits long and hence the network is scalable currently to a maximum of 32 nodes each capable of multi-Gb/s throughput access to the ring. It would be fairly straightforward to increase the allowed number of network nodes. The interface hardware of the network is designed to provide ease of connection to the PCI bus standard. An additional 5-byte header provides the

capability of using virtual channel identifiers (VCI) needed by other WAN networks such as ATM networks. Figure 4.1(a) shows the topology of the PONI network and Figure 4.1(b) shows the various components that constitute the PONI network interface. In a typical network, high-performance PCs constitute the host computers. These PCs are connected to the high-bandwidth network using the NIC. The maximum total network physical layer length in a 32-node network is 3.2 km (due to the link length limitation of 100 m of the HP-PONI module), corresponding to a total fiber latency of less than 20 μs.

The custom-designed link adapter chip (LAC) implements a MAC protocol. The LAC replaces the earlier described P2P on our network interface card shown in Figure 3.19 on page 74. The PONI network has been designed to support the high bandwidth and delay sensitive requirements of multimedia traffic such as uncompressed live video and voice, while at the same time providing adequate support for other traditional data traffic such as file transfer and e-mail. The MAC protocol used in the PONI network is discussed in Section 4.3.

4.3 PONI network protocol

The following section describes the MAC for a slotted ring network implemented on the LAC. The physical layer of the PONI slotted ring network consists of 10 signal lines - eight parallel Gb/s data lines for the serialized data streams resulting from a 32-bit host computer interface, one line for the clock and one line for a frame control signal. The clock is transmitted along with the control and data lines, thereby avoiding the need for clock extraction which is a necessity in conventional serial links. The clocking mechanism in the PONI network is a distributed scheme that is based on the one used in

FDDI [22]. Slot boundaries are indicated using the frame control line which is enabled to a logic high value for the entire duration of a slot as shown in Figure 4.2(a). Short idle gaps (a minimum of 8 bytes) are inserted between slots. The elasticity buffer located in the input receiving stage of the LAC uses the idle gaps to synchronize between the clock received from the network and the local chip clock. The differences in received network clock and local clock frequencies may lead to expansion or shrinkage of the idle gap. A smoother module located after the elasticity buffer in the datapath preserves a minimum idle separation between slots. The slot busy/free status is indicated by a single bit in the header. Access to a free slot is negotiated using the MAC.

Ring initialization is performed by a ring master which also performs error monitoring functions during normal ring operation. While the ring master could be selected actively by the network through an election procedure, we select the master by setting a control register on the LAC of one of the hosts on the network, through the device driver. Since the same LAC is used at all nodes, any host is potentially capable of being a ring master. The addresses for the various nodes in the ring are allocated during the ring initialization process initiated by the master. This is achieved using a hop-count field which is incremented by all nodes during initialization along the ring propagation direction. Subsequently, the master loads the ring with slots to enable normal ring operation.

The ring master can optimize the network for a variety of traffic patterns by adjusting the size and number of slots on the ring. These parameters are configured onto the control registers of the master via the device driver during ring initialization. A packet in the host is broken into smaller cells. A data cell fits within slots of the PONI ring and encapsulates

ATM adaptation layer (AAL) protocol data units. The slot size can be as small as 16 bytes and as large as 1 kB, which is also the size of the transmit and receive buffers on the chip. The PONI network does not rely on the physical layer diameter to accommodate the required number of slots. Instead, the smoother located on each chip provides buffer space adequate to hold the desired number of slots. The result is a flexible, ultra-high bandwidth network capable of supporting a variety of cluster applications from scalable servers to the delivery of multimedia data in a workgroup environment.

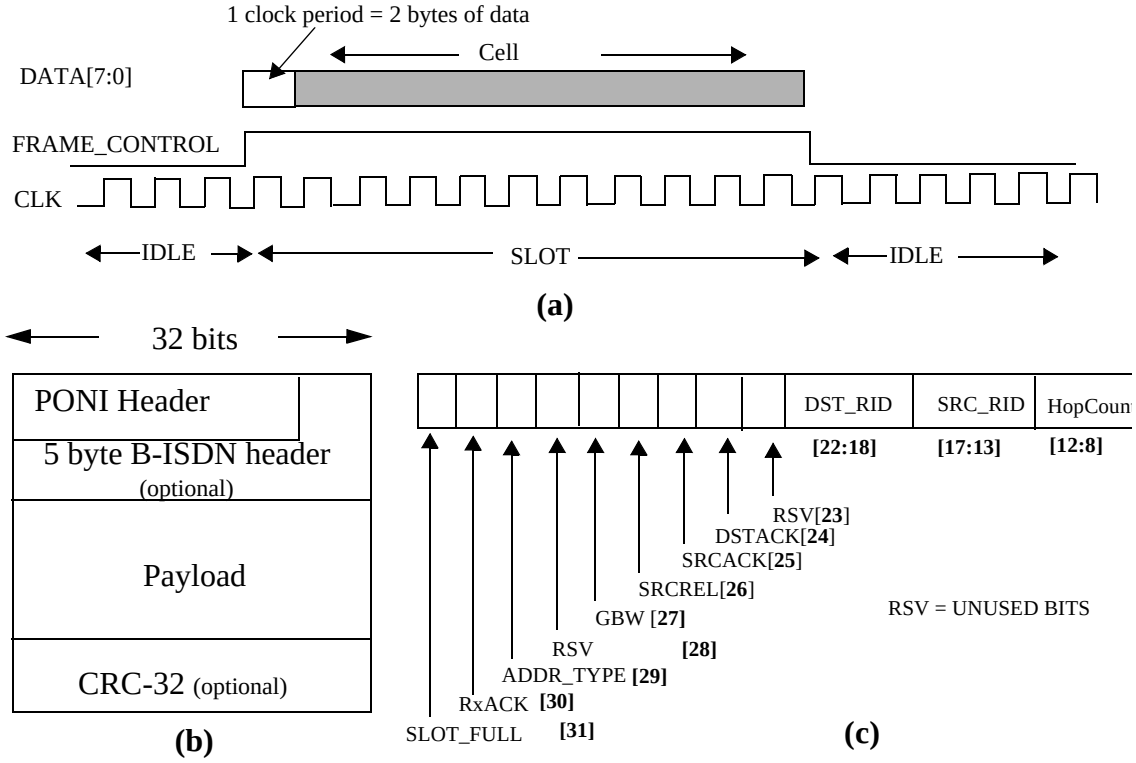


Figure 4.2: (a) Clock, frame control and data line format on the high-speed lines. Slot duration is marked by the frame control line. There are two bits of data on each of the 8 parallel data lines in every clock cycle. (b) Cell format with a 3-byte PONI header and an optional 5-byte B-ISDN header. (c) PONI header format. The SLOT_FULL bit indicates if a slot is busy or free. Transmission access rights are negotiated based on the GBW and SRCREL bits.

The header portion of a network cell as shown in Figure 4.2(b) is 3 bytes wide and precedes the 5 byte B-ISDN header (which is optional) and the AAL protocol data unit. The PONI network header shown in Figure 4.2(c) contains a minimal set of address and control information for slotted ring operation. The first field consists of eight bits of control information. The slot busy/free status is indicated by the SLOT_FULL bit. The next two fields contain the 5-bit source and destination short node addresses. The last field contains the hop-count which is used by the master during normal ring operation for monitoring purposes. The PONI network supports two forms of addressing - one that uses the short node addresses and can be used for rapid decoding, and a second that uses the B-ISDN virtual channel identifier (VCI) address. The VCI address table entries are configured and can be subsequently modified using the device driver. The ADDR_TYPE bit indicates which of the two is used. The LAC can currently recognize up to 1024 addresses with provisions to expand by a further 2048 addresses. It is intended that the VCI address space be used to provide broadcast and multicast capabilities.

Packet removal is a critical issue for any slotted ring. The designed network should balance the competing goals of maximizing bandwidth available to requesting nodes and ensuring fairness of access. A cell could be removed from the network either at the source node or at the destination node, with the source or destination node being allowed to reuse the slots for another transmission. Both schemes, if not correctly monitored, could result in a few nodes utilizing all the network resources and depriving other nodes of transmission opportunities.

The PONI network implements a source removal scheme with or without source reuse. The alternative destination removal scheme with spatial reuse provides higher overall network utilization and bandwidth. Various fairness schemes for such networks have been proposed, however the hardware implementation is more complex. Our goal was to implement the MAC protocol on a single chip that could be accommodated within a high-performance general-purpose package that we had designed. The source removal scheme we use results in simpler hardware implementation which makes it easier to optimize for higher speeds, and simplifies the task of testing. In a small workgroup cluster, applications still have access to very high network bandwidths.

The PONI network provides four priority classes using two bits - the GBW bit and the SRCREL bit. Currently, only two of the four possible priority classes are used. The SRCREL bit being set indicates that slot access is controlled under the source release protocol, while the GBW bit being set indicates that slot access is negotiated under the higher priority guaranteed bandwidth protocol. The network operation under the two schemes is as follows. The source release protocol guarantees fair network access. Under this scheme, after the cell is successfully received by the destination, the source resets the slot full bit. This slot cannot be reused by the source immediately and is passed on to the next downstream neighbor. The guaranteed bandwidth protocol provides applications with a guaranteed upper bound on access latency, since in a network with *n* nodes, the source release protocol can cause a worst case access latency that is *n* times greater. It emulates a constant bit rate connection. Under this scheme, the ring is configured with some of the total number of slots being allocated to specific nodes. The number of such

slots can be flexibly controlled by the ring master. A node can transmit only on those slots whose header source address field matches its own ring identifier. A returning slot whose contents were received successfully can immediately be reused by the source in this mode.

The remaining PONI header bits are used to provide hardware assistance for realizing application performance gains. The RxACK bit is set by the destination on successful reception of a cell. The master monitors the ring for cells recirculating indefinitely, either due to receiver problems or due to corrupted header bits. The payload of short cells is padded in hardware to accommodate cells smaller than the configured slot size. There are three unused header bits to accommodate future additions.

The modules comprising the datapath that implements the digital logic have been constructed from a 0.5 μm CMOS standard-cell library that we developed. They were individually and exhaustively tested using a Verilog-based switch-level simulator. A switch-level simulation of a ring consisting of three nodes was performed using key functional vectors that verify ring initialization, protocol correctness under normal ring operation and error-handling. The size of the PONI LAC chip is 10.2 mm x 7.3 mm. The minimum node latency is less than 150 ns, with additional latency, if any, resulting from that deliberately inserted using the adjustable smoother buffer memory.

4.4 Worst-case throughput analysis of the PONI network protocol design

In the following section, we present a simple, deterministic analysis of the worst-case throughput of a node in the PONI network protocol design. The purpose of this analysis is to establish a lower bound on network bandwidth available for applications at a node. Traffic at transmitting nodes is assumed to be generated continuously in a deterministic

pattern so that nodes transmit data on every accessible free slot. A more detailed analysis of the network protocol performance would include a description of the dynamic behavior of the network under random traffic, and is outside the scope of discussion of this paper.

In a ring consisting of *n* nodes, let the total number of slots be N_{total} which consists of N_{sr} number of slots accessed using the source release protocol, and the remaining slots N_{gbw} accessed using the guaranteed bandwidth protocol. The round-trip latency of the network is T_L. The maximum available system bandwidth at a node is equal to

$$BW_{max} = \frac{Number\ of\ data\ bits}{Number\ of\ data\ bits + Number\ of\ idle\ bits} \times Link\ data\ rate$$

The total number of bits on the ring depends on the link data rate and the network latency including the adjustable node latency. The number of idle bits depends on the minimum required for correct operation of the elasticity buffer to accommodate a pre-specified clock variation between adjacent nodes. The worst case access latency for a node in a network in a steady state with *k* ring nodes that are currently active using the source release protocol (with no source reuse) is $(k+1)T_L$. If BW_{sr} is the total system bandwidth available under the source release protocol from N_{sr} slots, the bandwidth $BW(i)_{sr}$ available to node *i* averaged over an interval equal to this latency is given by

$$BW(i)_{sr} = \frac{BW_{sr}}{k+1}$$

The latency for a node to access a slot under the guaranteed bandwidth protocol is exactly one round-trip latency, T_L. The bandwidth $BW(i)_{gbw}$ available to node i for every slot that it has access to under the guaranteed bandwidth protocol is given by

$$BW(i)_{gbw} = \frac{BW_{max}}{N_{total}}$$

If $n(i)_{gbw}$ represents the number of slots available to the node i under the guaranteed bandwidth protocol, the total bandwidth available to node i is equal to

$$BW(i) = \frac{n(i)_{gbw}}{N_{total}} \bullet BW_{max} + \frac{N_{total} - N_{gbw}}{N_{total}} \bullet \frac{1}{k+1} \bullet BW_{max}, k \leq n$$

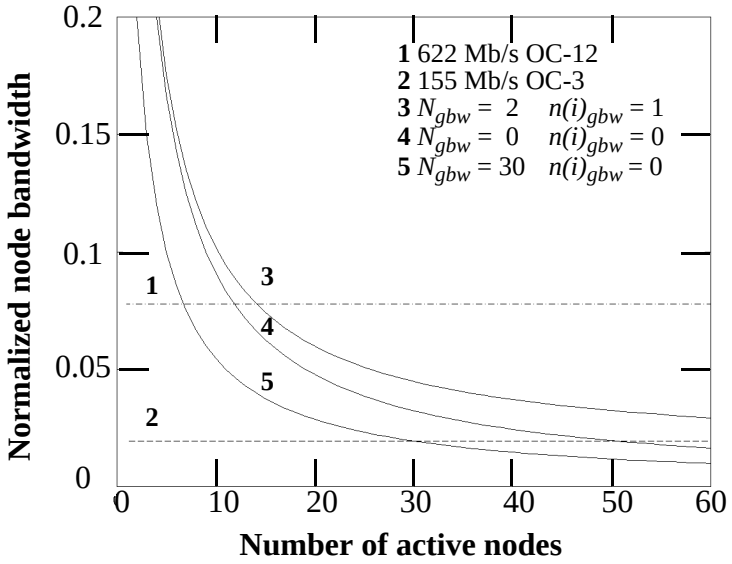


Figure 4.3: Calculated node bandwidth normalized to assumed ring network bandwidth of 8 Gb/s total bandwidth with increase in number of active nodes for N_{total} = 75, and (1) ATM using 622 Mb/s OC-12 link (2) ATM using 155 Mb/s OC-3 link (3) N_{gbw} = 2 $n(i)_{gbw}$ = 1 (4) N_{gbw} = 0 $n(i)_{gbw}$ = 0 (5) N_{gbw} = 30 $n(i)_{gbw}$ = 0. The bandwidth allocated to a node varies depending on ring configuration. It could be greater than (curve '3') or less than (curve '5') the bandwidth available under a purely source release scheme (curve '4') thereby providing adaptability to workgroup needs.

Figure 4.3 shows the bandwidth that is available to a node as a fraction of the total available bandwidth, BW_{max} as the number of active nodes in the network increases. Each node for ATM is assumed to be connected to a centralized switch box using a full-duplex link. As can be seen, an increase in the number of active ring nodes results in a decrease in the bandwidth available to a node. By allocating some slots in the guaranteed bandwidth mode, a node can access more bandwidth than it otherwise would have had access to under a purely source release protocol, so long as the following relationship is observed in a ring with *k* active nodes:

$$\frac{n(i)_{gbw}}{N_{gbw}} > \frac{1}{k+1}$$

If each of the *k* active nodes in the guaranteed bandwidth mode use equal proportions of the guaranteed bandwidth, the above relationship is always satisfied. A node that uses a larger share will result in some nodes obtaining less bandwidth than available in a purely source release scheme. The aggregate system throughput, obtained by summing individual node bandwidths, is given by the following expression:

$$System\ throughput\ =\ BW_{max} \bullet \left(1 - \frac{N_{total} - N_{gbw}}{N_{total}} \bullet \frac{1}{k+1}\right),\ k \le n$$

Higher system throughputs are obtained when using some slots in the guaranteed bandwidth protocol than when all slots are accessed using the source release protocol.

4.5 A brief comparison of PONI with a few LANs

A previous network whose access control has similarities to that used in the PONI network was the Cambridge Fast Ring (CFR) [13] implemented in 1986. This was a 100 Mb/s slotted ring with two modes of operation - the normal mode and the channel mode. Due to the lower link speed, the CFR had restrictions on the slot size (data field of 32 bytes) as well as the total number of slots. In the CFR network protocol, a node was allowed to use only one slot at a time while in the PONI network, a node is allowed to use multiple slots. In the initial implementation of CFR, channel mode was not implemented.

The node latency in CFR was in the microseconds range. The later Cambridge Backbone Network (CBN) [23] was operational at 512 Mb/s and relaxed the one slot per revolution restriction. The CFR and CBN used relatively expensive ECL technology to achieve high line speeds and single-mode optical fiber for the transmission medium.

Examples of current fiber-based local area networks are Gigabit Ethernet and ATM. Gigabit Ethernet [24] is specified to run at a maximum data rate of 1 Gb/s over 550 m when using multimode fiber. The network can operate in half-duplex or full-duplex modes. Collisions due to contention in the half-duplex mode which uses the CSMA/CD MAC protocol degrade the available 1 Gb/s bandwidth. Collisions can be avoided using the full-duplex mode. A multiported Gigabit Ethernet switch is however necessary to fully avoid collisions in addition to the NICs in each host.

In contrast, the PONI network uses a single NIC installed in each host connected to the network medium. This is sufficient to implement a full-duplex network protocol that incorporates admission control features and simultaneous collision-free access by multiple network nodes. The LAC chip in PONI is designed to provide a higher net link data rate of 16 Gb/s. The PONI network is cost-scalable. The total cost of the PONI network when amortized over the high available bandwidth makes it a potentially attractive alternative to provide gigabit access to the desktop in small to medium-sized clusters.

ATM has a complex Quality of Service (QoS) allocation scheme to handle multiple traffic classes with widely varying requirements. While such complex and expensive bandwidth reservation schemes are essential in a WAN environment, less sophisticated protocols are sufficient and cost-effective in a high-bandwidth small LAN environment

that PONI targets. Broadcast and multicast modes are easier to support in the shared-medium PONI ring network as opposed to switch-based networks such as ATM or full-duplex collision-free Gigabit Ethernet. The PONI header format provides WAN compatibility through support for interoperability with ATM.

The PONI ring network has potentially lower latency than centralized switches making it an attractive choice for cluster applications. The measured latency for a 64-byte cell through an Ethernet switch between two switch ports in the same chassis and under minimum load is 17 μs [69]. The latency for a 32-node PONI network under the guaranteed bandwidth protocol scheme, irrespective of total network load, has the following components:- (a) worst-case slot access latency of 6.4 μs (b) the time taken to load or unload a 64 byte packet from FIFOs at a TTL speed of 62.5 MHz is 256 ns (c) internodal fiber latency of 50 ns over a distance of 10 m per node (d) a node's input receive port to output transmit port latency of 150 ns. The node latencies for the source and destination nodes aggregate to that of an entire node. In a 32-node network, this gives a worst-case latency of 13.1 μs for a destination that is immediately upstream to the source. Thus, by using small networks with less than 32 nodes the round-trip network latency can be kept low. Larger networks can be constructed using a multiple hierarchy of interconnected rings as described in earlier work on ring networks [13].

4.6 Reliability

The issue of reliability in ring networks wherein failure of a node in a single (unidirectional) ring leads to ring breakdown is not addressed in this thesis. FDDI has well-established techniques to perform electrical and optical bypassing of faulty nodes.

Optical bypass switches for parallel fiber-optic links have not been implemented though such technologies are available currently. An example of a reliable unidirectional ring network configuration is shown in Figure 4.4 where host computers are connected to a central concentrator in a star configuration through intermediate optical bypass switches. In the bypass state, the switch automatically transitions from the inserted state to the bypass state (currently in less than 25 milliseconds for serial optical links) thus maintaining ring integrity. There is thus the potential for constructing ring networks that are tolerant to failure of ring nodes.

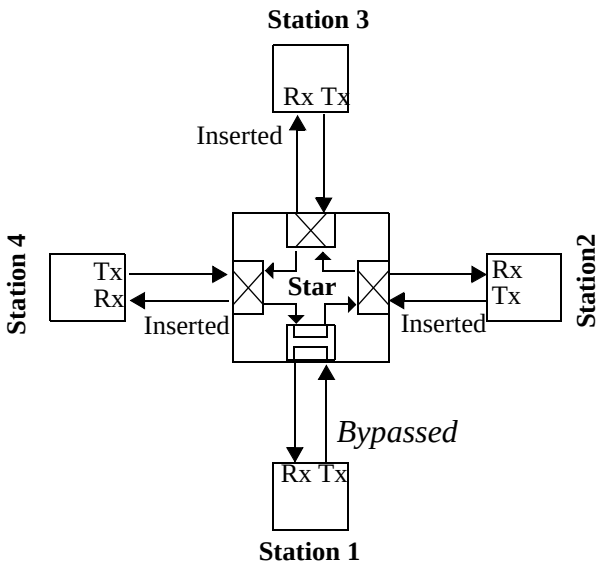


Figure 4.4: Example configuration of a reliable unidirectional ring constructed of single-attach nodes connected to a concentrator in a star configuration through optical bypass switch interfaces. Station 1 is in the bypassed mode while stations 2, 3 and 4 are connected to the ring network. Optical bypass switches enable reliable unidirectional rings practically independent of data rates. Optical bypass switches for serial optical links exist. Switches for parallel fiber-optic links have not been implemented though such technologies are available currently.

4.7 Summary

In this chapter, we have described the architecture of a slotted-ring network known as the PONI network targeted at network data rate of over 8 Gb/s. The PONI network uses a ten-wide multimode fiber-ribbon as the physical medium. Eight of the ten signal lines are used for transmitting data while the remaining two are used for transmitting clock and control information. The PONI network provides two quality-of-service traffic priority classes for sharing bandwidth resources - the lower priority source release protocol and the higher priority guaranteed bandwidth protocol. Packets are released at the source which cannot reuse the slot it just released. The PONI ring network differs from previous ring networks chiefly in the use of parallel fiber-optics as the physical medium and from ring networks of the early 1990s in the use of single-chip CMOS-based network interface solutions. Advantages of the network architecture described are equally applicable to parallel fiber-optic networks based on wavelength division multiplexing (WDM) technology.

Chapter 5

LAC

5.1 Introduction

As outlined in Chapter 4, we chose to implement a slotted-ring network due to its potential for broadband implementation as well as cost-effectiveness arising from its simplicity relative to alternative schemes such as those based on centralized switch-based and shared bus-based networks. The simplicity of the ring network implemented using conventional CMOS technology enables higher data rates than would be possible in more complex systems.

In subsequent sections of this chapter, we describe the architecture and implementation of the link adapter chip for the PONI network known as LAC. The link adapter chip achieves a network aggregate data rate in excess of 2 GB/s with digital logic clocked at over 500 MHz speeds achieved entirely using conventional CMOS technology interfacing to parallel fiber-optics. This is to our knowledge the highest data rate achieved in a dedicated shared medium ring network reported to date. There are other parallel link ring networks implemented using WDM for the metropolitan area network such as the

two-wavelength HORNET [59][60]. Here, two wavelengths are currently used to implement a ring network with a specified link rate of 2.5 Gb/s on each independent ring network where an entire packet is transmitted onto a single ring.

Previous high-speed ring networks for the local area network include the CRMA-II [86] which achieves 2.4 Gb/s using ECL and CMOS technologies and single-mode optical fiber physical medium. Higher speeds have been achieved in system area networks such as the Sebring Ring [93] which achieves a unidirectional speed of 532 MB/s using a 16-bit wide bus clocked at 266 MHz. Each of the Coherent Toroidal Interconnect (CTI) rings used in the HP-Convex Exemplar supercomputer system SPP 2000 series [94][95] has a bandwidth of up to 660 MB/s. The HP Exemplar V2600 system in the HP 9000 series [96] has a higher rate of 960 MB/s. The Cray T3E uses a GigaRing channel based on the SCI [31] protocol for processor I/O interconnects as seen in Figure 5.1. The ring is clocked at 600 MB/s (32 bits at 150 Mb/s rates per signal upgradable for up to 250 Mb/s per signal) data rates but capable of providing bandwidths up to a GB/s. In March 2001, an IEEE study group known as the Resilient Packet Ring study group was formed to define a standard for 10 Gb/s ring networks.

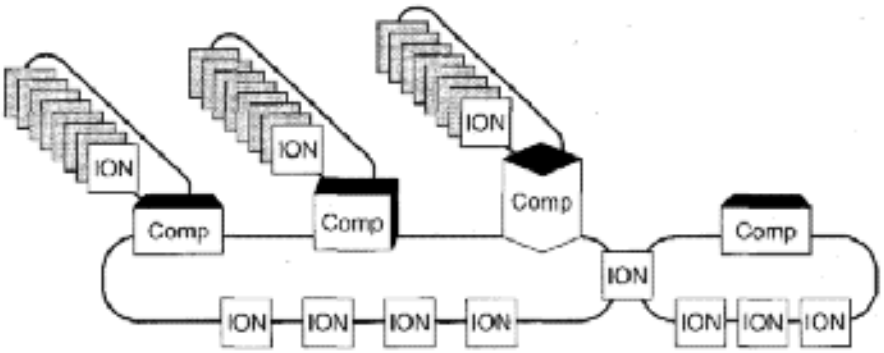


Figure 5.1: Schematic of I/O nodes in a Cray T3E supercomputer system interconnected by a GigaRing Channel. The GigaRing is clocked at 600 MB/s and capable of up to 1 GB/s data throughput.

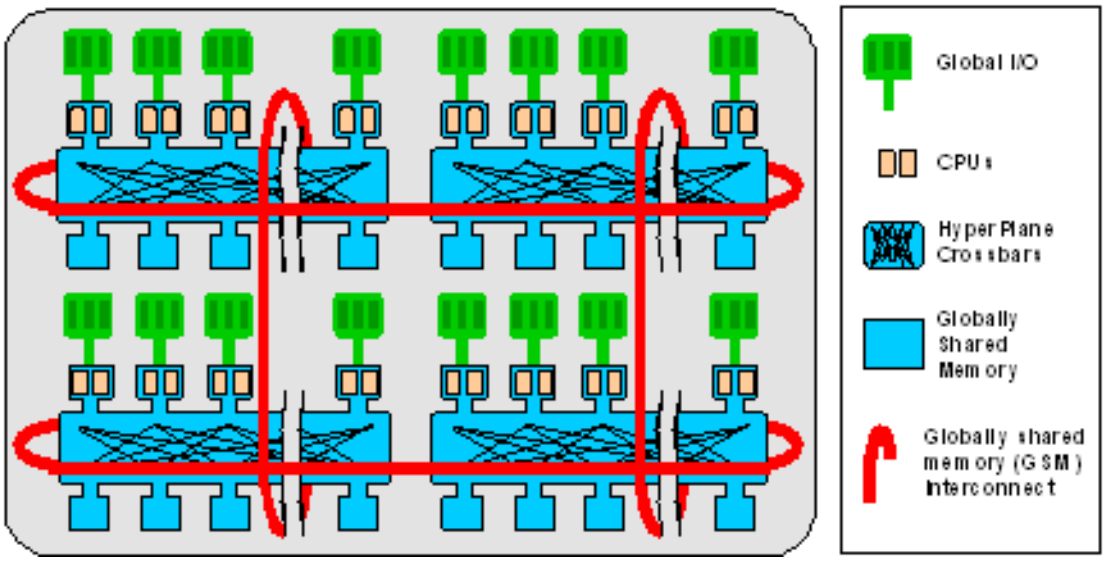


Figure 5.2: Architecture of the HP/Convex V2600 Exemplar architecture, a supercomputer used as a high-end unix server. The globally shared memory interconnect consists of four Coherent Toroidal Interconnect (CTI) rings with a bandwidth of 960 MB/s per ring.

5.2 LAC microarchitecture

This section describes the design and implementation of the link adapter interface chip (LAC). The LAC contains all of the high-speed circuitry needed to interface directly to the PONI optoelectronic module. It also contains the media access control (MAC) functions of the PONI network and provides data buffering on-chip with separate 1 KB receive and transmit FIFOs.

The physical layer high-speed interface circuitry in the LAC is the same as that used for the P2P described in detail in section 3.3 on page 46. The physical layer consists of a ten-wide parallel fiber-ribbon, eight of which are data lines, one is a clock and one is the frame control line. The data put onto the physical layer is of differential reduced voltage swing LVDS (low voltage differential swing) format. Data that is 32-bits wide is received at the high-speed transmitter inputs and multiplexed onto eight output data lines. At the receiver, this data is again demultiplexed to produce 32-bit wide data.

Independent 32-bit wide Rx and Tx ports are provided to the host computer interface. These data ports have a programmable operating clock frequency of the digital logic clock frequency divided by 2, 4, 8 or 16 up to a maximum of 50 MHz (due to limitations of a reverse-clocked TTL interface) and are TTL compatible for ease of system integration with external commercially available buffers. Host control of the LAC is accomplished via a control bus (Cntrl) which can access control and status registers within the chip along with the VCI address table.

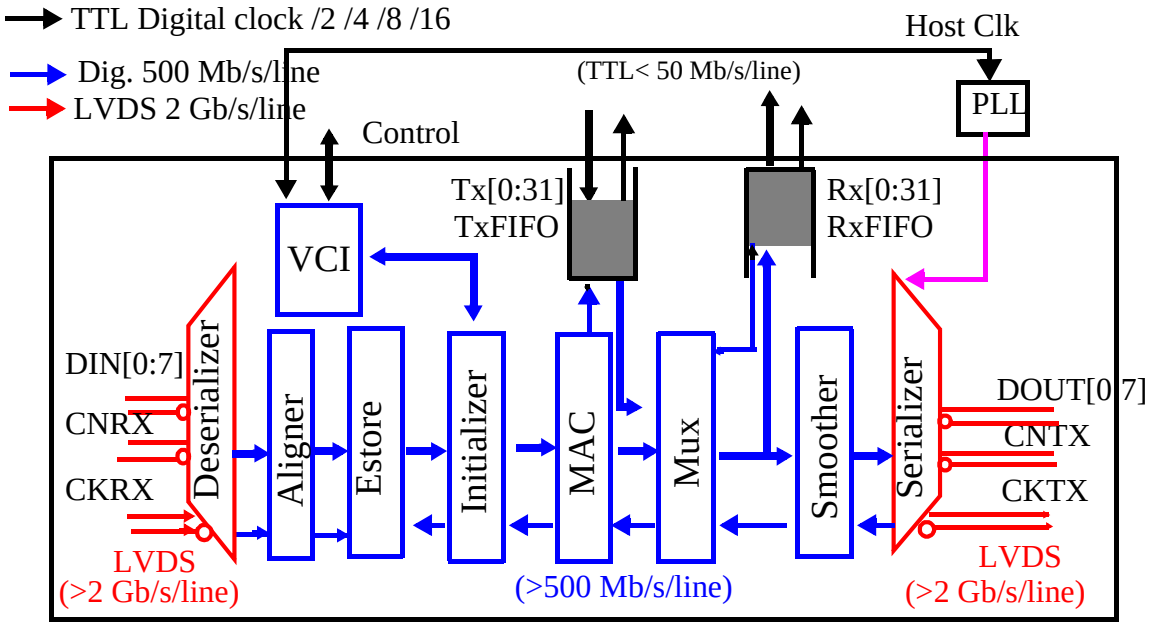


Figure 5.3: Block diagram of LAC. Figure 5.3 shows the top level block diagram of the LAC. The interface to the PONI module consists of a pair of 10-wide Rx and Tx LVDS ports. These ports directly connect to the module and are de-skewed on-board for 1 GHz operation. The high-speed ports consist of one clock, one control and 8 data channels. No line encoding is used for the data or control channels. The positions of TxFIFO and RxFIFO in the datapath should be interchanged if implementing a destination removal protocol with spatial reuse. For our source removal scheme with no source reuse, there is no loss in performance for the current implementation.

A separate phase lock loop (PLL) is mounted on the interface card which locks to the host's system clock. The control bus interface operates at this clock frequency to ensure correct control of the LAC during initialization. Since the network is source clocked, the transmit clocks are also locked to the host's frequency and used to synchronize the PONI transmit port of the chip. A clock for the digital logic circuitry is generated from the

transmit clock using a toggle flip-flop divider circuitry as described in Chapter 3. Clocking in the PONI network is a distributed asynchronous scheme. Incoming data to the chip is clocked by the upstream generated receive clock (CKRX) and used to deserialize the data stream. To accommodate the bit shifts in the received deserialized data which may occur due to the receiver's deserializer being phase mismatched with its upstream transmitter's serializer, there is an aligning module. The incoming deserialized stream is word aligned with respect to the frame valid control signal and then written into the elastic store (estore). The estore is an asynchronously clocked module with the write port clocked using a digital clock derived from the received network clock while the read port is clocked using the locally generated digital clock. Thus, the data is synchronized to the local clock following the estore. Subsequently, the received packet is propagated along the datapath so that the header can be decoded.

If a host has a packet to transmit, it first queues the packet into the on-chip FIFO. The MAC then waits for an empty slot on the ring. As the empty slot enters the ring it is marked as a full slot and filled with the outgoing packet. As the TxFIFO is drained the host can concurrently fill up the on-chip FIFO with another packet for the network.

Multiple packets can reside in the on-chip FIFOs depending upon the slot size. The total capacity of the on-chip FIFOs is 1 KB. If the slots are smaller than this size (e.g. 56 Bytes - as would be the size of an ATM cell with the 3 byte LAC header) then multiple packets can be queued on-chip. However, once the RxFIFO of the LAC is full,

subsequent network packets addressed to the node will not be accepted and will have to be retransmitted. If the Tx FIFO is full, then the host must wait until it is drained before more packets are queued.

In the datapath currently implemented, the TxFIFO precedes the RxFIFO. This was done since it was originally intended to implement a Cyclic Redundancy Checksum (CRC) unit in the datapath. Since the CRC unit is fairly large and consumes a lot of power, it was intended to share the same unit for both the transmit and receive operations on a packet. However, for simplicity, the CRC was omitted from the final implementation. In a ring network which uses a destination release link protocol with spatial reuse, the RxFIFO should precede the TxFIFO in the datapath. However, since we implement a source release protocol with no source reuse, the performance is not affected by the placement of FIFOs.

The penultimate stage in the datapath before the serializer is the smoother module which regulates the idle spacing between slots and maintains it at the minimum value required for proper operation. Finally, the data is shipped out from the high speed serializer end. The datapath is highly pipelined for high frequency processing with single logic gate delay per pipeline stage. In the following sections, the operation of each of the individual blocks in the LAC datapath will be discussed in greater detail.

5.2.1 Aligner

At the receiver, the eight high-speed parallel data lines are demultiplexed to produce 32-bit wide data. This data is word aligned in hardware by the aligner module shown in Figure 3.12 on page 63. The operation of the aligner, which is also used in the P2P chip,

has been discussed previously in section 3.4.3 on page 62. Alignment is required when the demultiplexed signals at the receive port are not phase aligned with the multiplexed signals of the transmit port. Output data from the aligner is written into the estore memory.

5.2.2 Elasticity buffer (Estore)

The aligner output is written into the elasticity buffer (estore) memory. The function of the estore is to interface between two plesiochronous systems (corresponding to two clock domains of nearly the same frequency), one derived from the clock received from the network and the other generated locally derived from a local host computer's clock. The estore compensates for phase differences between the two clocks, as well as small frequency differences and variations.

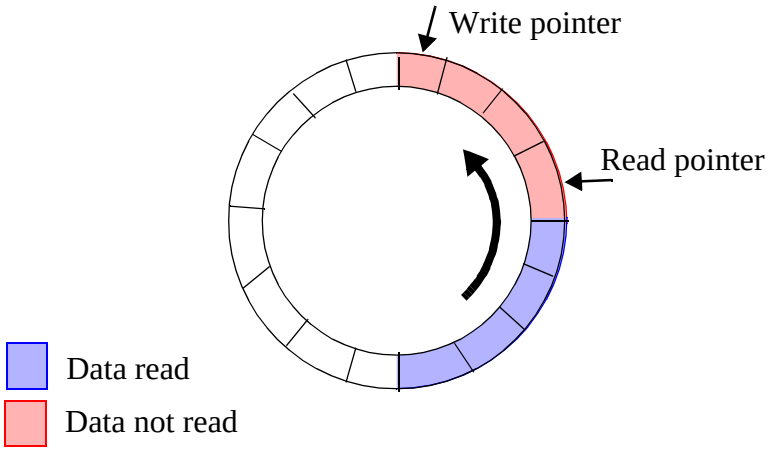


Figure 5.4: Logical diagram of the estore buffer. The estore is a 16-word deep buffer with independent and asynchronously clocked write and read ports. Read operation commences after a programmable preset number of words are initially written into the buffer.

The estore used in the LAC is 16 words (a word is 32 bits wide) deep. The memory element used is a 64 byte (or 16 32-bit words) dual-ported SRAM. The logical operation of the estore depicted in Figure 5.4 is as follows. Prior to the start of a frame, the read and write pointers point to the same memory location. On receiving a slot (indicated by the frame control line going HIGH), the write operation commences while the read pointer is still stationary. When the number of words written into the buffer equals a programmable preset value, the read operation commences. This initial separation between the write and read pointers is introduced to account for small frequency differences between them which can otherwise cause an underflow (where read pointer overtakes the write pointer) or overflow (where write pointer wraps around the circular buffer and overtakes the read pointer). At the end of a frame, the write operation ceases. Meanwhile, the read operation continues until it has completed reading all the data written into the buffer, and stops at the memory location addressed by the write pointer. Thus the idle symbols between slots are used to recover from any clock variations in the two clock domains as is necessary to overcome the initial separation between write and read pointers.

The maximum frequency difference that the estore can accommodate is given by the expression

$(\Delta f/f) \times$ (max. slot size in words) = Initial lag of faster pointer behind slower pointer

where, $\Delta f/f = (f_+ - f_-)/f_+$ is the fractional clock frequency difference between the faster (f_+) and slower (f_-) pointers, slot size is the maximum size of the slots used in the ring in words, and the depth of the estore is also measured in the maximum number of words it can hold (here, 16). An overflow occurs when write pointer overtakes the read pointer and

an underflow occurs when read pointer overtakes the write pointer. The allowable slot size is a maximum when initial separation of write and read pointers is a maximum (up to half the depth of the estore, here 8). Assuming a maximum slot size of 256 words (given by the size of the TxFIFO and RxFIFO), the maximum allowable clock difference is 3%. In a network composed of computers with identical clock frequencies, the clock at each host varies in frequency only due to the variations in crystal frequency which is less than 1%. For example, the measured clock frequency variation among Hewlett-Packard HP 735 workstations was observed to be about 0.1%. Hence, by adjusting the programmable write pointer offset, frequency variations in the asynchronously clocked network can be accommodated by the estore design.

The estore operation can however create a shrinkage in the idle gap between slots, which if uncorrected, can lead to ring failure by corrupting the slots. This idle gap size is maintained by another module in the datapath, namely the smoother the need for which was specified in the FDDI architecture [22].

5.2.3 Initializer

Data read out of the estore memory propagates to the initializer pipe stage. The initializer module performs various tasks during initialization of the ring, such as detecting whether an upstream neighbor's clock (Figure 5.6) is active, assigning ring addresses for individual nodes and slot creation as specified by the states of the initializer state machine shown in Figure 5.5. The various stages of the Initializer operation are described in Table 5.1.

The PONI ring has a master node that initiates the initialization process for all nodes connected to the ring. In a practical network, the node that will be the master is negotiated in hardware so that a new master can be dynamically elected if a ring node that was originally the ring master fails and/or is removed from the network. However, in our prototype implementation, an external control line is used to control whether a node is a master or a slave, i.e. it is hardwired. After the LAC has been reset, the master node initiates the procedure of detecting whether all nodes in the ring network are active by transmitting a toggling pattern on one of the data lines. Each of the slave nodes in turn detects this toggling pattern and repeats it to its downstream node. When the loop is complete as detected by the master receiving a toggling pattern from its upstream neighbor, it proceeds to initialize the ring nodes with their ring addresses. To do so, it transmits an initialization packet with a 5-bit hop count field in its header. Each downstream node increments this field and assigns itself that node address as a ring node identifier (ID) and passes the packet on to its own downstream neighbor in the ring. When the packet completes one round trip, it is removed from the ring by the master which then initiates the creation of slots separated by idles. When all the slots for the ring have been generated, initialization is complete and normal ring operation can commence.

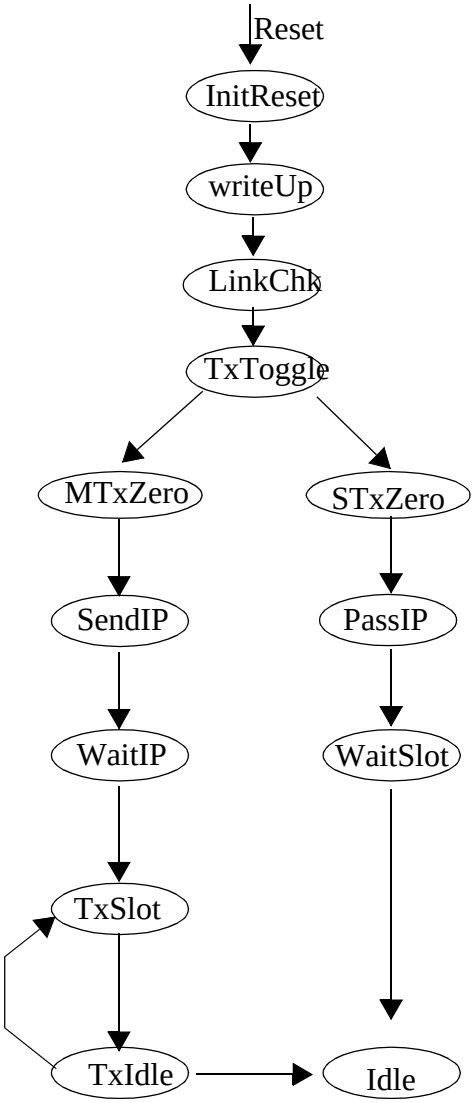


Figure 5.5: States of the initializer state machine. On reset, the state machine enters the Initreset state. The initializer state machine is responsible for ring initialization with tasks such as checking if the network is up. If so, the master node generates the slots for the ring network. Once slots have been generated, the final state is the Idle state where regular ring operation can commence.

State	Description
Initreset	Initializer state machine enters this state on reset, wait until host interface indicates that the various LAC registers have been programmed such as slot size, smoother delay size etc.
writeup	Introduce desired delay in smoother buffer module to initialize the ring size to accommodate the required size and number of slots
linkchk	Wait in this state until it is known that link with upstream neighbor is active indicated by ascertaining if upstream neighbor's clock is up (if master) and a toggle pattern on the first signal line of the eight-wide incoming stream is received (if slave)
txtoggle	Master initiates a toggling pattern on one of the eight outgoing signal lines on the physical layer when its upstream neighbor's clock is determined to be active. Slave initiates a toggling pattern when it receives a toggling pattern from its upstream neighbor.
mtxzero	Master initiates deactivation of toggling of slaves by replacing the toggle pattern with an all-zero pattern
stxzero	When slave node stops receiving an incoming toggle, it stops transmitting a toggle and transmits all zeros as well.
sendip	When master's upstream node has stopped transmitting an idle pattern, it initiates ring node address initialization by transmitting an initialization packet.
passip	Slave passes on the initialization packet, increments the hop count field in the packet header and assigns itself a ring address corresponding to an incrementing of the hop count field
waitip	Master waits in this state until the initialization packet has made one complete round trip so that it can commence slot generation
waitslot	slave transitions to this state after assigning itself a ring node address and waits for the first incoming slot
txslot	Transmit slot
txidle	Transmit idles to create idle gap between slots

Table 5.1: Description of various states of initializer state machine

State	Description
Idle	Ring has been initialized with slots and regular ring operation commences

Table 5.1: Description of various states of initializer state machine

The operation of the received-clock-active circuitry (Rclkok) is shown in Figure 5.6. The incoming received clock is divided by four and retimed using the local clock. The local clock then takes four successive samples of this retimed signal and looks for a zero-to-one transition between adjacent pairs.

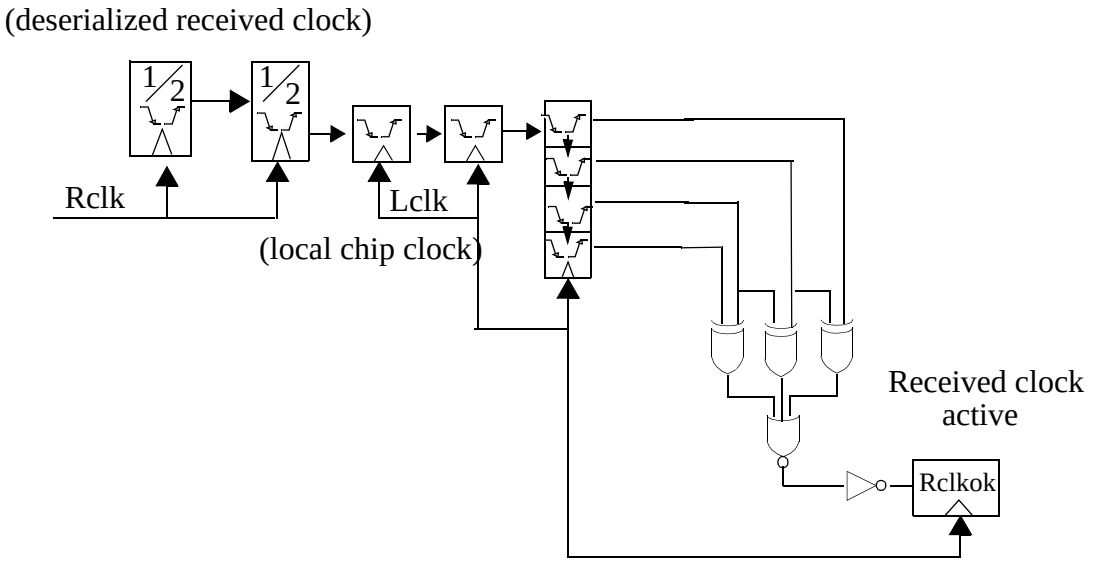


Figure 5.6: Clock detection circuit. The above schematic represents the circuitry used for detecting whether the upstream ring neighbor's clock is up. Received deserialized clock (Rclk) is divided by four and sampled using the local clock (Lclk). A bit transition between at least one adjacent pair is then looked for to indicate that received clock is active.

The ring can be disabled if at any time any ring node detects its upstream clock to be inactive. Subsequently, it enters the InitReset state and slots in the ring will be inactivated. Currently, in our prototype implementation, we have not designed for a ring recovery. However for practical operation, a ring master will have to be dynamically re-elected and the ring subsequently re-initialized with slots.

5.2.4 VCI

The PONI ring has provisions for two modes of addressing - one that uses the ring node address assigned during initialization and the second that uses a virtual channel identifier (VCI) address such as is required in ATM networks. The VCI memory holds the address table required for this mode of operation. Further, it also provides limited programmability to the LAC through a control interface port to the host computer by means of control registers that control various parameters used in the LAC such as slot size and idle size. It also reports status in the LAC such as errors occurring due to overflow in the estore to the host.

The VCI memory is a RAM constructed out of a 1 Kb dual-ported SRAM. The memory is organized as four interleaved banks that are 32 rows tall. Each bank is one byte wide. This address space is used for both the VCI address table as well as the control and status registers. The interface to a host computer uses TTL signal levels. There are 7 address lines to address the VCI memory as well as eight bidirectional data lines to

provide a byte-wide data interface for reading from and writing into memory. The host computer supplies a TTL clock (< 50 MHz) to act as the clock for the host computer's interface with the VCI block.

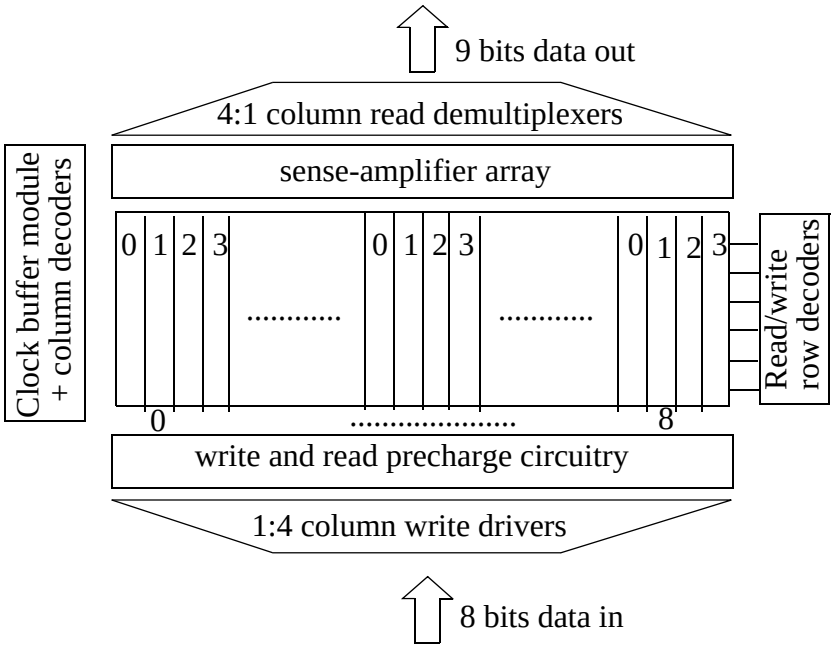


Figure 5.7: Block diagram of VCI memory. The VCI memory has 32-rows of four byte-wide banks. A seven-bit write address is used to address a byte of the memory for write operations. A ten-bit read address is used to address a byte of memory for read operations with the additional three-bits used to select a single bit of the byte-wide output.

Accesses to the VCI memory are random as opposed to the sequential method of access seen in a FIFO. In the VCI memory, only one clock phase is available for precharging the bitlines. The column bitlines are precharged using the precharge circuitry shown in Figure 5.8 on every alternate phase of memory clock pc_en (when clock is high). Correspondingly, there is only one clock phase available for writing into memory cells or reading from memory cells. This operation takes place in the low phase of clock signal

line pc_en in contrast with FIFOs where a whole clock cycle is available for writing to and reading from memory. A shorter evaluation time implies that bitline swings at sense-amplifier inputs will be smaller. Hence, achievable speeds in a RAM are lower than that of a FIFO of the same column height.

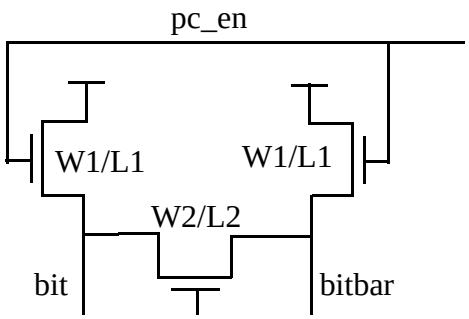


Figure 5.8: Circuit schematic of precharge circuitry used for write or read bitlines (bit and bitbar), in the VCI memory. The precharge enable line used is the clock used by the VCI memory. Hence, there is one clock phase to pre-charge a bitline whose total load is close to 300 fF.

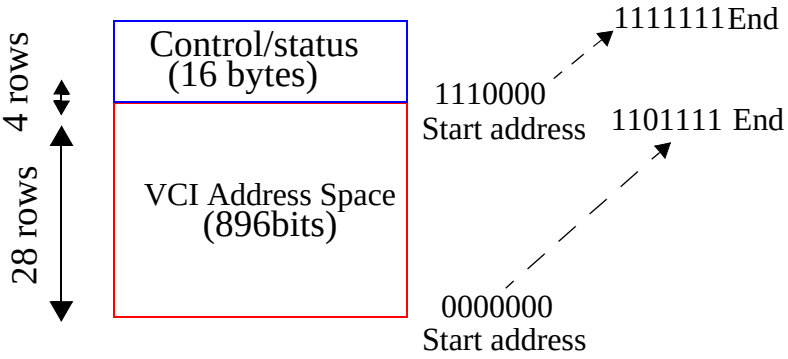


Figure 5.9: Address space in VCI. The VCI is a 128 byte (1 Kb) memory. Of this, 16 bytes are allocated for the control/status registers and various registers used by the datapath such as the slot size, idle size etc. The remaining 896 bits are used for the VCI address space thus yielding 896 addresses. A seven bit address is used for addressing the memory.

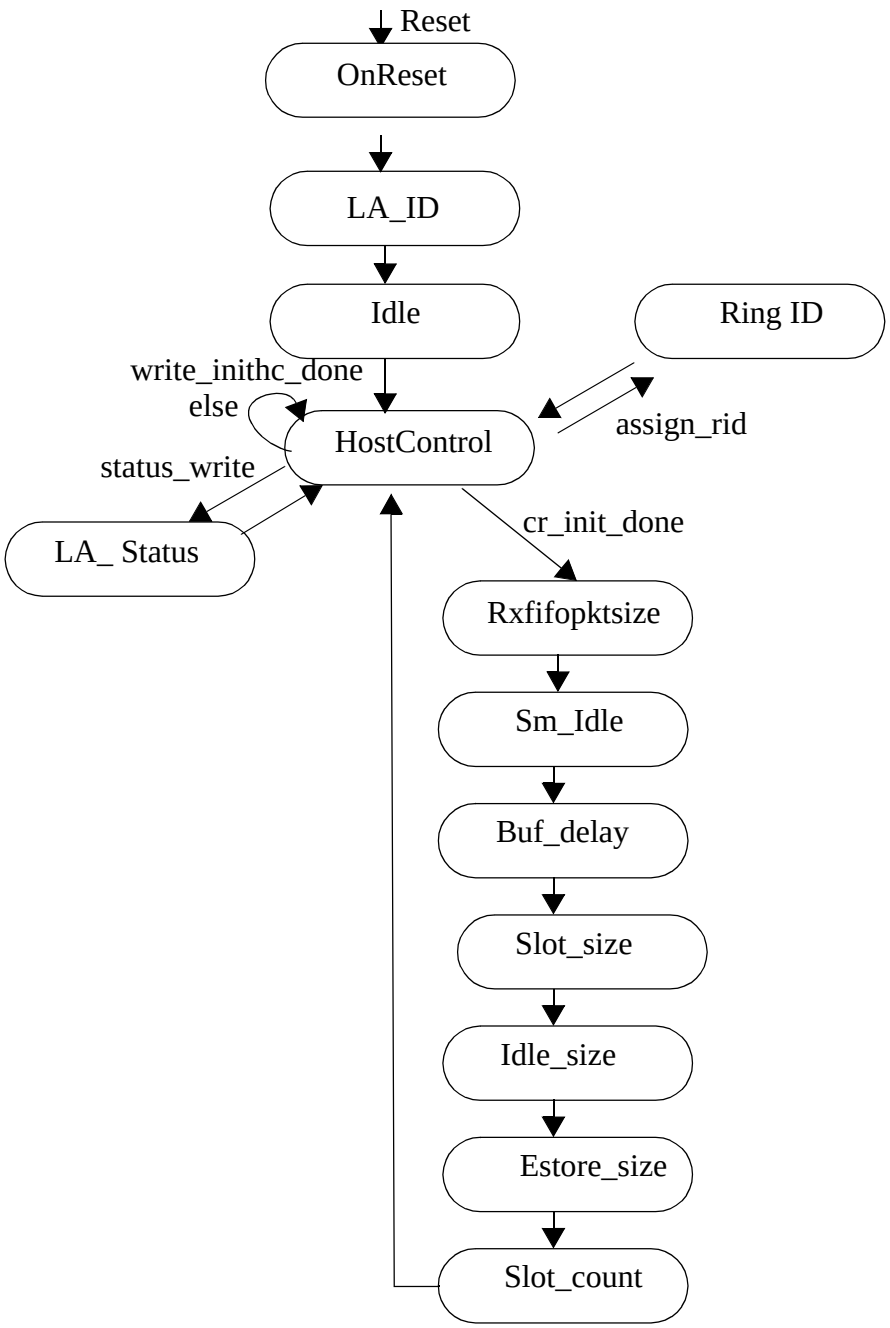


Figure 5.10: Control register state machine implemented as a Moore state machine. The state machine loads various registers needed by the LAC such as smoother idle size, buffer delay size, slot size, idle size and so on. It also provides the address used for writing status information or reading control information.

The operation of the finite state machine used to access the control/status register space of the VCI block is shown in Figure 5.10. On applying a reset, the LA_ID register which represents a tag used to address the chip to be the LAC is read from memory. Subsequently, the various registers used by the LAC such as slot size, idle size etc. are read from memory. The status register is written into whenever the LAC has any status to report such as an error in the elasticity buffer or smoother buffer module. The various states of the state machine are described in Table 5.2.

State	Description
OnReset	State entered into upon applying chip reset
LA_ID	Read the ID of the LAC hard-wired into the chip - this ID can be used to identify that the chip on the network interface card is for a slotted-ring network as implemented by LAC
Idle	Wait in this state until host control register in VCI memory block is initialized
HostControl	Constantly read from host control register in this state
LA_status	Write into status register to report LAC status to external host
Rxfifopktsize	Initialize register used by RxFIFO to determine packet size needed to evaluate if RxFIFO has sufficient buffer space to receive the packet
Sm_Idle	Initialize register used by smoother to determine minimum idle gap spacing to be maintained between slots
Buf_delay	Initializer register used by smoother to insert buffer space into ring adequate to hold the required number of slots
slot_size	Initialize register containing size of slot used in ring network

Table 5.2: state description of control register state machine

State	Description
idle_size	Initialize register containing size of idle gap used in ring network
estore_size	Initialize register determining latency of elasticity buffer as necessary for a predetermined allowable frequency variation between upstream and immediate downstream neighbor nodes
slot_count	Initialize register determining number of slots in ring network
Ring_ID	Report ring node ID assigned during ring initialization to external host

Table 5.2: state description of control register state machine

5.2.5 Medium access control (MAC) unit

Input the medium access control module which implements the ring medium access control (MAC) link layer protocol is received from the initializer module. Access to the slots is negotiated using the MAC protocol described in Section 4.3. On any received packet, the following actions can be performed based on MAC protocol as indicated in the MAC state machine shown in Figure 5.11: transmit, receive, pass and strip. If the slot is empty and if the node has appropriate transmission rights, it can load a packet from its TxFIFO onto the slot. If the slot contains a packet destined for the current node and the current node is able to accept it, it receives the packet. If the node was the originator of the slot that was successfully received or if the slot contains erroneous packet information, and if the current node is a master, then it empties the slot. In all other cases, the node passes along the slot.

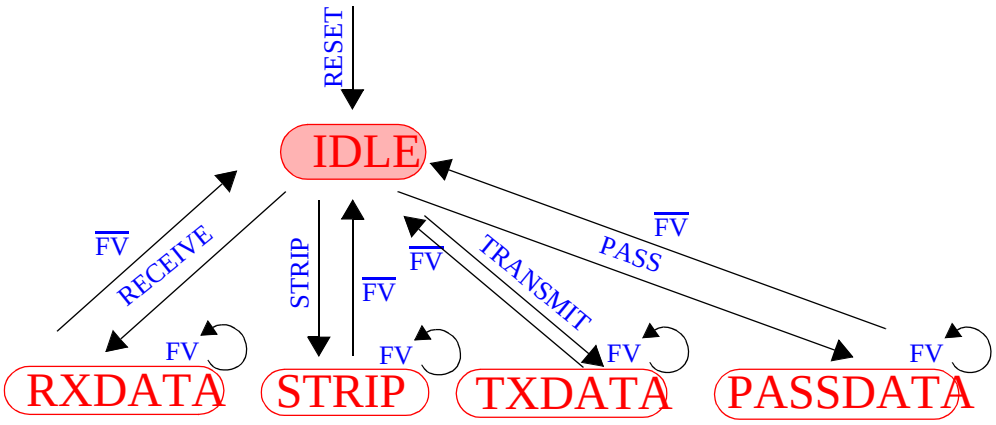


Figure 5.11: state machine for the medium access control designed as an output encoded Moore state machine to maximize speed of operation. There are four possible operations on any incoming slot - receive the packet (RxDATA), load a new packet (TXDATA), pass the packet (PASSDATA) or empty the slot (STRIP). Transitions to each of the four states are under conditions RECEIVE, STRIP, TRANSMIT and PASS as designed for in the network protocol. The state machine returns to the IDLE state when frame (FV) is low, or during an idle interval.

On any packet, only one of the transmit, receive, strip or pass functions can be performed. This enables the medium access control state machine to be implemented as an output encoded Moore state machine maximizing its speed of operation. The state encoding is thus 0000 (IDLE), 1000 (RXDATA), 0100 (STRIP), 0010 (TXDATA) and 0001 (PASSDATA). The state variables thus themselves act as the outputs needed for datapath logic in each of the transmit, receive, strip or pass modes. A packet being successfully received by the master which has a ring node ID of 0 (specified by header bits 22:18) is shown in Figure 5.12. A packet being successfully received by the slave which has a ring node ID of 1 (specified by header bits 22:18) is shown in Figure 5.14.

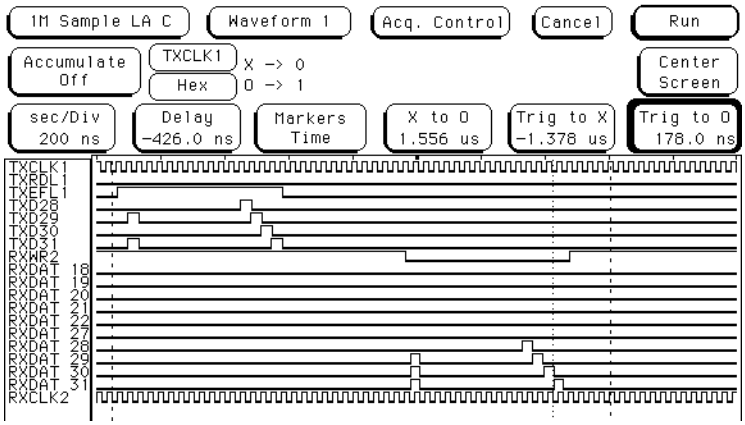


Figure 5.12: Measured TTL output viewed on an HP 16500B logic analyzer system of packet received by master with ring ID of 0. High-speed serializer/deserializer clock frequency is 1 GHz, digital logic clock frequency is 500 MHz and TTL clock is programmed to run at 1/16 of digital logic clock frequency (31.25 MHz). The start of the received packet is the value of received data RxDAT at the first clock edge of clock RxCLK after control line RxWR2 goes low. Received header bit RxDAT<30> being high shows acknowledgment of successful receipt.

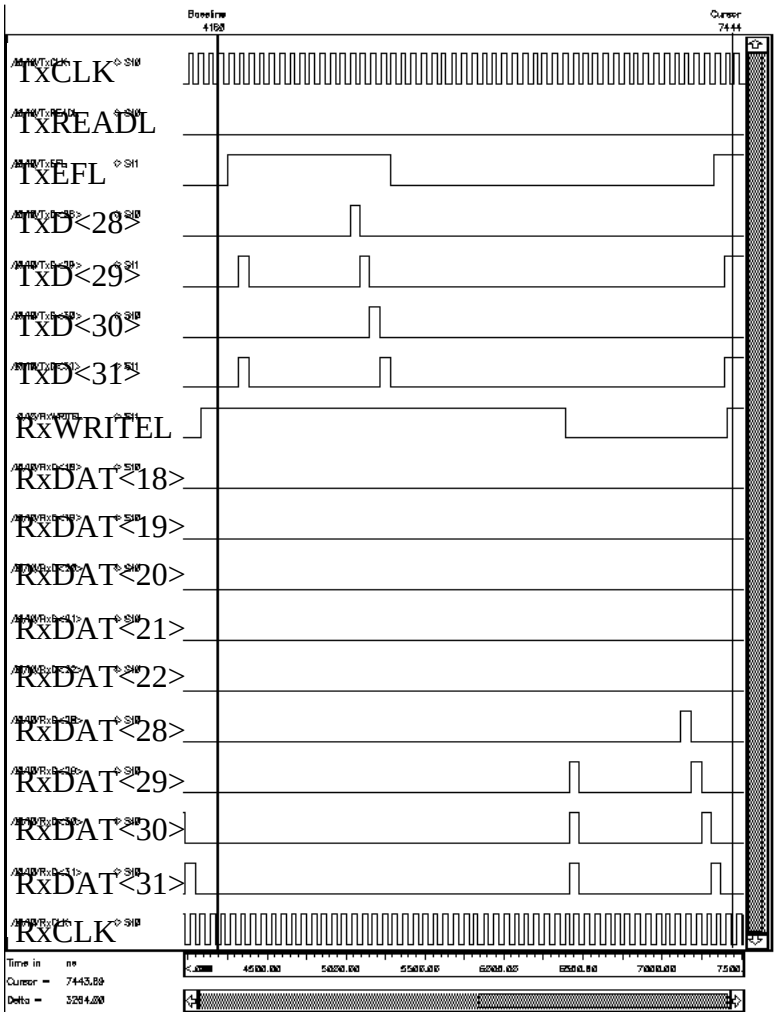


Figure 5.13: Simulation using Verilog switch-level simulator for master receiving a packet for the same parameters described in Figure 5.12. Measurements confirm expected simulated behavior. External loop-back wire delay of measurement setup was not used in simulation.

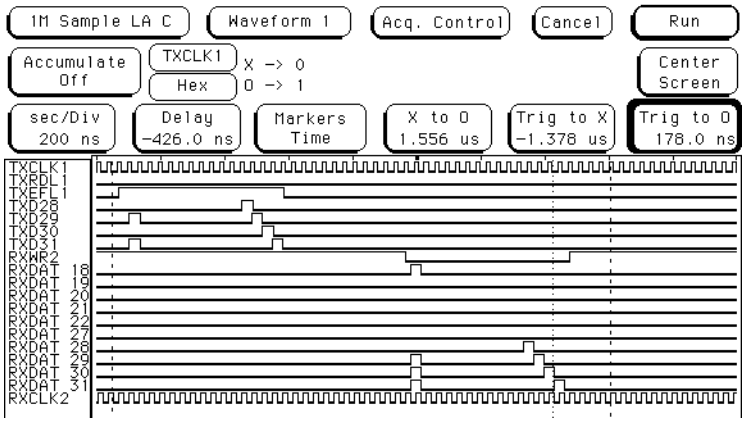


Figure 5.14: Measured TTL output viewed on HP 16500B logic analyzer system of packet received by slave with ring ID of 1. High-speed clock is 1 GHz, digital logic clock is 500 MHz and TTL clock is programmed to run at 1/16 of digital logic clock (31.25 MHz). The start of the received packet is the value of received data RxDAT at the first clock edge of clock RxCLK after control line RxWR2 goes low. Received header bit RxDAT<30> being high shows acknowledgment of successful receipt.

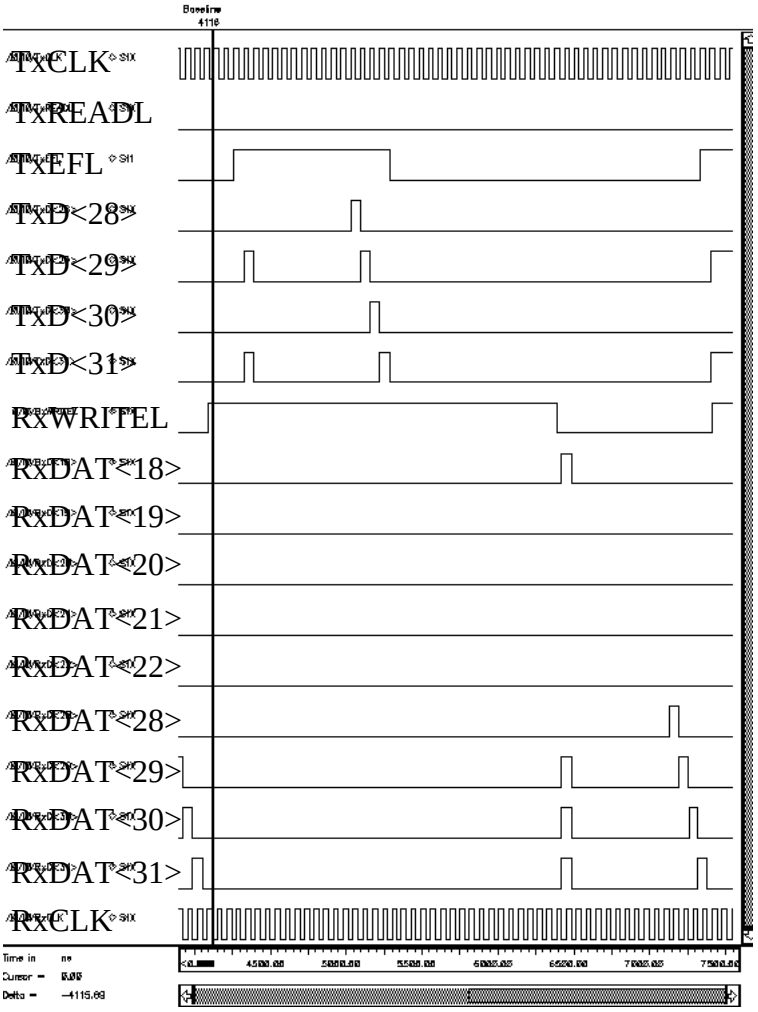


Figure 5.15: Simulation using Verilog switch-level simulator for slave receiving a packet for the same parameters described in Figure 5.12. Measurements confirm expected simulated behavior. External loop-back wire delay of measurement setup was not used in simulation.

A slot being passed by a node because its header bit are all zeros is shown coming out of the high-speed serializer unit in Figure 5.16. A packet being stripped by the master because its header bits show that the ring has an invalid source address (header bit 25 is high) is shown in Figure 5.17. Normally, any packet is stripped by the sender after

successful receipt. When a packet passes by the master, it stamps header bit 25 high and the source of the packet resets this bit low. If no node on the ring is able to match its address to the source address specified in the header, this bit remains high. When it passes by the master a second time, it is then stripped. A packet can also be stripped if it has made more than 32 revolutions around the ring as specified in the hop count field to avoid deadlock of resources. The sender can also pre-program the hop count field before transmitting a packet onto the network enabling packet removal from the ring after fewer revolutions.

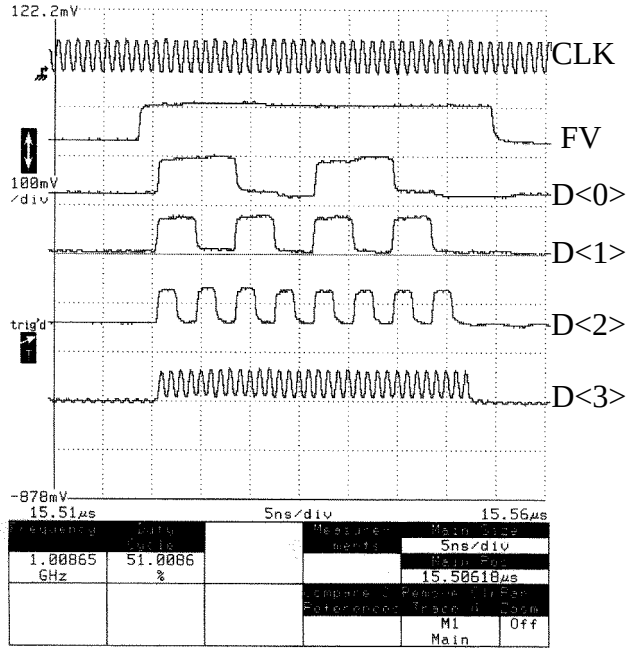


Figure 5.16: Measured serializer output viewed on a Tektronix 11801B digital sampling oscilloscope of packet passed by a node. Time is shown on the X-axis at 5 ns/division while signal amplitude on the Y-axis is 1 V/division attenuated by 20 dB. The signals shown are clock (CLK) at 1 GHz speed, frame control (FV), and four data lines D<0:3> corresponding to the data-path bits data<31:16>, attenuated by 20 dB. The packet is passed on since the first packet is empty (first word in the packet contains only zeros).

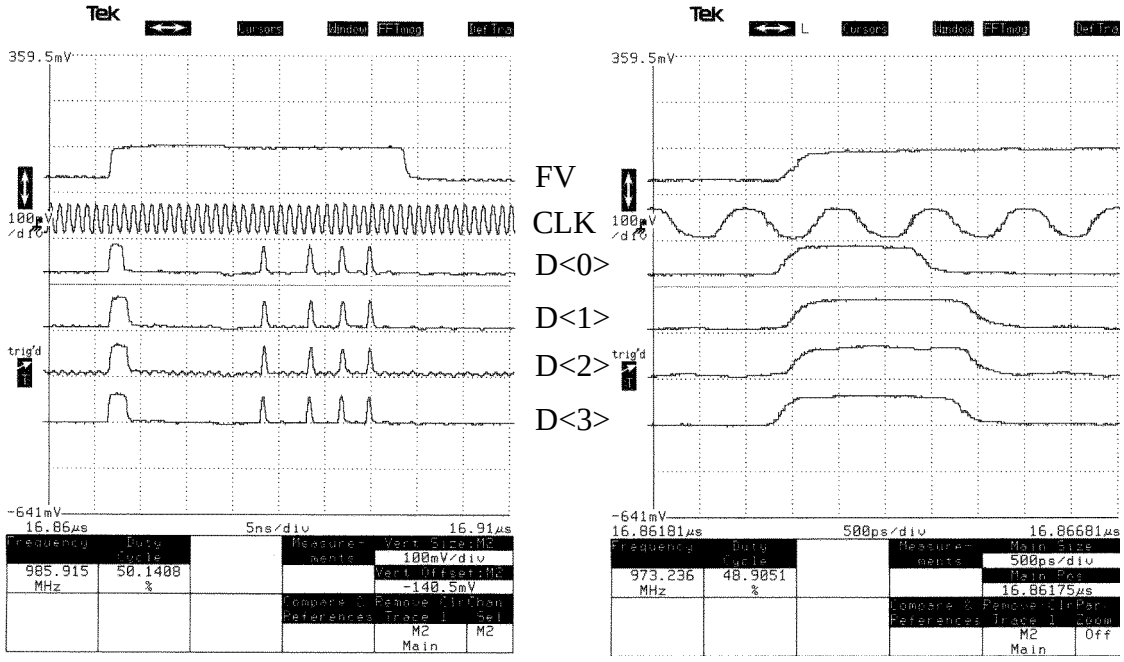


Figure 5.17: Measured serializer output viewed on a Tektronix 11801B digital sampling oscilloscope showing packet having been stripped by the master at 1 GHz clock frequency. The figure on the right is an enlarged view of the start of the packet. Signal amplitude on the Y-axis is 1 V/division attenuated by 20 dB. Time on the X-axis is 5 ns/division for the figure on the left and 500 ps/division for the figure on the right. As can be seen, the first three bits at the start of the packet in data line D<0>, corresponding to TTL lines TxDATA<30:28> are high while the fourth bit corresponding to TTL line TxDATA<31> or the slot/full empty bit is low indicating that the slot is empty. The packet is stripped because of an invalid source address specified in the header (header bit 25 is high).

5.2.6 TxFIFO

The TxFIFO interfaces with the host on one side and the LAC datapath on the other side. The TxFIFO memory is of size 1 KB and is constructed out of 4 banks of interleaved memory in 64 rows of dual-ported SRAM. Operation of the TxFIFO is discussed in detail in section 3.4 on page 53.

5.2.7 RxFIFO

The receive RxFIFO interfaces with the host on one side and the LAC datapath on the other side. The RxFIFO memory is of size 1 KB and is constructed out of 4 banks of interleaved memory in 64 rows of dual-ported SRAM. Operation of the RxFIFO is discussed in detail in section 3.4 on page 53.

5.2.8 Multiplexer (Mux) pipe stage

The principal function of the multiplexer (mux) pipe-stage is to select between incoming data from the ring and data from the TxFIFO awaiting transmission onto a slot. Header modifications such as setting the receive acknowledge bit and adjusting the hop count field are also performed in this pipe stage. Output from the multiplexer pipe-stage is fed into the RxFIFO as well as the smoother buffer module.

5.2.9 Smoother

The smoother buffer module constitutes the penultimate stage of the LAC datapath. It performs two functions. Firstly, it ensures that a minimum inter-slot idle symbol spacing is maintained due to the idle gap spacing being potentially altered by the elasticity buffer operation. Secondly, it provides buffer space adequate to hold the required number of slots on the ring.

The smoother is composed of a 1 KB FIFO constructed using dual-ported SRAM. Its write and read pointers are separated during initialization to provide buffer space adequate to hold slots in the ring network. When a short idle gap is detected by the read pointer, it stalls at an idle location until the idle gap size has been extended to the minimum allowed. The pointer separation for write and read pointers is thus temporarily increased by an

amount tracked by a “debit” counter. The state machine for the debit counter is shown in Figure 5.18 (a) and the various states are described in Table 5.3. The debit counter and the spacing between the write and read pointers need to be restored to the initial value when possible else there would be a pointer overflow in the smoother. The spacing is reduced whenever a long idle is detected at the write input port to the FIFO. Then, “credits” are issued corresponding to the extent to which the idle gap exceeds the minimum required idle spacing as is necessary to compensate for the debits incurred earlier. The state diagram for the state machine controlling the credit counter is shown in Figure 5.18 (b) and the states are described in Table 5.4. The spacing between write and read pointers is reduced by the write pointer stalling while writing idles into the buffer for such time as there are non-zero net debits (difference between debits and credits). When a frame is received at the write input port, write pointer again proceeds with writing data into the buffer and any remaining net debits will have to be compensated for the next time a long idle is received.

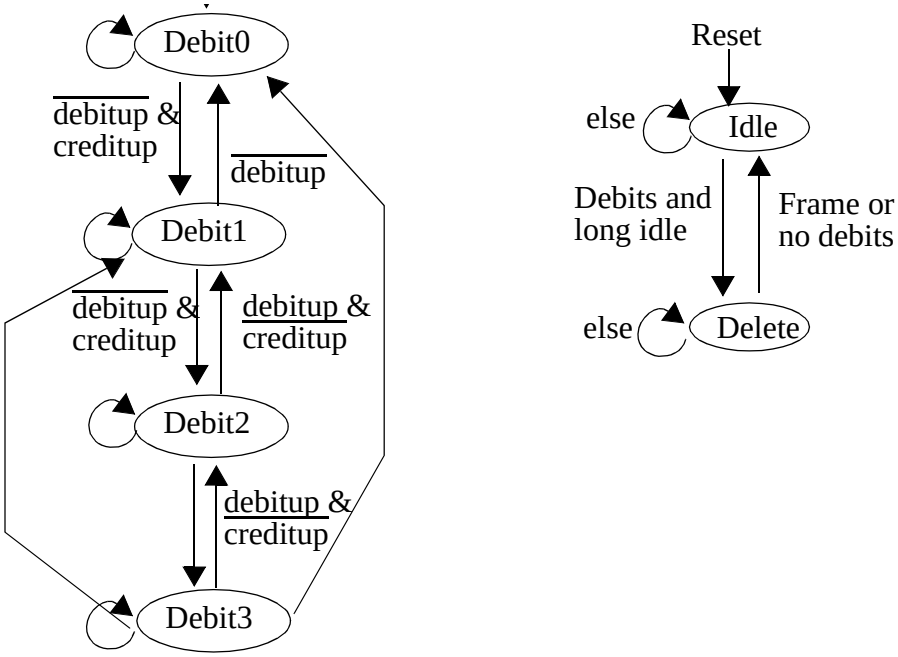


Figure 5.18: State machine for (a) debit state machine, which keeps track of number of times read pointer has stalled at an idle, designed as a two-bit Mealy state machine and (b) credit state machine in smoother, which keeps track of number of times write pointer has stalled while writing an incoming idle, designed as an output encoded Moore state machine

State	Description
Debit0	No debits incurred by read pointer
Debit1	One debit incurred
Debit2	Two debits incurred
Debit3	Three debits incurred

Table 5.3: Debit state machine of smoother

State	Description
Idle	Retain incoming idle gap
Delete	Delete idles from incoming idle gap

Table 5.4: credit state machine for smoother

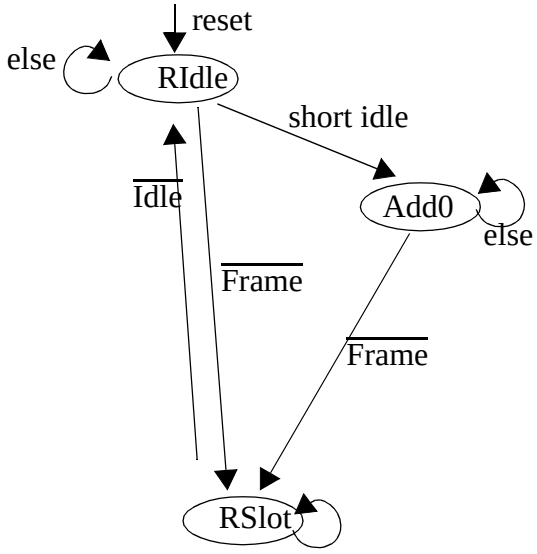


Figure 5.19: State machine for read pointer designed as a Mealy state machine. While reading, the read pointer either reads the slot (RSlot) or idle (RIdle) contents while incrementing its position, or stalls at an idle to restore minimum specified idle size (Add0).

State	Description
RIdle	Monitor and pass outgoing idle of smoother output
Rslot	Pass slot contents from smoother output

Table 5.5: Smoother read pointer state machine

State	Description
Add0	Add idles to a short smoother output idle gap to maintain minimum idle spacing as specified in smoother registers

Table 5.5: Smoother read pointer state machine

To optimize use of total network bandwidth, the residual idle gap (the idle gap between the last slot and the first slot on the ring) has to be minimized. To do so, it is necessary to count the total number of bits that the ring network can hold comprising of the bit capacity of each LAC and the frequency-dependent bit capacity of the interconnecting fiber medium. The bit capacity introduced by the smoother has to be adjusted based on this computation. This has not been implemented on our prototype implementation.

5.3 Digital logic clocking

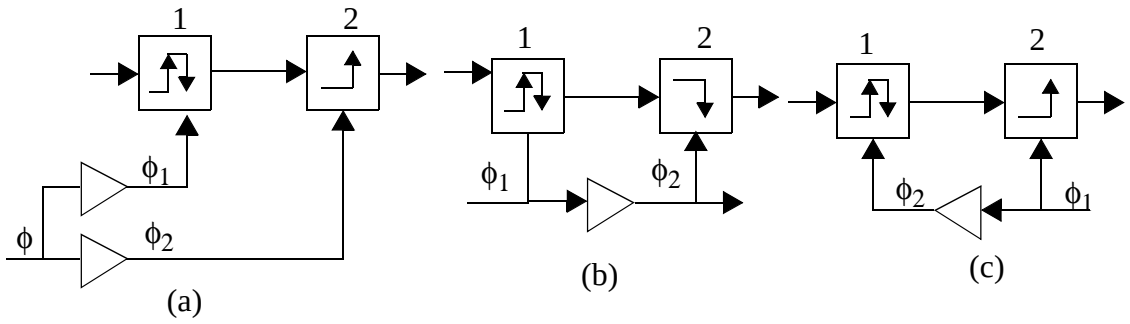


Figure 5.20: shows different clocking styles used for the digital logic. The schematic in (a) shows a matched-delay clocking style such as is used within the standard-cell region of the controller where data flows from an n-latch/p-latch to a p-latch/n-latch where clock delays for the two latches are matched. The schematic in (b) shows a forward-clocking style where data and clock flow in the same direction from an n-latch/p-latch to an n-latch/p-latch. The schematic in (c) shows a reverse-clocking style where data and clock flow in opposite directions from an n-latch/p-latch to a p-latch/n-latch.

There are three different clocking styles used for the digital logic as shown in Figure 5.20. They are (a) zero-skew or matched-delay clocking (b) forward clocking and (c) reverse clocking strategies.

In a zero-skew scheme, data clocked out of the first gate is connected to a second gate or latch. The clocks for either gate have negligible skew between each other in comparison with the logic gate delay. The timing diagram for a zero-skew clocking interface is shown in Figure 5.21 (A) and (B). The diagram in Figure 5.21 (A) shows data from a gate feeding a second gate operating off the opposite phase. The logic style used in the LAC is TSPC with typical gates and latches as shown in Figure 3.11. In TSPC logic,

data input to a logic gate has to be set up before the evaluation phase and remain stable during that phase to prevent corruption of data. Data input to a TSPC latch however can change in the evaluation phase of the latch so long as it meets setup constraints of the latch. Hence, the maximum speed of operation of the interface in Figure 5.21 (A) is given by inverse of the clock period,

$$t_{cycle} = 2(t_{d1} + t_{s2})$$

where t_{d1} refers to delay of the first gate including any delays due to output buffers while t_{s2} refers to setup time of the second gate. The maximum speed of operation for a logic gate feeding a latch as shown in Figure 5.21 (B) is however given by inverse of the clock period, t_{cycle} given as

$$t_{cycle} = t_{d1} + t_{s2}$$

In a forward-clocked scheme, data from a gate drives a second gate or latch whose clock is delayed with respect to the clock of the first gate. The timing diagram for such an interface is illustrated in Figure 5.21 (C) for a logic gate driving a latch, where latch clock is delayed by an amount t_{buf} corresponding to a clock buffer delay, the constraint for correct operation is

$$t_{buf} - t_{s1} < t_{cycle}/2 + t_{d1}$$

so that there is no clock and data race condition.

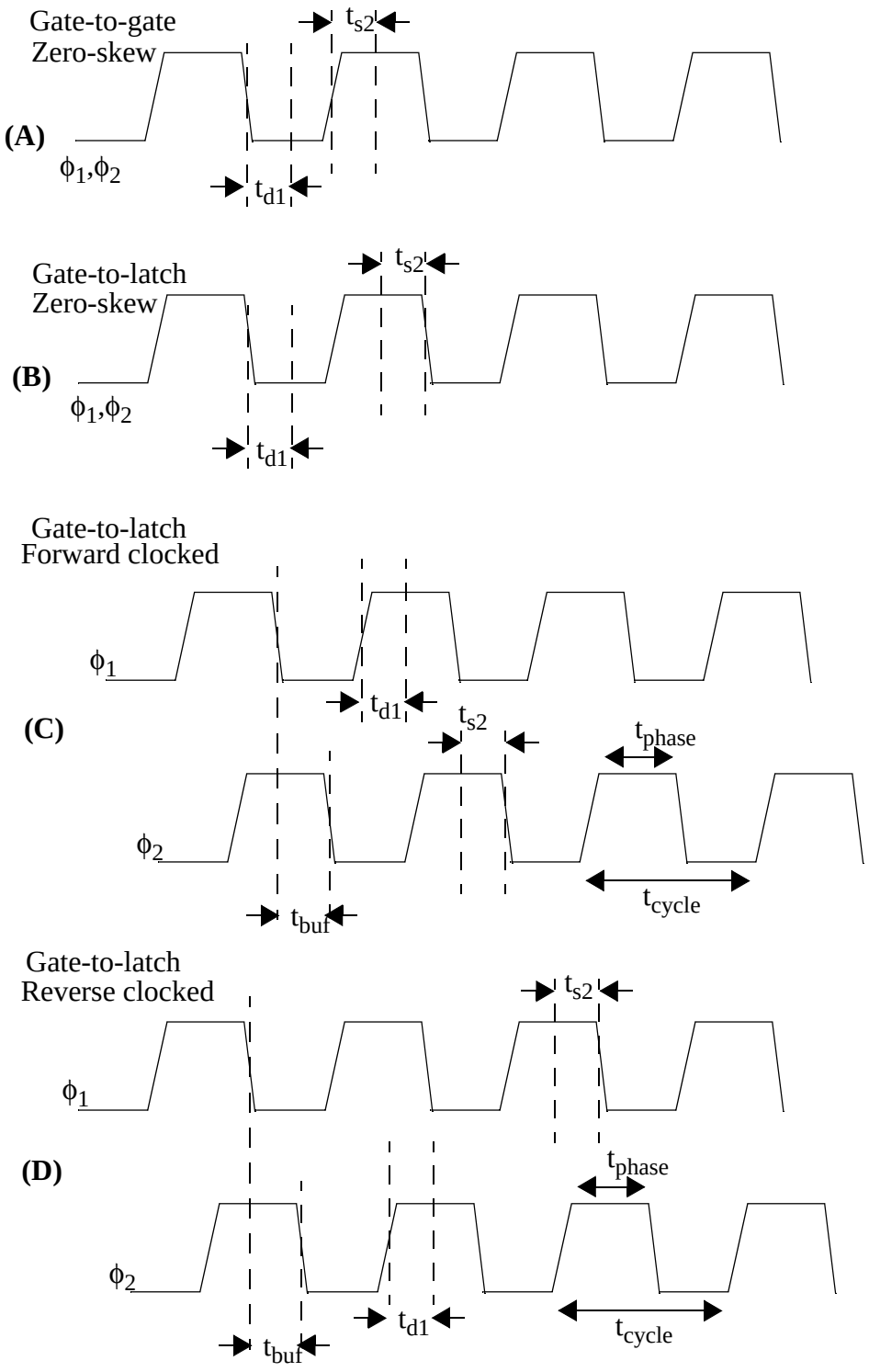
In a reverse-clocked scheme, clock and data flow in opposite directions as shown in Figure 5.20 (a). For cascaded gates operating on different phases of clock (such as n-latch feeding a p-latch or vice versa) a buffer delay is introduced for clocks of the two gates, with a timing diagram as shown in Figure 5.21 (D). The maximum speed of operation is given by the inverse of the relation,

$t_{cycle} = t_{d1} + t_{s2} + t_{buf}$

Hence, lower speeds are achievable using a reverse-clocked scheme as would be achievable in a zero-skew clocking scheme. The trade-off however is one of simplicity, since enforcing a deterministic delay is simpler to implement than a scheme with matched delay where extensive characterization has to be performed for variations in operating conditions such as process, temperature, capacitive load and voltage.

There are two principal clock domains in the digital logic. The first clock domain corresponds to the clock obtained by dividing the incoming network clock to extract a half-speed clock. The aligner and the write port of the estore clocks are derived from this clock. The second clock domain corresponds to a digital clock derived locally from a half-speed version of the high-speed transmitted clock. The rest of the LAC is clocked using this clock. Clock and data flow directions for the LAC are depicted in Figure 5.22. The deserializer-aligner, aligner-estore and multiplexer pipe-stage - RxFIFO interfaces are forward clocked. All other interfaces are reverse clocked as shown in the figure. Within each memory block controller and pipe stage, matched buffer-delay clocking is used.

Figure 5.21: The following figure is an illustration of different clocking interfaces. The figure (A) shows a zero-skew scheme where data output of the first gate is connected to a second gate, the clocks for the two gates showing negligible delay with respect to each other in comparison with the logic gate delay. The figure (B) shows the same clock delay configuration; however, data from the first gate is connected to a TSPC latch instead of a logic gate. The figure (C) shows data from the first gate connected to a TSPC latch where clock for the latch is delayed in comparison with the clock for the first gate. The figure (D) shows a reverse-clocked interface with clock for the first gate is delayed in comparison with the clock for the following TSPC latch. The parameter t_{d1} is the delay of the first gate, the parameter t_{s2} is the setup time of the second gate, t_{buf} is the buffer delay, t_{phase} is a clock phase time and t_{cycle} is a clock period.


Gate-to-gate
Zero-skew
(A)
ϕ_1,ϕ_2
t_{s2}
t_{d1}
Gate-to-latch
Zero-skew
(B)
ϕ_1,ϕ_2
t_{s2}
t_{d1}
Gate-to-latch
Forward clocked
(C)
ϕ_1
ϕ_2
t_{d1}
t_{s2}
t_{phase}
t_{buf}
t_{cycle}
Gate-to-latch
Reverse clocked
(D)
ϕ_1
ϕ_2
t_{s2}
t_{phase}
t_{d1}
t_{buf}
t_{cycle}

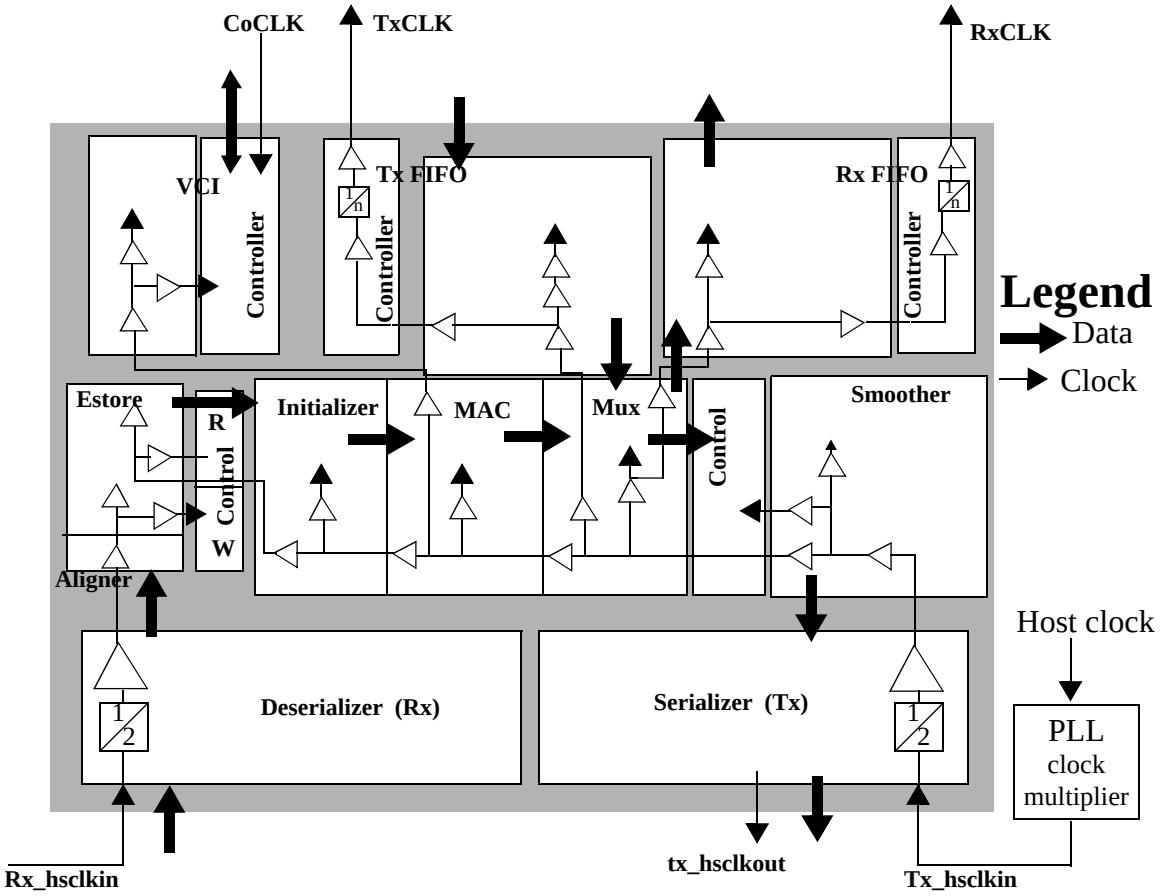


Figure 5.22: Illustration of the clock and data flow directions for digital logic circuitry. The received network clock, Rx_hsclkin latches data into the deserializer and a divided clock is used to derive the clocks for the aligner and the write port of the estore. The rest of the LAC digital circuitry is clocked from a clock obtained by dividing the serializer's high-speed input clock, Tx_hsclkin. This clock originates in the smoother and is distributed to the rest of the chip with buffers inserted periodically in the datapath. The deserializer-aligner, aligner-estore and multiplexer pipe-stage - RxFIFO interfaces are forward clocked. All other interfaces are reverse clocked as shown in the figure. A reverse clocking strategy is a simpler scheme of clock distribution as opposed to a zero-skew clocking scheme. Within the memory block controllers and the pipe stages, a zero-skew clocking is used. The thick solid arrows show direction of flow of data. The signals TxCLK, RxCLK and CoCLK are used to clock the TTL interface with the host.

The use of a reverse clocking strategy for global clock combined with the use of latches to re-time data across datapath blocks as opposed to a zero-skew global clocking strategy is an effective and simple method of achieving over 500 MHz digital operation. It is useful to compare this clocking strategy with the zero-skew global clock distribution implemented in the 600 MHz Alpha microprocessor [92] in six-layer metal 0.35 μm technology. This chip uses a global clock distributed from the center of the die to distributed global clock drivers in buffered H, X and RC trees. To ensure that the global clock distribution satisfied the timing requirements, a thorough analysis of skew and edge-rate under worst-case operating conditions needed to be performed with simulation-based characterization of global clocks to account for variations in process, temperature, voltage and capacitive load.

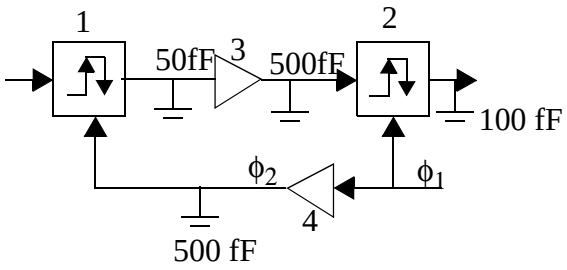


Figure 5.23: Schematic of typical interface across pipe blocks in LAC datapath. The timing constraints imposed by this reverse-clocked interface limits achievable data rate in the LAC to nearly 500 MHz digital operation.

The reverse-clocked interface for global clock distribution to various datapath blocks in the LAC limits the achievable data rate as will be demonstrated in measurements in section 5.7 on page 142. A typical such interface is shown in Figure 5.23. The simulated

delays using slow libraries at 85^0 C at 3V power-rail swing are as follows: clock-to-output logic gate delay of the first latch of nearly 0.5 ns combined with output buffer delay of 0.45 ns, clock buffer delay of 0.55 ns and a setup time for the second latch of 0.5 ns. This gives a net reverse-clock interface delay of nearly 2 ns, limiting the block interface speed to 500 MHz digital frequency.

5.4 Finite state machine design

Control operations may be performed in an organized manner using finite state machines. The operation is represented by means of a state transition graph with labelled states and conditions causing transition from one state to another. Finite state machines (FSM) are of two kinds - Mealy state machines [90] and Moore state machines [91] schematically illustrated in Figure 5.24. In Mealy state machines, the outputs depend on both the current state values stored in state registers as well as the inputs to the state machine. In Moore state machines, the outputs depend on the current state only. Mealy state machines have less delay (such as fewer pipeline stages) to produce outputs and quite often use fewer states. Functionally equivalent Moore state machines on the other hand may use more states and feature higher output delays. However, they are simpler to implement and hence yield higher speeds of operation.

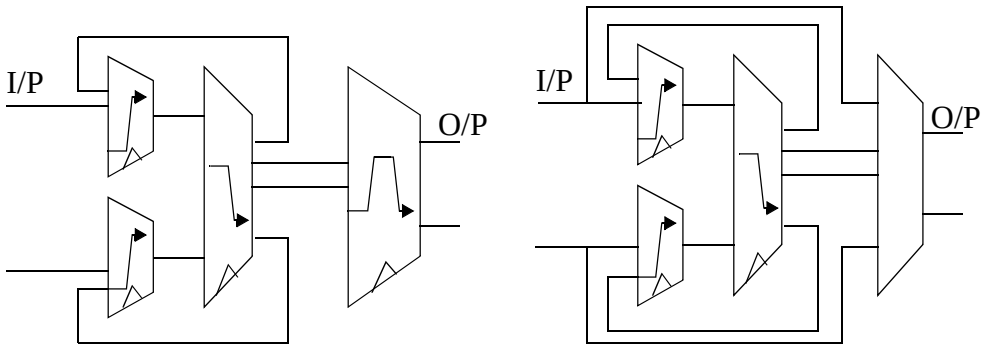


Figure 5.24: Schematic of (a) Moore state machine and (b) Mealy state machine. In a Moore state machine, outputs (O/P) are derived from the state variables only. In a Mealy state machine, outputs are derived from the state variables as well as inputs (I/P) to the state machine. Moore state machines may use more states, the outputs may experience more pipeline delay, but result in simpler output implementation enabling higher speed of operation.

5.5 Design techniques for high-speed datapath design in LAC

Digital logic operation at over 500 MHz is achieved in the LAC by various techniques some of which are listed below:

- Use of dynamic true single phase clocking (TSPC) logic as opposed to static logic. Dynamic logic uses fewer transistors than static logic and hence has smaller parasitic load capacitances. Secondly, the use of a single-phase clock as opposed to two-phase clock does not result in any clocking overhead and allows for high-speed clocking.
- Use of n-logic gates with inverted clocks to realize p-logic gates. Since mobility of n-transistors is higher, an n-transistor is faster than a p-transistor of the same dimensions. A smaller n-logic gate results in lower clock load to achieve the same speed.

- Use of a deeply pipelined datapath design with single-logic gate delay per pipeline stage.
- Use of Moore state machines to design finite state machines wherever possible instead of Mealy state machines. Though Mealy state machines result in smaller pipeline delay, since outputs depend on the state variables as well as state transition inputs, the logic used to realize them is more complex (since a Boolean table for realizing a certain function depending on n variables is of size 2^n thus increasing table size exponentially with each additional variable). Moore state machines, where the output depends only on the current state variables and not on the input variables can hence be used to obtain higher speeds due to the lower design complexity at the cost of increased pipeline delay.
- All state machines are designed using standard cells as opposed to using a PLA-based approach. This is achieved by minimizing state machine implementation complexity as described above so that conventional standard cells may suffice and also by using state encoding techniques such as an output encoding scheme used in designing the medium access control state machine. By designing using standard cells one bypasses overheads that would be encountered in moving signals across two different clock domains - the standard cell region and the PLA regime.

- Exploitation of architectural features of blocks to maximize their speed of operation. For example, since FIFOs have a deterministic sequence of access, as much as a whole clock cycle can be allowed for precharging and evaluation as opposed to a single clock phase available for a RAM to which accesses are random. Thus, bitline swings can increased maximizing sensitivity of memory read port sense-amplifier output stage.

Some of the techniques described above are illustrated in the design of a fast eight-bit counter. A logical schematic of a one-bit slice of the counter is shown in Figure 5.25. The counter uses only n-logic gates or inverted clock n-logic XOR and XNOR gates.

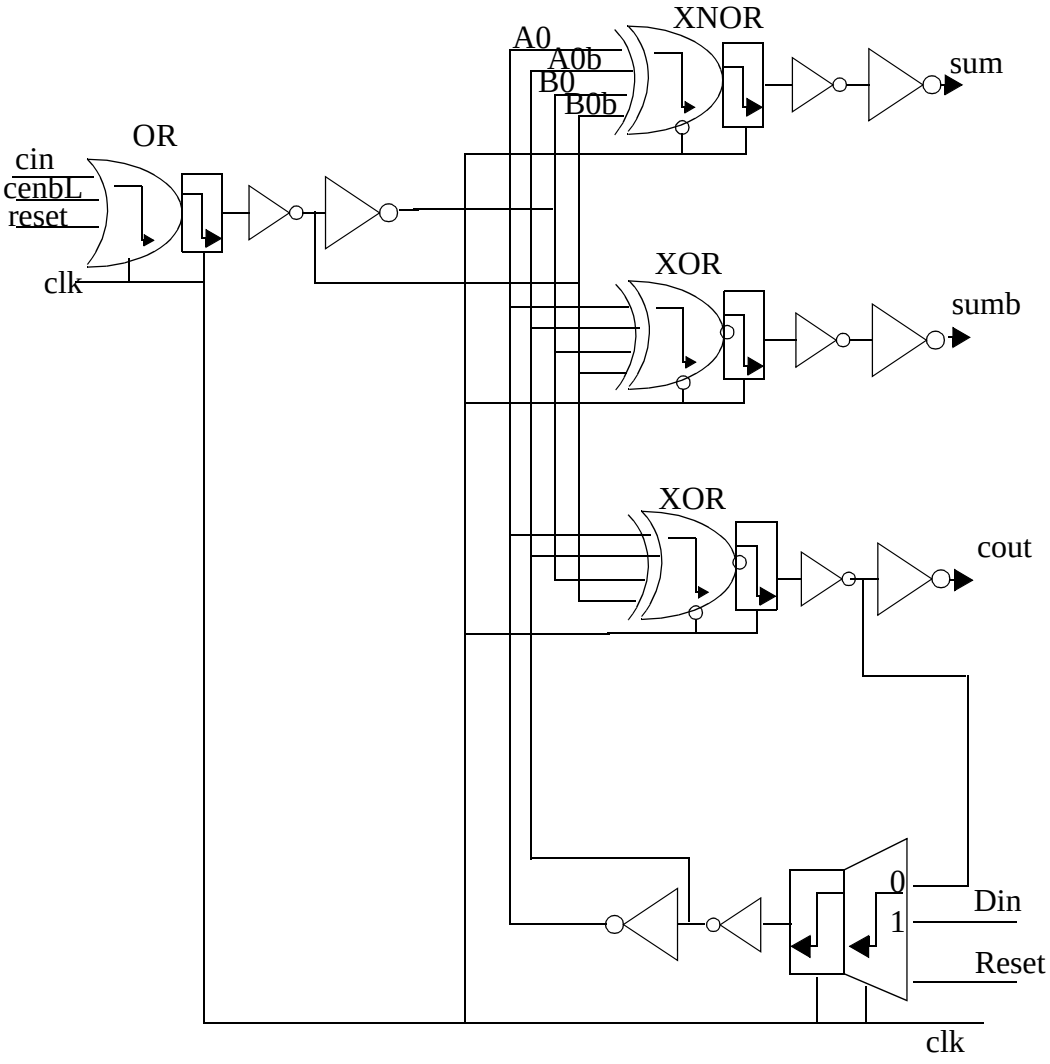


Figure 5.25: Design of a one-bit slice of a fast eight-bit counter. The figure shows a counter slice cell with separate outputs for the sum (sum), carry (cout) and inverted sum (sumb) bits. High-speed is achieved by using multiple parallel output units which optimizes place-and-route of circuitry for simultaneously realizing eight-bit counters and performing counter output comparisons. Each of the counter outputs drives less than 500 fF in total load. P-logic gates are realized by locally inverting clock of n-logic gates. Other inputs to the counter slice are reset (Reset), clock (clk), counter enable (cenbL) and carry input (cin).

5.6 Layout

The LAC was fabricated in a 0.5 μm CMOS process provided by MOSIS using the HP-AMOS14TB process. This is an n-well process with 3 metal and one poly layer and operates with a nominal 3.6 volt power supply. A photograph of the die is shown in Figure 5.27. The size of the die is 10.2 mm x 7.3 mm and is mounted in a ceramic QFP 244-pin package designed at USC for this project. Analog, digital and TTL power are isolated on the chip. There are 113 pads for the digital logic and TTL power and 116 pads for digital and TTL ground. There are 31 pads for analog power and 36 pads for analog ground. The number of transistors on the chip is close to 380,000. The ground connections for the digital circuitry and the high-speed circuitry are also separated. The substrate connection for the digital circuitry clock buffers is isolated from the ground connection for the rest of the digital circuitry. Substrate contacts and guard rings are placed to protect sensitive memory circuitry such as sense-amplifiers from substrate-coupled noise. A guard ring is also placed around the high-speed circuitry to isolate it from the rest of the digital circuitry. In the high-speed circuitry, each of the ten channels has a distinct power and ground connection to avoid power supply degradation due to resistive drops. Decoupling capacitors between power and ground were placed on the chip wherever space allowed. A standard cell-based approach is used for the digital logic circuitry. A custom layout of the high-speed interface circuitry and the FIFO memory blocks was performed. Place-and-route of the transmit and receive controllers and all top-level routing was performed using the Cadence Cell Ensemble tool and the extracted layout was simulated using HSpice at a junction temperature of 80^0 C.

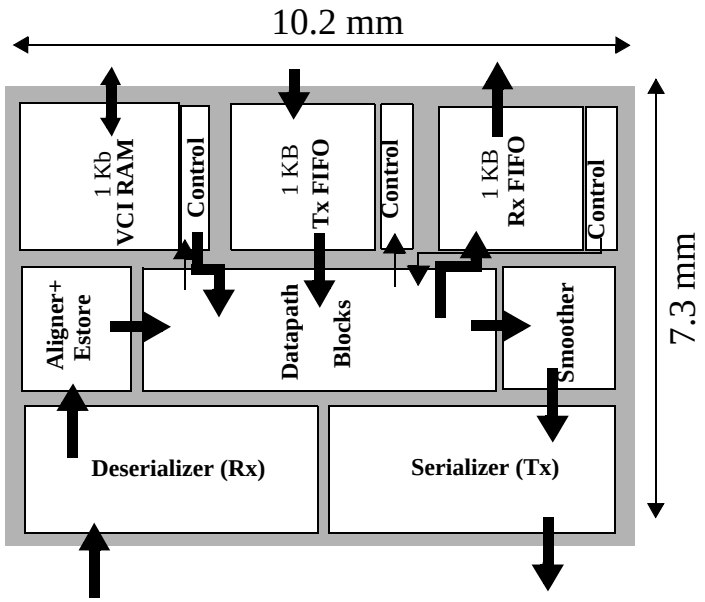


Figure 5.26: Floorplan of the LAC showing the logical placement of various blocks. The LAC measures 10.2 mm x 7.3 mm. Network data is received by the deserializer, aligned by the aligner and synchronized by the estore. It passes through the datapath blocks and smoother and exits through the serializer. Packets are transmitted from the TxFIFO, received into the RxFIFO and address lookups are performed using an address table stored in the VCI RAM.

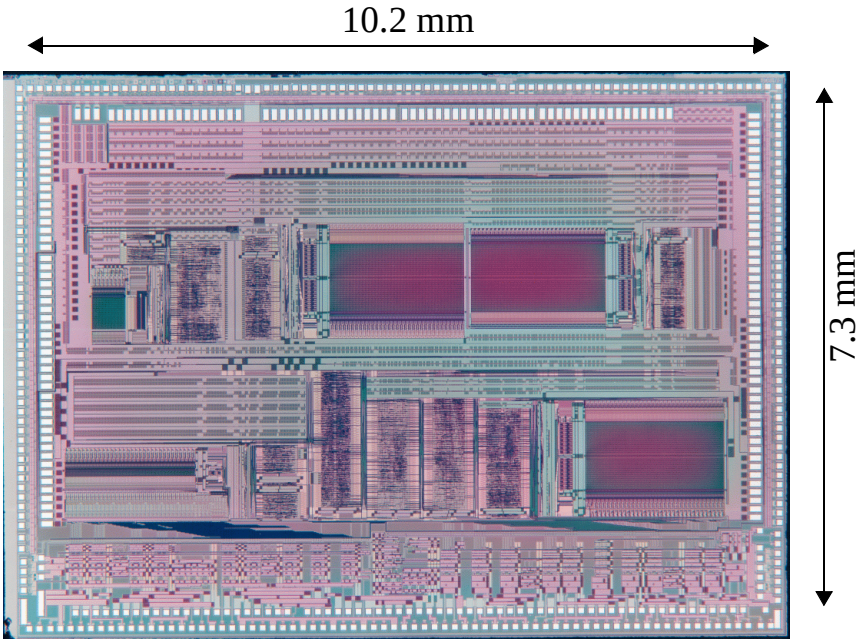


Figure 5.27: Die photograph of LAC. The LAC was fabricated in 0.5 µm 3-layer metal HP-AMOS14TB CMOS process. The design was submitted on 8/17/2000 and the fabricated chip was received on 11/10/2000. It measures 10.2 mm x 7.3 mm and contains nearly 380,000 total transistors.

Parameter	Value
Process technology	0.5 µm CMOS (HP-AMOS14TB)
Tox	9.7 nm
Interconnect	3-level Aluminum
Substrate	P-epi with n-well
Supply voltage	3.6 V (nominal)
Max data rate	2.1 Gb/s per signal line
Transistor count	~ 380,000

Table 5.6: Summary of LAC chip features

Parameter	Value
Peak power	10.5 watts
Die size	10.2 mm x 7.3 mm

Table 5.6: Summary of LAC chip features

5.7 Test setup and measurement results

The measurement setup used is shown in Figure 5.30 and Figure 5.31. Figure 5.30 shows a single node connected in loop-back configuration using 24 inches long electrical RG4U microcoax copper cables with low-cost 3M Shielded Controlled Impedance (SCI) stake assemblies. The cables with stake assembly have an insertion loss of 2 dB at 2.5 GHz and a -3 dB bandwidth of nearly 3.5 GHz as shown in Figure 5.28. When the cable is inserted into an 8-mil wide 50 ohm FR-4 printed circuit trace that is 8 cm long, the insertion loss as seen from Figure 5.29 is 4 dB at 2.5 GHz with a 3 dB bandwidth of nearly 1.8 GHz. For two back-to-back boards connected using a copper cable, the total FR-4 trace length is on the order of 10 inches. As will be seen in the following chapter from Figure 6.13, the insertion loss due to 9.5 inches of FR-4 trace alone at 2.5 GHz is 3 dB. Hence, total loss for the two boards inclusive of QFP package loss can be as high as 6-7 dB at 2.5 GHz and -3 dB bandwidth less than a GHz. Figure 5.31 shows two nodes connected back-to-back using electrical copper cables. Test vectors from the TTL inputs were supplied using an HP 16500B logic analysis system. An HP 70841B pattern

generator and an HP 70842B error detector supply the differential input clock which can vary from 50 MHz to 1.3 GHz. A Tektronix 11801B digital sampling oscilloscope is used to view the high-speed signal outputs of the serializer.

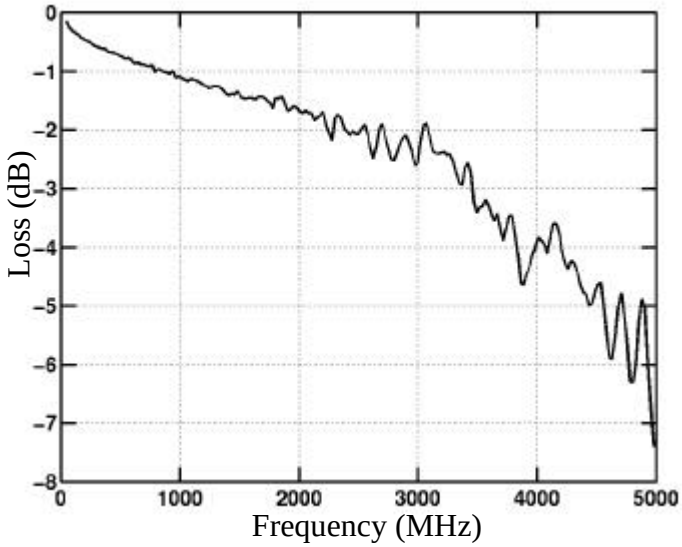


Figure 5.28: Insertion loss measurement of RG178 microcoax copper cables with 3M SCI stake assembly. The cable is 24" long and has an insertion loss of 2 dB at 2.5 GHz and a 3 dB bandwidth of nearly 3.5 GHz.

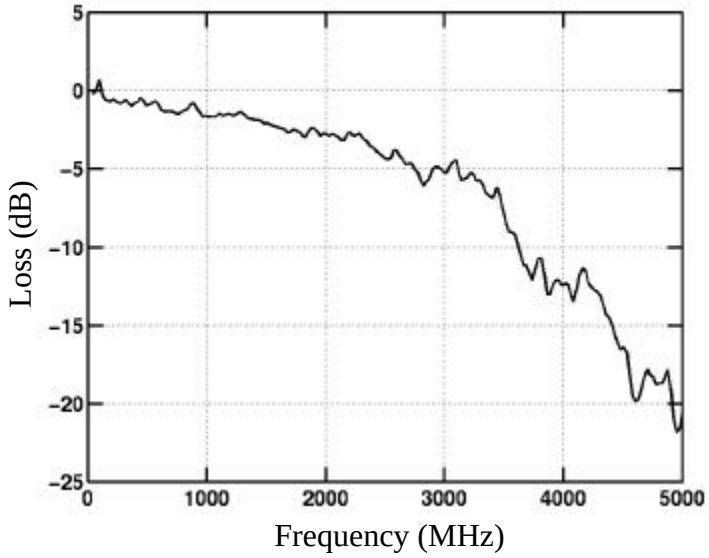


Figure 5.29: Insertion loss measurement of 24" of RG178 microcoax cable with 3M SCI stake assemblies and 8 cm of 8 mil wide 50 ohm FR-4 printed circuit trace. At 2.5 GHz, the insertion loss is nearly 4 dB with a -3 dB bandwidth of near 1.8 GHz.

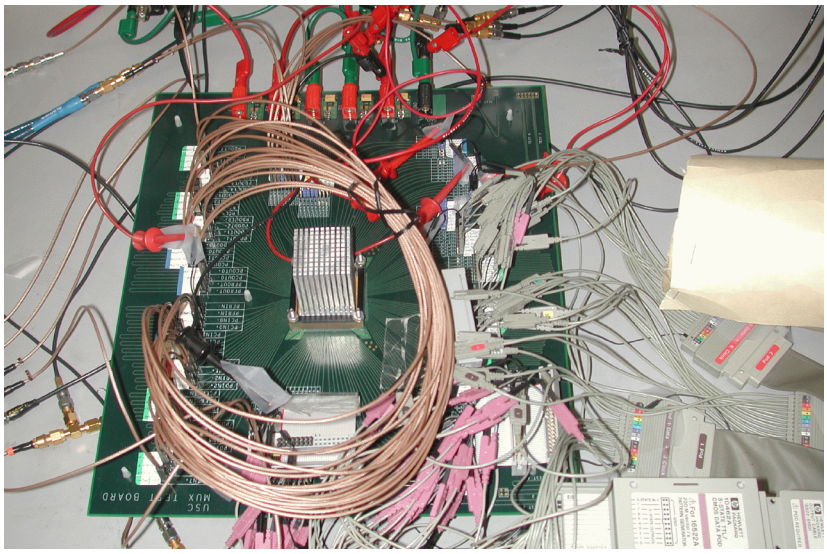


Figure 5.30: Photograph of packaged LAC mounted on an FR-4 printed circuit board and connected in loop-back configuration using 24" long RG178 micro-coax copper electrical cables with low-cost 3M SCI stake assemblies. A heat sink is mounted on the package using thermal adhesive. An HP 16500B logic analysis system supplies the TTL inputs to the chip and monitors the TTL outputs.

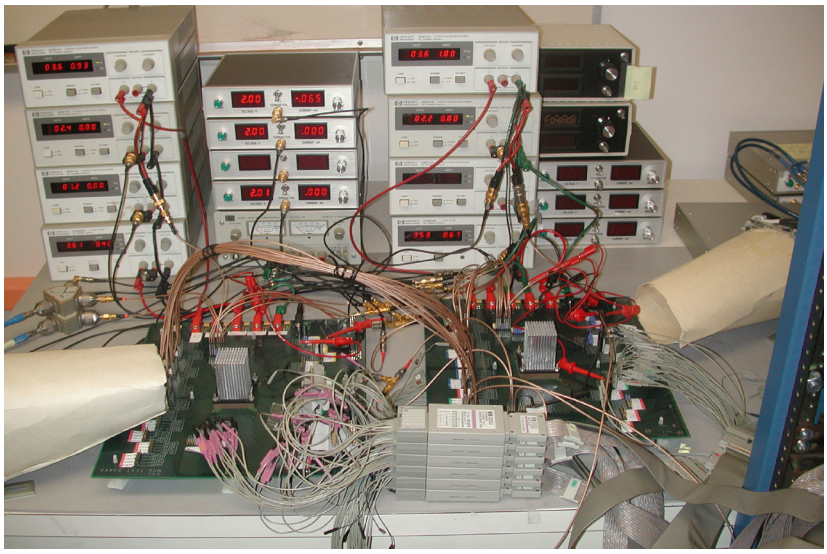


Figure 5.31: Photograph of two LAC chips connected in back-to-back configuration using 24" of RG178 copper microcoax electrical cables with 3M SCI stake assemblies. The power supplies used are visible in the background.

The high-speed outputs viewed on a digital sampling oscilloscope featuring the clock, frame control line and four data lines are shown in Figure 5.32. The figure on the left shows output data lines functional at 2 Gb/s per signal line at a supply voltage of 3.6 V. This results in an aggregate network data rate of 16 Gb/s. The datapath was tested to be functional at up to 2.1 Gb/s per signal line (or a digital logic clock speed of 525 MHz). Beyond this frequency, there are errors which are most likely due to retiming across blocks (elasticity buffer, initializer, medium access control, multiplexer and smoother blocks) as seen in circuit simulation. When supply voltage is raised to 4.15 V however, correct operation is again observed at up to 2.5 Gb/s per signal line. Serializer/deserializer blocks are not functional at higher speeds at this supply voltage.

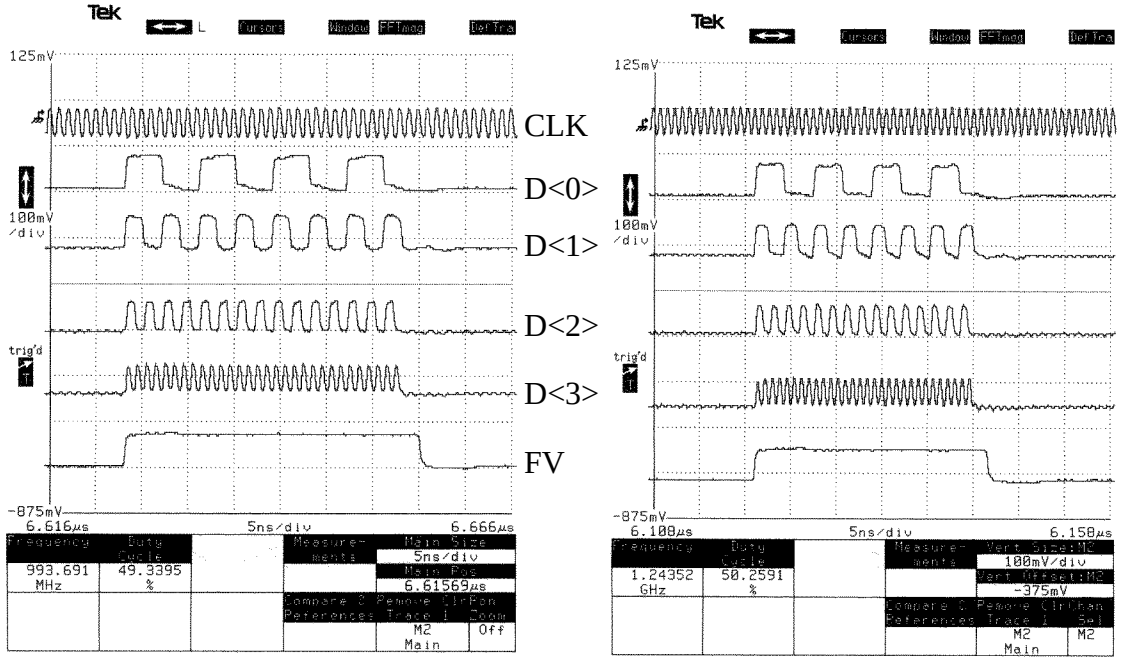


Figure 5.32: Measured high-speed output showing clock (CLK) at 1 GHz speed, frame (FV) and four data lines D<0:3> at 2 Gb/s per signal line for the figure on the left at 3.6 V supply voltage. The figure on the right shows the same signals with clock at 1.25 GHz and data lines at 2.5 Gb/s per signal line at 4.15 V supply voltage.

The measured power consumed by the LAC is plotted in Figure 5.33. As expected from switching activity calculations, it shows a nearly linear scaling with operating digital frequency. The current drawn at DC is nearly 0.34 A. In the P2P chip, there is DC power consumption of nearly 0.13 A. In this section, we now investigate sources of these currents flowing through various memory components as seen from circuit simulations. From simulations, there is a DC power consumption however of nearly 300 µA per memory cell along a row activated by wordline in an inactive bank (precharging ON). This results in a static current consumption of nearly 30 mA for the three 33 bits-wide

banks of memory as seen from simulation using the slow libraries at 85^0 C and 3 V power rail swing. An additional simulated RMS current of nearly 300 μA is consumed in each sense-amplifier shown in Figure 5.35 due to DC current flowing to ground. This is because of the action of the transistor equalizing the arms of the cross-coupled inverter pair resulting in a simulated current consumption of nearly 40 mA for a 132-bit wide sense-amplifier. Hence, the total simulated static current dissipation in the memory blocks is nearly 70 mA for each memory block. In the P2P, there are two such memory blocks - one each in the transmit and receive buffers, for a total simulated static current drawn of nearly 0.14 A, showing good agreement with measured value of nearly 0.13 A. Short-circuit current due to finite pulse rise times in the controller logic gate inputs have not been considered. The differences may be due to simulated conditions (power supply of 3.6 V, operating temperature, slow library corner etc.). In the LAC, there are four FIFO blocks - the transmit, receive, smoother and elasticity buffers. This results in a simulated static current drawn of roughly 0.3 A. In the VCI, there are 32 columns which result in simulated sense-amplifier current of nearly 10 mA. Simulated static current to ground through memory cell during precharging of 32-columns accounts for a further 8 mA of current leading to a net simulated static current in VCI of 18 mA. Hence, simulated total static current in LAC is nearly 0.32A showing good agreement with measured value of 0.34A.

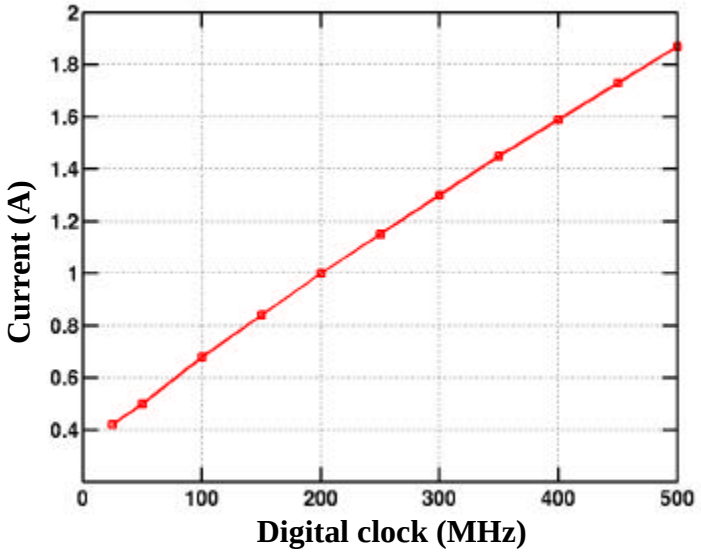


Figure 5.33: Measured digital logic power consumption versus digital clock frequency at a power supply of 3.6 V. Peak digital power consumption is 6.73 W at 500 MHz digital clock frequency. TTL clock runs at a sixteenth of the digital clock frequency. The dominant source of static power consumption is in the memory blocks through memory cells activated by a wordline and in the cross-coupled inverter pair of the sense-amplifiers.

Power dissipation in the sense-amplifier can be reduced by turning on the tail current source only prior to sensing operation. To reduce power dissipation in the memory cells, one possible scheme would be to precharge column bitlines only for one clock cycle prior to bitline evaluation as opposed to having bitlines always precharging when the bank is not operational as is implemented currently. Static power reduction in the memory requires further exploration.

As seen, the power dissipated in the digital logic varies linearly with frequency due to its contribution coming mainly from dynamic switching activity. In other words power dissipation on the LAC is due to total switching capacitance (approximately 860 nf from

LAC measurements). In CMOS design, there is a trade-off between circuit power consumption and delay (or achievable data rate) and hence it is desired to minimize the power-delay product.

Using the same CMOS process, more complex systems such as microprocessors achieve lower clock speeds [99]. For example, the Intel Pentium Pro processor achieves a clock speed of 150 MHz at peak power consumption of nearly 30 W for a die size of 306 mm^2 [1]. Other processors such as the Cyrix/IBM 6x86 processor (clock speed of 133 MHz) [3] and the AMD Am486 processor (clock speed of 120 MHz) [2] also achieve less than 150 MHz clock speeds for the same fine-line dimensions.

To gain a rough measure of the efficacy of this design, we use the power-delay product normalized to the area of the chip (as being representative of the load capacitances in the chip) to compare the LAC against the Intel Pentium Pro processor implemented in the same technology. The goal is to minimize this value or to maximize its reciprocal. The Pentium Pro processor with a die size of 306 mm^2 and 5.5 million transistors consumes nearly 30 W of power to achieve 150 MHz clock speeds. This gives an area/power-delay figure of 1530 mm^2 MHz/W. The LAC measuring 10 mm x 7 mm in die size with 400,000 total transistors consists of a front-end serializer-deserializer core and a back-end digital logic. The height of the serializer/deserializer core is approximately 1 mm. The remainder of the chip of close to 6 mm is dedicated to the digital logic. The area/power-delay figure for the LAC digital logic occupying area of 10 mm x 6 mm is hence nearly 4456 mm^2 MHz/W demonstrating at the very least that this is a commercially viable design point.

The P2P implementation demonstrates that achievable data rate using this design point is limited by the achievable data rate in FIFO buffers, the limitation arising from the bitline swing to the sense-amplifier inputs. Crossbar switch-based networks have large buffer requirements due to contention. In ring networks, contention is avoided because ring nodes cannot transmit unless there is a free slot. Hence, ring networks are more conducive to achieving higher data rates because buffers can be designed to run at higher clock speeds. However, larger buffers than the current implementation will be needed for larger protocol data units such as the 1500 bytes used in Gigabit Ethernet. But more importantly, receive buffer size on the LAC will have to be increased to allow for multiple nodes to communicate to a single node since host I/O interface speeds do not increase at the same rate as on-chip clock speeds.

Larger buffers are typically implemented by partitioning into smaller blocks since the height of the block determines the bitline delay. However, there is a trade-off between overall memory area and partition size. The 200 MHz Digital Alpha implemented in 0.8 μm CMOS used 64 rows [101] for the cache design, though 128 rows have been used in the 450 MHz 0.18 μm CMOS-based PowerPC microprocessor as well [102]. Theoretical analysis of SRAM delay scaling with size [100] in 0.25 μm CMOS indicates that the optimal height of an SRAM partition to minimize delay is 32 rows. In generations of 0.13 μm CMOS and below, subthreshold leakage currents from access transistors of memory cells on a bitline will prevent the bitline from discharging through the active memory cell and limit the block height to a maximum of 64 rows [107].

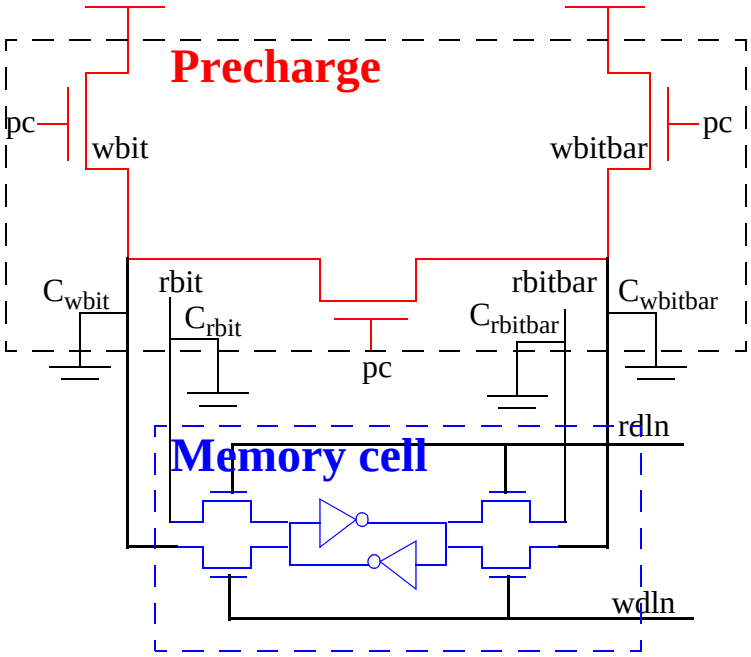


Figure 5.34: Schematic of precharge circuitry enabled by precharge control line (pc) in combination with write bitlines (wbit and wbitbar), read bitlines (rbit and rbitbar) and memory cell. There is static power dissipation through the memory cell activated by a write wordline (wdln) or read wordline (rdln).

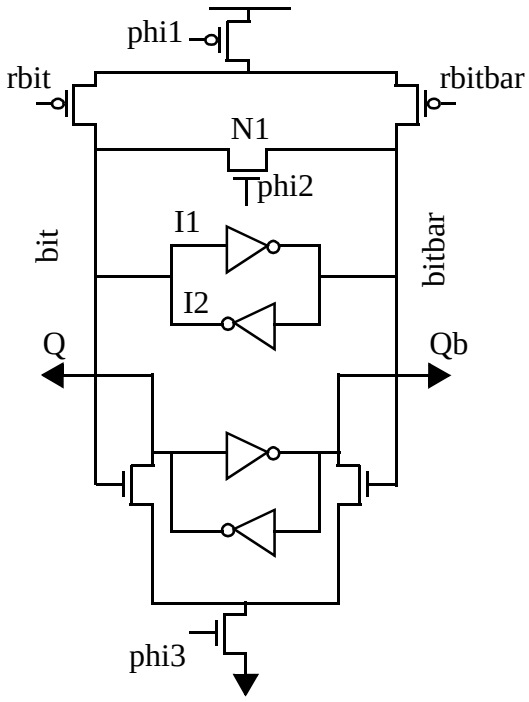
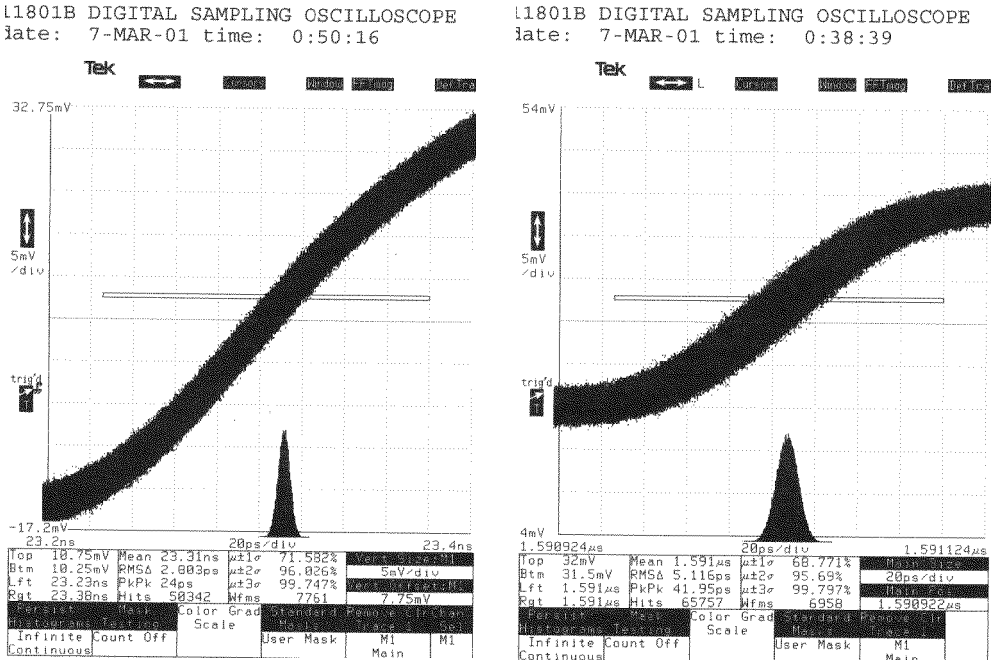


Figure 5.35: Circuit schematic of the sense-amplifier used to sense read bitline memory outputs, rbit and rbitbar. When phi2 is high, the transistor N1 equalizes the arms of the cross-coupled pair bit and bitbar. In the LAC, phi1 is always low. Hence, there is DC current flowing through the n-transistors of inverters I1 and I2 with simulated value of nearly 300 μA per sense-amplifier at 85^0 C and 3 V power-rail swing.

Serializer output clock jitter measurement measured for over 65000 hits (over 5 minutes time) is shown in Figure 5.36. The jitter with reference to negative clock output as seen from Figure 5.36 (a) has an RMS value of less than 3 ps with a peak-to-peak value of 24 ps representing jitter from the oscilloscope box. The source-referenced jitter as seen from Figure 5.36 (b) has an RMS value of nearly 6 ps with a peak-to-peak value of less than 42 ps. Hence, RMS clock jitter is not more than 3 ps. As seen from Figure 5.37, the dominant contribution of clock jitter is the clock source. The LAC clock output is at least

as good as the source clock. From the relation, frequency f = 1/T, where T is the clock period we can derive the relation $\sigma_f = \sigma_t/T^2$, σ_f is the standard deviation in frequency and σ_t is the standard deviation of jitter. At 1 GHz clock, the frequency variation over 20 minutes is 3 kHz. As seen from the figure, a 3σ value of 9 kHz at 1 GHz clock frequency provides information about frequency variation to within 99.8% accuracy. Peak-to-peak jitter is less than 42 ps which results in an equivalent peak-to-peak frequency variation of nearly 42 kHz.

Jitter over long time scales is not a significant issue for parallel links in terms of clock recovery, but only in terms of the representative frequency variation that it represents. It is only the cycle-to-cycle jitter or jitter over short time scales that is significant for latching data at the receiver. Jitter considerations for latching received data will however be significant for equivalent serial links of 16 Gb/s (or bit time of 31.25 ps) data rates as also parallel fiber links realized in CMOS processes below 0.1 μm technology.



(a) (b)

Figure 5.36: Jitter measurements for over 65000 hits on serializer output clock at 1 GHz high-speed clock speed. The figure (a) shows positive clock jitter with reference to negative clock output of differential pair and the figure (b) shows clock jitter with reference to BERT clock source. RMS jitter is less than 6 ps in either case with approximately 2.8 ps resulting from jitter from the box.

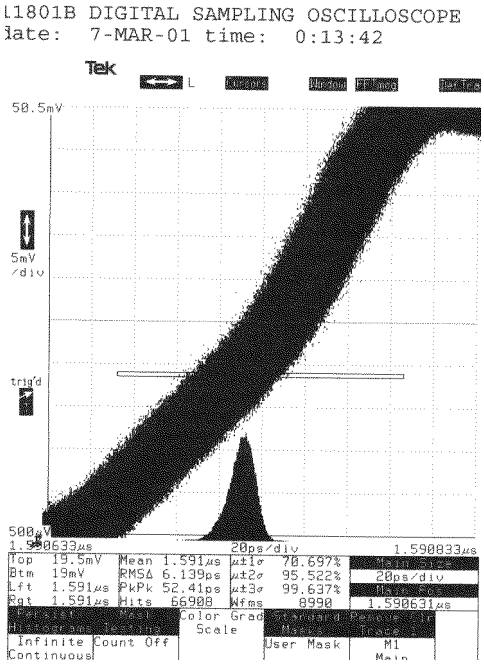


Figure 5.37: Source clock jitter referenced to trigger clock output of BERT. The source clock jitter has an RMS value of 6.2 ps and Pk-Pk value of 52.4 ps at 1 GHz indicating that contribution to LAC clock output is primarily from the source clock to the LAC.

5.8 Summary and future work

In this chapter, a CMOS network interface chip that implements a slotted-ring network was described. The network interface chip achieves measured data rates of over 16 Gb/s in 0.5 µm CMOS technology in a die of size 10.2 mm x 7.3 mm at a power consumption of less than 10.5 W. This is an experimental validation of the broadband capabilities of CMOS-based interfaces to parallel multimode fiber.

The network interface chip integrates the link transceiver components with the digital logic. The transceiver portion of the chip uses low-swing differential voltage format circuitry to implement the serializer and deserializer units. The digital logic controllers are implemented using dynamic true single phase clocked structures and achieve digital clock speeds of over 500 MHz operation. The controller logic is deeply pipelined with single logic gate delay per pipeline stage. Components in the digital logic include FIFOs, RAM and finite state machines. Global clock is distributed using a reverse clocking strategy from the serializer unit down to the elasticity buffer read port. A reverse clocking strategy is an effective method of distributing digital clock which when combined with using retiming latches for receiving data across the initializer, MAC, multiplexer and smoother blocks, is a simple and effective method of achieving 500 MHz digital logic operation. The reverse clock overhead however limits operation to nearly 500 MHz and clock rates of 575 MHz could be achieved by using a zero-clock distribution scheme, as seen from measured results on the P2P chip described in Chapter 3. The measured ring network data rate is the highest obtained to date, to the best of our knowledge enabled by parallel fiber-optic technology.

The performance of the ring network with scaling in CMOS technology is studied in the following chapter.

Chapter 6

Network performance scaling with technology

In this chapter, we study the effect of scaling electrical technology on the performance of ring networks. In Section 6.1, we present some issues that will be encountered in scaling the implemented LAC to finer dimensions. In Section 6.2, we outline a proposed modification to the elasticity buffer that is uniquely suited for ring networks. The motivation here is to maximize available network bandwidth in the presence of small neighboring node clock frequency variations (order of 0.1%). In Section 6.3, a simple analysis of printed circuit board-level interconnects seems to indicate that direct transmission of unequalized unencoded data onto printed circuit traces at data rates of 10 Gb/s on multiprocessor backplane interconnects may not be feasible. Reflection losses and transmission line losses will likely have to be compensated for transmission at over 5 Gb/s data rates.

6.1 Scaling of CMOS technology

In this section, we briefly describe how with scaling in CMOS electronics technology several Gb/s per signal line speeds can be achieved in CMOS technology, potentially resulting in ring networks of over 100 Gb/s network data rates. From constant field

scaling theory, inverter delays scale linearly with process dimensions. From simulations using HSpice at 85^0 C junction temperature and 10% voltage rail-drop though, the fanout-4 delay of an inverter, τ_{FO4}, driving a load equivalent to that of four inverter gates is 220 ps in 0.5 μm CMOS technology and 130 ps in 0.25 μm CMOS technology. This implies a speedup of 70% or potential speeds of 850 MHz in 0.25 μm CMOS technology for a 500 MHz 0.5 μm CMOS implementation. In the LAC, the reverse-clocked interface shown in Figure 6.1 limited achievable speeds to 500 MHz. A straightforward shrink of transistor dimensions to 0.25 μm CMOS however yields a reverse-clocked interface speed closer to 750 MHz. This implies that latches and inverters have to be resized in 0.25 μm CMOS technology to yield a linear scaling of speed with transistor geometry.

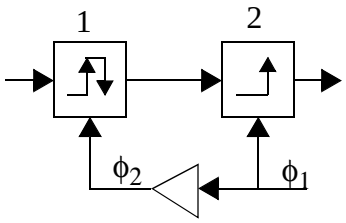


Figure 6.1: Reverse-clocked timing interface across pipestage blocks which limits achievable data rate in the current LAC implementation to 500 MHz digital clock frequency in 0.5 μm CMOS.

It is also becoming apparent that in microprocessors logic and cache speeds will not scale at the same rates with scaling [103] since sense-amplifier sensitivity will degrade with scaling due to an increased impact of variations in threshold voltage. As fine-line dimensions become smaller, the threshold voltage variations due to dopant concentration variations in the channel increase [133],[104],[105],[106]. For processes down to the 0.1

μm CMOS generation, the threshold voltage mismatch between closely spaced transistors in a sense-amplifier scales as $L^{-1/2}$ where L is the channel length, and it deteriorates even further beyond that. This makes the offset voltage in the arms of the cross-coupled inverter pair of the sense-amplifier shown in Figure 3.10 worse with succeeding generations. As outlined in [121], the offset voltage of the cross-coupled pair is related to the threshold voltage mismatches as

$$v_{offset} = (g_{mn}\,\Delta v_{thn} + g_{mp}\,\Delta v_{thp})R_{eff}(t)$$

where v_{offset} is the offset voltage, g_{mn} and g_{mp} are the transconductances, Δv_{thn} and Δv_{thp} are the threshold voltage mismatches of the n and p transistors constituting the cross-coupled pair and $R_{eff}(t)$ is the time-dependent effective resistance seen between the two arms of the inverter pair. Neglecting $g_m R_{eff}$ product scaling it depends primarily on the effective threshold voltage mismatch, which as described earlier deteriorates with decrease in channel dimensions. Hence, the offset voltage is not just non-scalable with geometry scaling, it actually gets worse. In the theoretical analysis on scaling in random access memories [100], threshold voltage mismatch is assumed to be constant with scaling; however since it actually gets worse with smaller geometry the results of memory scaling will further deteriorate than indicated. Below 0.1 μm, the threshold voltage variations become further worse due to short-channel effects such as the threshold voltage, V_{th} itself showing a dependence on channel length L in addition to its usual dependence on channel doping concentration, N_a. Hence achievable speed in memory blocks such as FIFOs and RAMs will not improve at the same rate as that of logic cells. Sense-amplifier circuits will have to be designed to compensate or mitigate the impact of process variation

mismatches [118][119][134][118]. A thorough analysis of FIFOs with scaling for area, speed, power and optimal organization has to be performed to reveal trends in scaling of FIFO speed which then impacts achievable network data rates. Further, subthreshold leakage currents also increase with scaling [132]. It may be necessary to use dual-Vt processes for constructing speed-critical cells using low threshold voltage and other cells such as SRAM cells using higher threshold voltage [107].

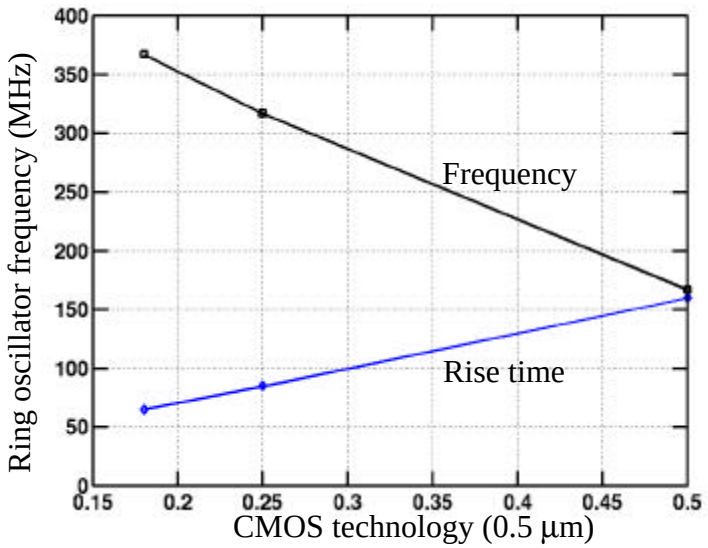


Figure 6.2: Simulated HSpice value for inverter rise times and frequency of oscillation in a 31-stage ring oscillator for 0.5 µm, 0.25 µm and 0.18 µm CMOS technologies shows nearly linear change with process dimensions.

Parameter	0.5 µm	0.25 µm
tsetup	0.5 ns	0.4 ns
tbuffer	0.5 ns	0.35 ns

Table 6.1: Simulations for scaling of LAC

Parameter	0.5 μm	0.25 μm
tlat	0.5 ns	0.3 ns
τ_{FO4}	220 ps	130 ps
Peak reverse-clock timing frequency	500 MHz	740 MHz
31-stage Ring oscillator frequency	167 MHz	317 MHz
Ring oscillator rise time	160 ps	85 ps

Table 6.1: Simulations for scaling of LAC

Historic trends in microprocessor design have shown clock frequency doubling with each generation due to deeper pipelining and advanced circuit techniques [103]. The semiconductor industry association roadmap projects a high-performance on-chip local clock frequency of 3.5 GHz for the 0.1 μm technology and 10 GHz for the 0.05 μm technology. Hence, it may be possible to realize as much as 20 Gb/s per signal line. Newer circuit topologies will be necessary to cope with effects such as subthreshold leakage currents and reduced noise margins.

6.2 Network bandwidth utilization

In this section, we highlight issues arising from the bandwidths enabled by scaling of CMOS technology. Firstly, the potential implication of the multi-Gb/s data rates achieved is that even in low-cost networks of computers spread over a building in a tightly coupled

system of computers (frequency variations within 0.1%), the packet integrity on the ring network and its utilization is impacted by the frequency variations between adjacent nodes.

In the PONI network, slots are separated by idles. The idle symbols between slots in the network represent a loss of available bandwidth in the network. Idles are necessary to synchronize the received network clock with the local host clock in the elasticity buffer due to the plesiochronous nature of operation of the system. A slot size that is large compared to the idle size would maximize the available network bandwidth.

However, the slot size is constrained by the maximum allowable frequency variation between adjacent nodes for correct operation in the smoother. As explained in section 5.2.2 on page 102, if f_w is the frequency of the "write" clock received from the network and f_r is the frequency of the "read" clock of the local host clock and H is the half-size of the estore (or separation between read and write pointers when read operation on the FIFO commences), the maximum slot size, S, allowed to prevent overruns is given by the expression,

$$S = Hf_w/\Delta f$$

where $\Delta f = (f_w - f_r)$ is the difference between write and read clock frequencies. The idle gap has to be at least H bytes long to allow synchronization to take place (so that read pointer can re-synchronize with the write pointer at the end of a slot). Hence, for each slot and idle combination, the network capacity needed is (S+H) or equivalently $H(1 + f_w/\Delta f)$.

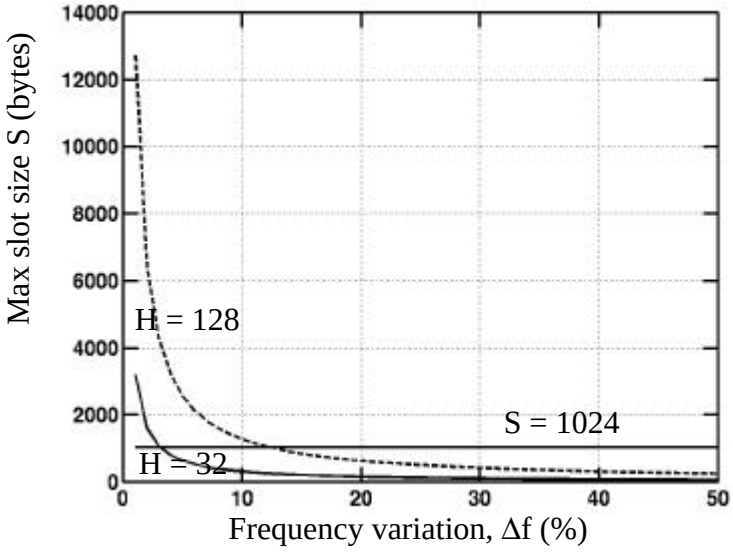


Figure 6.3: Variation of maximum allowed contiguous network slot size, S with frequency variation, Δf between adjacent ring nodes. Maximum slot size is when separation between write and read pointers in estore is set at H bytes where 2H is the depth of the estore. For H = 32 bytes and frequency variation of 0.1%, maximum allowed slot size is 32000. However with dissimilar nodes such as a PC with a 33 MHz host clock and another with a 66 MHz host clock from which the high-speed clocks are realized, for H = 32 bytes, maximum allowable slot size is 32 bytes. Hence, for dissimilar ring nodes, to maximize slot size used, initial write and read pointer separation in estore (and idle size) has to be increased.

The smoother buffer in an LAC node can be used to provide a frequency-independent capacity to hold slot bits in the ring network. The LAC pipeline delay also introduces additional bit capacity of *ndelay* bytes. The present implementation has a node delay of nearly 60 clock cycles exclusive of that added by the smoother.

The fiber interconnect between LAC nodes also represents additional bit capacity. Since speed of light in fiber is approximately 2 x 10^8 m/s, assuming a length L meters of fiber between any two nodes with a data rate of f Gb/s, the bit capacity per node segment is 5fL bytes on an 8-wide fiber-ribbon.

Total byte capacity of a network of *n* nodes is hence

$$T = \sum_{1}^{n} (ndelay + smdelay + 5fL)$$

which is also the total number of slots and idles in the network. Assuming that fiber capacity is not used for slots and all the bit capacity is inserted by the smoother buffer as one slot per node, utilization of network bandwidth is given by the expression,

$$U = \frac{5fL}{5fL + (Slot + Idle)}$$

For the current implementation, slot size is 1 kB and idle gap is 16 bytes. Assuming a node delay of 1040 bytes (accounting for 1024 bytes slot and 16 bytes idle), the network utilization with variation in link data rate is as shown in Figure 6.4. Firstly, larger slots yield higher network utilization. More importantly, unused fiber bit capacity results in considerable wastage of available bandwidth. With this scheme, as much as 65% of

bandwidth could be wasted at 20 GB/s network rates, while the wastage could drop to 50% on doubling slot size. This hence implies that slot size should be increased and fiber bandwidth should be used to optimize network usage.

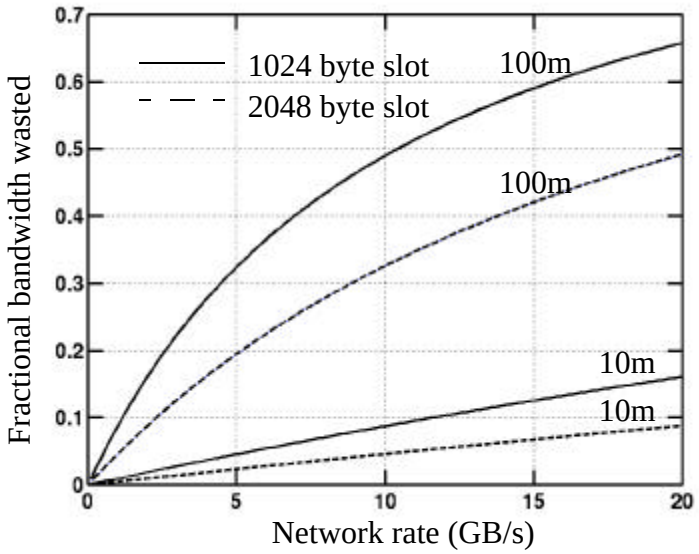


Figure 6.4: Variation of network bandwidth wasted with frequency assuming fiber bandwidth is not used for slots. With this scheme, as much as 65% of bandwidth could be wasted at 20 GB/s network rates, while it could drop to 50% on doubling slot size. This hence implies that slot size should be increased and fiber bandwidth should be used to optimize network usage.

Message size in the network is important for network performance. Network traces [112][73] have shown the existence of a bimodal pattern in network traffic where 90% of the network traffic consists of small packets of less than 200 bytes while only 10% of the traffic consists of large packets such as 8192 byte TCP packets. Overheads dominate for smaller packets while loss of small packets due to receiver bottlenecks may be more easily tolerated or handled. To maximize network usage a better scheme than to separate slots by

idles would be to code the slot train as one contiguous slot where start and end delimiters for each slot are encoded onto the frame control line. The frequency variations between nodes determines the appropriate slot size, as seen from Figure 6.3.

The current prototype implementation of the PONI network was designed for a maximum network size of 32 nodes, expandable up to 64 nodes. The reason for this is that ring node bandwidth scales inversely with the number of nodes connected to the network. From measurements made using raw data transfer [75] bypassing operating system overheads for a 166 MHz Intel Pentium-based Triton motherboard with 82430 VX PCI chipset (33 MHz) motherboard using Windows NT 4.0 operating system, we found that the maximum sustainable bandwidth over a PCI bus in a 1 Gb/s peak data rate interface is 300 Mb/s while throughputs obtained using actual applications may be less than 200 Mb/s. Hence, up to 64 nodes may be connected to a ring network where essentially bandwidth is free despite the rudimentary resource allocation medium access protocol that we have implemented. In a ring network that uses a more sophisticated destination release protocol with spatial reuse, the average bandwidth is doubled and network nodes may even be increased to 128 nodes.

For reliability purposes, ring nodes in a building network may be constructed by connecting to a concentrator that provides the ability to bypass failed nodes or to insert new nodes into a network. Hence, adjacent computers may be separated by as much as 100 m distances, the maximum link distance that the current optoelectronic link components are designed for. For a ring network with nodes separated by up to 100 m distances between adjacent nodes, the internodal bit capacity at 20 Gb/s per signal line

rates is as much as 10 kB which is far more bit capacity than that of the node latency in the LAC. Hence, most of the ring network bit capacity is stored in the fiber and it is critical that fiber bit capacity be used to optimize network utilization. Neglecting node capacity in comparison to the fiber bit capacity, a 128-node network has a bit capacity of 1280 kB. As explained earlier fractional frequency variation allowed by the estore is given by the relation, $S = Hf_w/\Delta f$; for a fixed fractional frequency variation of 0.1% and slot size S of 1280 kB, estore should buffer up to 80 words, or in other words provide up to 2.56 kB of buffer space. Else, the allowable fractional frequency variation between nodes scales down proportionately to 0.005% which is indeed the frequency variation on commercial host crystal clocks. An 80 clock cycle delay per node may be performance-limiting when nodes are spaced closer in the ring network. Commercially available low-cost uncompensated crystal oscillators typically used as low precision digital system clocks have a frequency stability in the range of 10-1000 parts per million while typical PC oscillators which are the more expensive temperature-compensated crystal oscillators typically used on computer systems are stable up to 50 parts per million. Oscillators accurate to 1 part per billion are also available though they are very expensive and better used for high-end telecommunication systems.

A low-cost serial link would be a similar CMOS-based chip interfacing to a single multimode fiber and hence would yield eight times lower data than in a parallel fiber-based link. Connecting the same host computers to such a link would mean that 8 times fewer computers can be connected to the ring to enable the same bandwidth per ring node. This problem may hence not be seen in low-cost CMOS-based serial links.

The solution for slotted-ring networks may be to ensure that the contiguous slot size is restricted to that allowed by the frequency variations between node clocks. In a system with nodes with large variations in clock frequencies (>> 0.005%), this will result in under-utilization of network bandwidth since buffering requirements in the elasticity buffer are higher resulting in larger idle gaps. A better solution may hence be to re-design the elasticity buffer to fragment an originally contiguous slot to maximize network utilization. In slotted-ring networks, packet discarding is not desirable since the slot integrity must be preserved; else, the garbled slots must be removed and the network re-initialized.

The elasticity buffer in the DEC FDDI [22] design for a 125 Mb/s token ring network which is a typical design prevents write and read pointers from pointing to the same location and to drop packets when there is an overrun or underrun. Such a scheme will result in slot corruption in the slotted-ring network which then needs the ring to be re-initialized with slots. Unlike conventional networks such as crossbar switches where the incoming stream is indefinite in length and packet losses due to buffer overruns are unavoidable, the length of slots in the slotted-ring network is a known, bounded value. This feature can be exploited to mitigate packet losses in the slotted-ring network.

The first step is to initialize the ring with slots in a contiguous manner to maximize bandwidth utilization. A possible method of maximizing network usage in a scalable manner is to recursively initialize the network with slots. Hence, the master loads one slot

onto the network and then counts succeeding bits to ascertain whether there is sufficient space for another slot and idle. The next slot is then added in the next roundtrip and the procedure is repeated until the network is filled with slots.

Next, the elasticity buffer should be designed to dynamically fragment a contiguous slot at the nearest slot boundary to prevent impending overflow or underflows. Thus, whole slots will be preserved and packets will not be lost and there is a dynamic arrangement of slots and idles in the network depending on the frequency variations among the various ring nodes. Due to the repetitive nature of a circulating ring, the probability of occurrence of slot corruption is diminished with time for random frequency variations between nodes.

A complete analysis of ring network performance requires a detailed analysis of performance of applications at individual nodes, scaling of clock jitter and hence frequency stability at lower CMOS dimensions, skew across parallel lines with scaling in technology and line encoding overheads. Statistical traffic models provide the flexibility of analyzing for traffic due to various kinds of current and anticipated multimedia applications.

Until recently, arrival processes have been mainly modeled using Poisson processes. Previously, it had been shown [110] from packet data collected on a token ring network that a “packet train” (or ON/OFF) source model was more appropriate in order to capture the observed burstiness in network traffic. The model did not gain acceptance because the authors did not present a formal basis for the modeling approach. Another study of

importance was measurement conducted on a 10 Mb/s Ethernet LAN [111][112][113] which showed that data is bursty across multiple time scales. However, this modeling approach accounted for burstiness over only a limited range of time scales.

A landmark study [108] however showed that link traffic in local area networks exhibits long-range dependence and is parsimoniously modeled using self-similar processes. The same was also shown to be true in wide-area networks [109]. The network traffic is better represented by self-similar processes when several estimated statistics exhibit power-law behavior over a wide range of time (and frequency) scales and correlation properties exhibit long-range dependence. Conventional traffic models such as Poisson processes, Markov-modulated Poisson processes, do not exhibit the power-law behavior. This then led to a revision of the models used for source arrival processes.

At higher speeds, propagation delay in the interconnect dominates total delay in comparison to switching node delay. As a result of multiple packets being in flight between nodes, congestion and flow control protocols will become necessary [88] even in the case of these small networks of computers spread over a carpet floor.

6.3 Limitations of electrical links

In this section, we discuss issues in board-level electrical interconnects with scaling in data rates. We also perform a rudimentary analysis of the length scales over which printed circuit traces enable adequate signal integrity.

6.3.1 Transmission line model

In this section, we review basic transmission line theory. A lumped equivalent circuit model of a transmission line is shown in Figure 6.5. The transmission line parameters are the resistance R, inductance L, shunt conductance G and capacitance C per unit length.

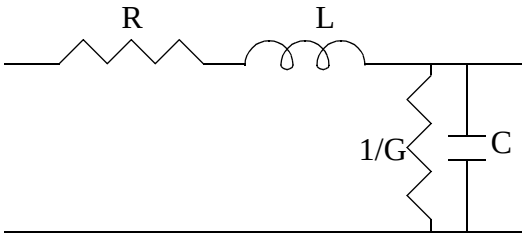


Figure 6.5: shows the lumped equivalent circuit model for a section of a transmission line. The parameter R is the conductor resistance per unit length, L is the conductor inductance per unit length, G is the dielectric conductance per unit length, C is the capacitance per unit length.

The solution for the electric potential in a transmission line [79] at a distance y along the transmission line is

$$V(y) = A\, e^{-\gamma y} + B\, e^{\gamma y}$$

where, the parameter γ is the propagation constant of the line, A and B constants evaluated using the boundary conditions. Assuming matched impedance termination and neglecting any reflections due to impedance mismatching, the potential due to the forward traveling wave is given by

$$v(y) = v_0\, e^{-\gamma y}$$

where, v_0 is the initial potential on the transmission line (y=0). Hence, the transfer function of the transmission line of a length L is given by

$$H(\omega) = e^{-\gamma L}$$
$$P(dB) = 20\log|H(\omega)| = -(20|\gamma|L)/(\ln 10)$$

At an angular frequency ω (=2πf rad/s), the propagation constant γ is given by the expression

$$\gamma = \sqrt{(R + j\omega L) \times (G + j\omega C)}$$

The propagation constant γ can be represented as

$\gamma = \alpha + j\beta$

$\alpha = \alpha_c + \alpha_d$

where, α_c contributes to conductor losses, α_d contributes to dielectric losses and β contributes to the phase of the traveling wave. Under low-loss approximation ($R << \omega L$, $G << \omega C$), the above terms can be expressed as

Conductor losses, $\alpha_c = R/2Z_0$

Dielectric losses, $\alpha_d = GZ_0/2$

Phase, $\beta = \omega/v_p$

where, v_p is the phase velocity of light and Z_0 is the characteristic impedance of the line given by the expression

$$Z_0 = \sqrt{\frac{(R + j\omega L)}{(G + j\omega C)}}$$

and G is the shunt conductance given by the expression

$G = \omega C \tan \delta$

where $\tan \delta$ is the loss tangent of the dielectric material. Under low losses

,

$$Z_0 = \sqrt{\frac{L}{C}}$$

which can be rewritten as

$$Z_0 = \frac{v_p}{C}$$

where the phase velocity v_p is given by the expression,

$$v_p = \sqrt{\frac{1}{\mu_0 \varepsilon_0 \varepsilon_{eff}}}$$

where μ_0 is the permeability, ε_0 is the electrical permittivity and ε_{eff} is the effective dielectric constant for the conductor-dielectric transmission line structure.

The resistance of the conductor, R varies with frequency due to skin effects. As frequency increases, the electrical field strength falls off rapidly inside the conductor and the current flows almost entirely in a conductor of cross-sectional depth δ from the surface of the conductor given by the expression,

$$\delta = \sqrt{\frac{1}{\pi\mu\sigma f}}$$

where μ is the permeability, σ is the conductivity of the conductor and f is the frequency of the signal. As depth of penetration of the fields within the conductor drops with frequency, its internal inductance also changes correspondingly. Based on the incremental inductance rule [82] the conductor losses in a microstrip can be evaluated.

6.3.2 Characterizing loss in microstrip transmission lines

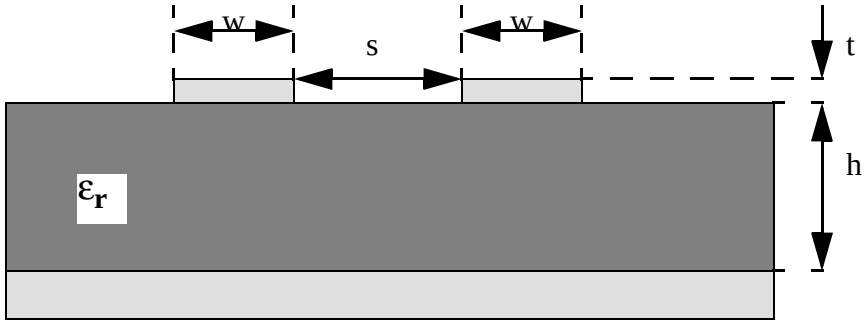


Figure 6.6: shows the geometric structure of a microstrip line. Here, the width of the conductor is w, its thickness is t, height from reference plane is h, spacing between adjacent conductors is s, and ε_r is the dielectric constant of the dielectric material.

An analytical expression for conductor losses of a microstrip line as shown in Figure 6.6 calculated based on the incremental inductance rule [82] in [80][81] is given by

$\alpha_c = (R_s A_\rho)/(Z_0 h)$ in dB/m

Here, Z_0 is the characteristic impedance of the transmission line, R_s is the specific resistance ($1/\sigma\delta$), h is the height of the conductor from the reference plane, A_ρ is a geometry factor given by the expressions (in dB)

For 0 ≤ w/h ≤ 1/2π

$$A_\rho = \frac{8.68}{2\pi}\left[\left\{1-\left(\frac{w_{eff}}{4h}\right)^2\right\}\left\{1+\frac{h}{w_{eff}}+\frac{h}{\pi w_{eff}}\left(\frac{t}{w}+\ln\frac{4\pi w}{t}\right)\right\}\right]$$

For 1/2π < w/h ≤ 2

$$A_\rho = \frac{8.68}{2\pi}\left[\left\{1-\left(\frac{w_{eff}}{4h}\right)^2\right\}\left\{1+\frac{h}{w_{eff}}+\frac{h}{\pi w_{eff}}\ln\left(\left(\frac{2h}{t}\right)-\frac{t}{h}\right)\right\}\right]$$

For 2 ≤ w/h

$$A_\rho = \frac{8.68}{2\pi}\frac{\left\{\frac{w_{eff}}{h}+\frac{w_{eff}/(\pi h)}{w_{eff}/(2h)+0.94}\right\}}{\left[\frac{w_{eff}}{h}+\frac{2}{\pi}\ln\left\{5.44\pi\left(\frac{w_{eff}}{2h}+0.94\right)\right\}\right]^2}\left[1+\frac{h}{w_{eff}}+\frac{h}{\pi w_{eff}}\left\{\ln\left(\frac{2h}{t}\right)-\frac{t}{h}\right\}\right]$$

where w is the width of the microstrip trace, h is the height from the ground plane, w_{eff} is an effective width of the trace with a correction term to the width to account for fringing fields arising from the finite width of the conductor and t is the thickness of the trace. Further, to account for conductor losses due to surface roughness,

$$\alpha_{c,\Delta} = \alpha_{c,0} \cdot \left\{ 1 + \frac{2}{\pi} \cdot \operatorname{atan}\left[1.4 \left(\frac{\Delta}{\delta} \right)^2 \right] \right\}$$

where δ is the skin depth of the conductor and Δ is the rms surface roughness of the conductor, $\alpha_{c,0}$ is the conductor loss for a perfectly smooth conductor.

At high frequencies, the microstrip radiates at all discontinuities. The radiation is divided into a free-space component and a surface wave component. The surface-wave radiation which is produced at all discontinuities in the microstrip line produces a surface wave in the ground plane which propagates in the transverse direction over the plane. The intensity of this wave increases with frequency. The surface-wave modes experience multiple reflections at the substrate boundaries and at strip transmission-line structures. This produces a parasitic coupling to all parts of the circuit thus affecting the response of the microstrip line. The measurement of the frequency response of a microstrip line is further complicated by the fact that microstrip lines do not propagate pure TEM waves, but instead propagate quasi-TEM waves. This leads to a discontinuity in launching from an SMA connector onto the microstrip line since SMA and coaxial cable lines propagate TEM waves.

The expression for Z_0 is as given in [77][78]. The expressions for L, C and G are calculated as explained in Section 6.3.1. The permittivity ε is calculated as $\varepsilon_{eff}\varepsilon_0$ where ε_{eff} is the effective dielectric constant for the conductor-dielectric structure due to the non-homogeneous dielectric (air above the conductor and PCB insulator below the conductor).

6.3.3 Loss measurements

The measured losses for microstrip traces on an FR-4 printed circuit board (ε_r = 4.5 and loss tangent tan δ = 0.02) are shown in Figure 6.8, Figure 6.12, Figure 6.13 and Figure 6.14. All traces have the following parameters: width, w of 8 mils, thickness t of 1.3 mils, height from ground plane h of 5 mils. Loss was measured using a 50 GHz HP 8510C network analyzer. The total loss for the microstrip trace due to conductor and dielectric losses calculated from the expressions listed in this section is compared against the measured loss given by the scattering matrix parameter S_{21} for a microstrip trace with a length of 12 inches as shown in Figure 6.8.

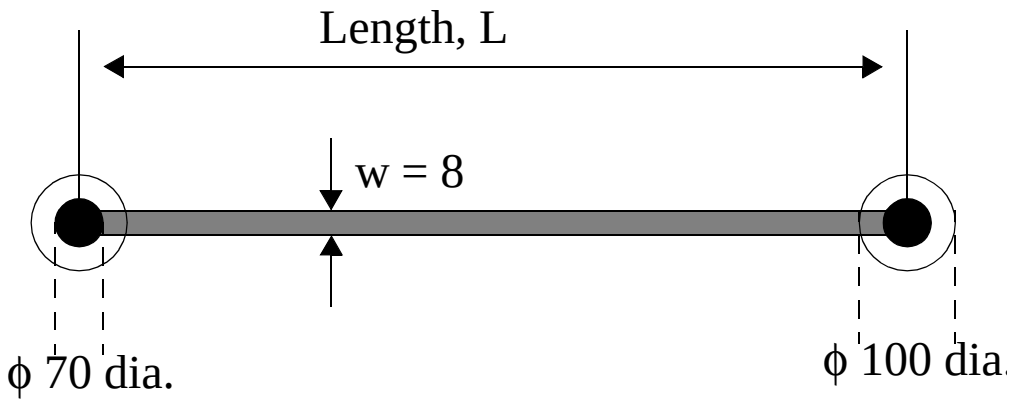


Figure 6.7: Top view of test microstrip trace with SMA connector launched signals. The SMA signal conductor is inserted into a via with drilled plated hole size of 70 mils and an outer pad size of 100 mils on top as well as internal layers. Clearance around the pad on internal ground and power layers is 10 mils. Width of the trace is 8 mils.

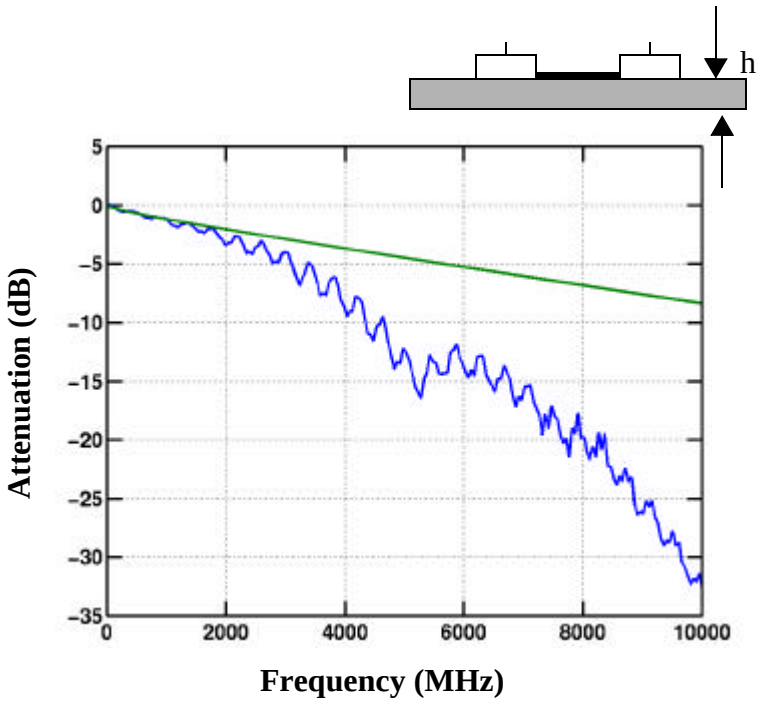


Figure 6.8: Plot of S_{21} measurements for a microstrip trace on FR-4 with SMA connectors on either end and line parameters width w = 8 mils, height h = 5 mils, length = 8.0 inches (20 cm) and simulated loss due to conductor and dielectric losses for FR-4 loss tangent of 0.02. The figure shows that actual loss exceeds simulated loss which is due to additional losses arising from reflections and radiations at the SMA-microstrip discontinuity. This additional loss is nearly 9 dB at 5 GHz.

As seen from Figure 6.8, the measured loss exceeds the theoretically computed loss. On measuring the reflected power given by the scattering matrix parameter S_{11} using an HP network analyzer, it is seen from Figure 6.9 that reflected power increases substantially with frequency with nearly total reflection beyond 8 GHz. The reflections

are due to the discontinuity at the SMA - microstrip interface where TEM waves launched from a coaxial cable now transition to quasi-TEM waves in the microstrip printed circuit board trace.

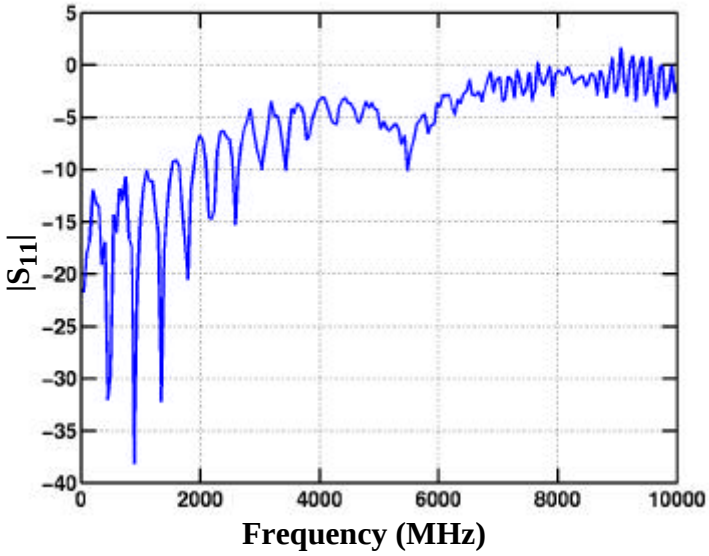


Figure 6.9: Plot of S_{11} measurements for a microstrip trace on FR-4 with SMA connectors on either end and line parameters width w = 8 mils, height h = 5 mils, length = 8.0 inches. Right-angle SMA connectors on either end of the traces are used to launch and collect a signal. The measurements indicate that reflections at the SMA-microstrip discontinuity result are nearly 100% at beyond 8 GHz indicating that a good launch onto PCB traces is necessary for achieving high-speed printed circuit board performance.

By comparing the difference in loss measured for two traces of different lengths, the effect of reflection loss can be subtracted from loss measurement as a first-order correction. In the S_{21} measurements shown in Figure 6.8, the spectrum shows ripples whose peaks are separated by approximately 400 MHz, which roughly corresponds to the round-trip time in a trace of length 8 inches pointing towards a reflection introduced

spectral dependence. The loss calculated for difference in losses between traces compared against the analytical computation is shown in Figure 6.12, Figure 6.13 and Figure 6.14. The analytical computation shows reasonable agreement with the measured loss values.

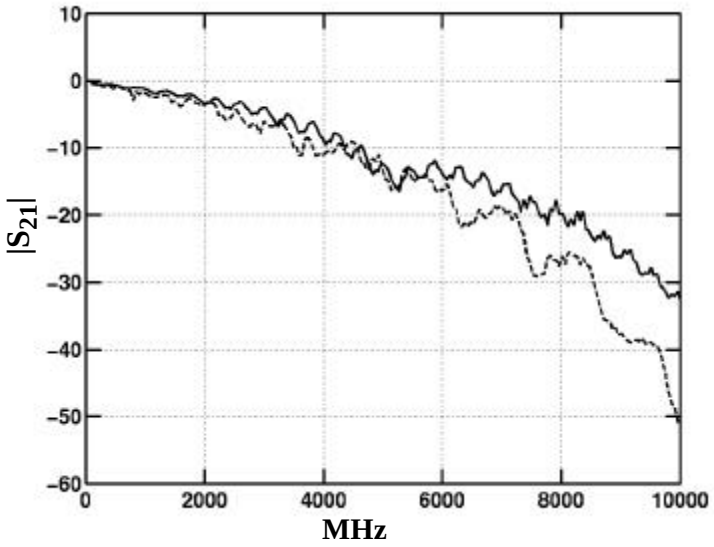


Figure 6.10: Plot of S_{21} measurement for trace of length 8 inches, w = 8 mils comparing trace with no vias (solid) to a trace which has two vias at 2.5 inches from either end (dashed line). Trace with no vias shows lower loss than trace with two vias.

Discontinuities such as vias also affect the transmission loss as seen from Figure 6.10. At 10 GHz lines with vias show as much as 20 dB higher loss. A trace with the SMA signal conductor extension is filed off where it extends below the board shows lower loss than one where it is intact as seen from Figure 6.11 indicating that radiation losses are also one of the loss components in transmission in addition to the reflection losses at the coaxial-microstrip transition.

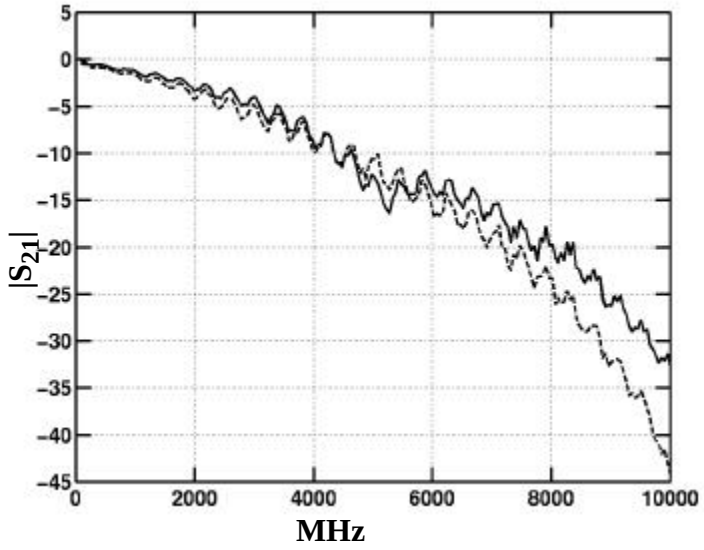


Figure 6.11: Plot of S_{21} measurement for trace of length 8 inches, w = 8 mils comparing trace with SMA signal conductor for vertical launch filed off (solid) to one where the signal conductor is intact (dashed). When signal conductor is filed off at the bottom of the board, the trace shows lower loss than otherwise indicating that the SMA radiates some energy otherwise.

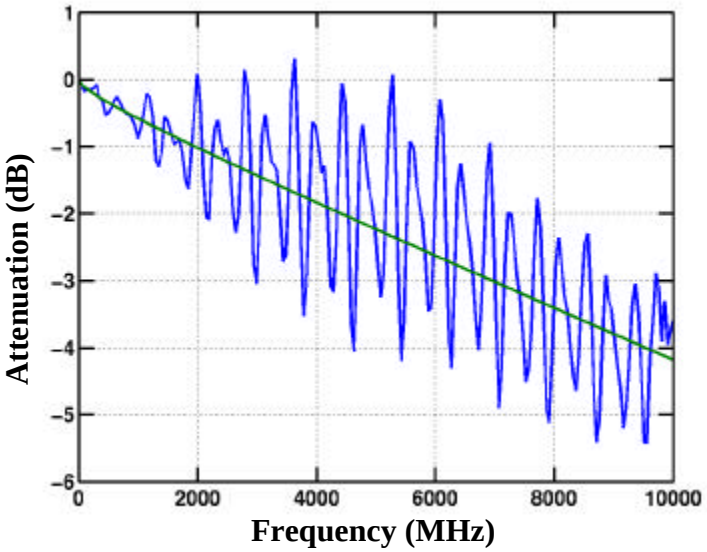


Figure 6.12: Plot of difference in measured S_{21} parameter of two microstrip traces on FR-4 with parameters w = 8 mils, h = 5 mils, length of first trace = 12000 mils and length of second trace = 8000 mils and simulated loss for trace of length 4000 mils and FR-4 loss tangent of 0.02. The variations in measured loss are due to reflections at microstrip-SMA launch discontinuity.

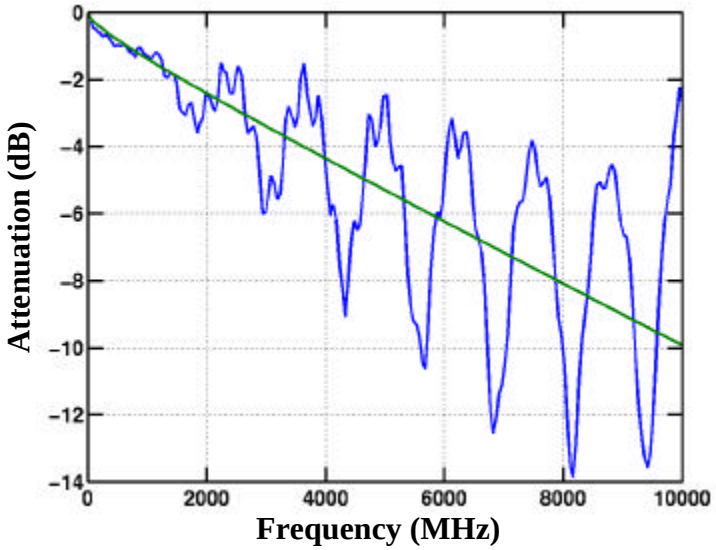


Figure 6.13: Plot of difference in measured S_{21} parameter of two microstrip traces on FR-4 with parameters w = 8 mils, h = 5 mils, length of first trace = 12000 mils and length of second trace = 2500 mils and simulated loss for trace of length 9500 mils and FR-4 loss tangent of 0.02. The variations in measured loss are due to reflections at microstrip-SMA launch discontinuity.

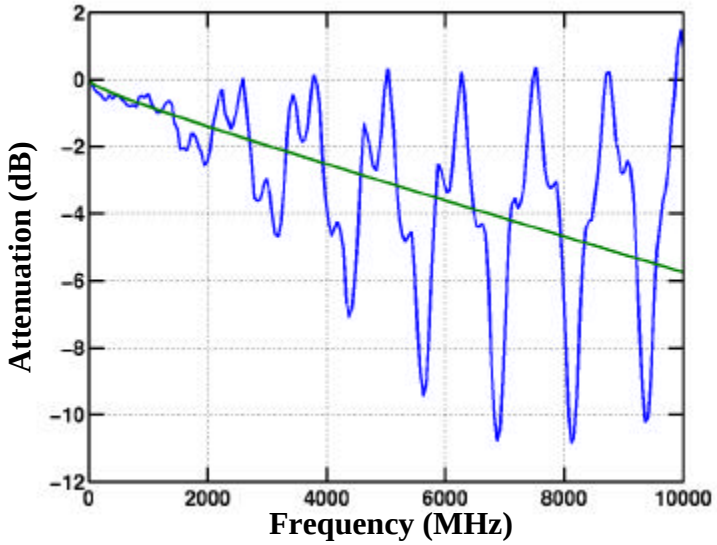


Figure 6.14: Plot of difference in measured S_{21} parameter of two microstrip traces on FR-4 with parameters w = 8 mils, h = 5 mils, length of first trace = 8000 mils and length of second trace = 2500 mils and simulated loss for trace of length 5500 mils and FR-4 loss tangent of 0.02. The variations in measured loss are due to reflections at microstrip-SMA launch discontinuity.

6.3.4 Form-factor in microstrip conductors

The achievable form-factor for an interconnect composed of microstrip conductors depends on the line-width and pitch of the traces. While narrower trace widths and spacing result in better form-factor, correspondingly conductor losses and crosstalk also increase.

6.3.4.1 Crosstalk

Crosstalk evaluation is performed by calculating the coupling coefficient. The coupling coefficient indicates the amount of signal coupled from a transmission line to another lying adjacent to it. Any signal traveling down a coupled transmission line pair is composed of two modes of propagation, which are even and odd in the case of symmetric

lines. The characteristic impedances in the two modes are designated as $Z_{0,e}$ and $Z_{0,o}$, for the even and odd modes respectively. Consider two transmission lines, A and B, lying adjacent to each other. A perturbation V traveling in A produces a voltage V_B in B. The coupling coefficient, which is the maximum coupling achieved (when the electrical length of the line is equal to $\lambda/4$), is given by

$$k = V_B / V$$

In terms of the odd and even mode impedances, the coupling coefficient can be written as

$$k = \frac{Z_{0,e} - Z_{0,o}}{Z_{0,e} + Z_{0,o}}$$

The amount of power coupled is given by

$$P = 20 \log k$$

If the maximum allowable coupling between pairs of lines is 20 dB, the maximum coupling coefficient is calculated to be k = 0.1. The expressions used in the calculation of $Z_{0,e}$ and $Z_{0,o}$ were obtained from [77] and [78]. For a 50-ohm line of ε_r = 4.5, t = 1.3 mils, spacing s = width w, the coupling coefficient k was found to be under 0.09 from w = 8 mils down to w = 3 mils. Thus, in the case of a microstrip line, if adjacent lines are spaced by more than a line width the coupling can be kept sufficiently low.

w (mils)	h (mils)	k
8	5	0.078
7	4.5	0.08
6	3.9	0.08
5	3.3	0.082
4	2.7	0.083
3	2.1	0.087

Table 6.2: Coupling coefficient of 50-ohm characteristic impedance microstrip trace of isolated microstrip pair on FR-4 for trace thickness t = 1.3 mils, width w, height h and spacing between coupled line pairs = width w of conductor. Coupling calculated for isolated transmission line pairs is less than 20 dB if lines are spaced at least one line width apart.

6.3.4.2 Loss in scaled PCB traces

Assuming FR-4 as the dielectric material, a loss calculation is performed for scaled microstrip lines. Neglecting the reflection losses when launching onto the microstrip trace, a loss calculation is performed for conductor and dielectric losses as shown in for 50-ohm characteristic impedance lines of widths 8 mils, 5 mils and 3 mils. The figure shows that loss per meter at 10 Gb/s signalling rate is between 40 dB and 55 dB per meter for these traces. The tolerable loss in a link depends on the receiver sensitivity. For a 20 dB allowable loss, direct transmission allows less than half a meter of transmission. Using techniques such as equalization [122] and data encoding can alleviate some of these concerns.

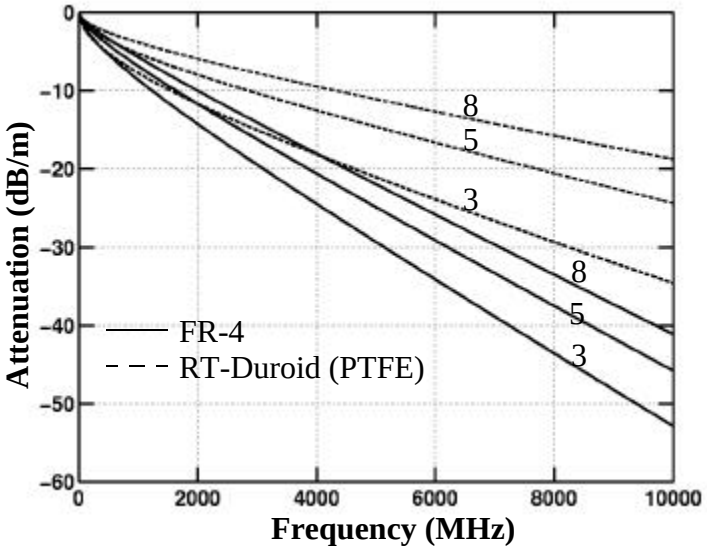


Figure 6.15: The figure shows loss per meter distance for 50-ohm microstrip traces on FR-4 (solid lines) with a loss tangent of 0.02) and RT-Duroid (dashed lines) with a loss tangent of 0.005 for trace widths 8 mils, 5 mils and 3 mils and trace thickness of 1.3 mils. For a 3-mil line at 10 Gb/s signalling speed in FR-4, loss is between 52 dB for a meter, indicating that direct transmission onto FR-4 microstrip trace with a 17 dB link budget has an upper bound of 33 cm. In RT-Duroid, for the same line, the upper bound is 50 cm. Reflection losses at launch discontinuity have been neglected. The corresponding numbers for 8 mil lines are 40 cm for FR-4 and 90 cm for RT-Duroid, again excluding reflection losses which could be significant.

Future microprocessors will dominantly handle 64-bit data words and maximum component density in chips is expected to double over the next five years [7]. To route signals to interconnect microprocessors on a backplane in a cabinet, transmitting 64 differential signals on a PC trace board may not be a practical solution. Hence, it may be necessary to serialize the internal traces to eight pairs of differential signal lines. Due to tighter spacing requirements on area arrays such as on BGA packages, pin-to-pin spacing may be as low as 3 mils. As seen from Table 6.3, at rates higher than 5 GHz a direct

transmission onto electrical signal lines is increasingly difficult and may need techniques like equalization [122] and data encoding. For a given BER, equalization could extend the achievable link length by 50% [137]. Hence, from an electrical loss perspective PCB traces are competitive with optics for distances less than a meter. However, a 1% variation in trace lengths (1 cm variation) can result in trace skews as high as 70 ps. Hence, in addition to loss compensation circuitry, skew compensation circuitry may also be necessary.

	5 GHz	10 GHz
FR-4 with measured vertical SMA launch onto 8 mil lines	< 50 cm (9 dB in reflections and radiations)	Not possible
FR-4, 8 mil lines, simulated microstrip loss only	< 77 cm	< 41 cm
FR-4, 3 mil wide lines, simulated microstrip loss only	< 57 cm (excludes reflection loss)	< 33 cm (excludes reflection loss)
PTFE Dielectric (RT-Duroid), 8 mil simulated microstrip loss only	< 140 cm	< 90 cm
PTFE Dielectric (RT-Duroid), 3 mil line, simulated microstrip loss only	< 80 cm	< 50 cm

Table 6.3: Microstrip electrical link performance for 17 dB link budget

6.4 Summary

In this chapter, performance of the network interface chip is studied for scaling in three dimensions. Firstly, a scaling analysis of CMOS feature sizes and the resultant increase in network data rates is performed. Secondly, the resulting impact of clock jitter on network utilization and suggested remedies is presented. Lastly, an analysis of scaling of printed circuit board dimensions and the effective distances for applicability of copper interconnects is discussed.

With CMOS scaling, network data rates of over 100 Gb/s can potentially be realized. As a result, local area network issues such as the effects of a long packet train causing synchronization problems at the elasticity buffer, flow and congestion control problems will be seen in building networks. Finally, relative to optical links, copper PCB interconnects do not have the form-factor advantage afforded by optics and may be competitive with optics only for distances less than a meter. In order to achieve 10 GHz clock speeds on computer system backplanes, it may be necessary to introduce optical interconnects more intimately integrated into the system.

Chapter 7

Conclusions

In this chapter, we summarize findings described in the dissertation. We also outline some suggestions as possible directions for future work.

7.1 Summary of dissertation chapters

This dissertation primarily addresses the applicability of slotted-ring networks for multimedia applications in clusters of computers spread over a “carpet floor”. Slotted-ring networks are particularly attractive for their topological simplicity and distributed switching in comparison with mesh-based networks such as those constructed using centralized crossbar switches. Congestion control, flow control, buffering requirements, broadcast and multicast are more easily designed in shared-medium ring networks. With emerging technology of parallel fiber-optic links and advanced CMOS processes, high data rates in slotted-ring networks may be obtained.

Parallel low-skew fiber enables direct transmission of clock in parallel with data which mitigates the need for the complexity and expense of clock and data recovery circuitry and active deskewing circuitry. The simplicity of a slotted-ring network interface chip which avoids the quadratic increase in switching element complexity with number of

ports in an N-port crossbar switch enables optimization for high digital logic data rates. Further, by eliminating noise-sensitive such as phase-locked loops (PLL) performance-reducing effects of digital switching noise on sensitive analog circuitry is reduced along with a reduction in total chip area. Coupled with broadcast and multicast advantages of shared-medium ring networks, this enables low-cost high-performance ring networks for carpet cluster applications.

We experimentally demonstrated this using a network interface chip which achieves measured data rates of over 16 Gb/s in 0.5 μm CMOS technology at a total power consumption of less than 10.5 W. Details of the implemented network protocol are found in Chapter 4 while full details of the implementation of the network interface chip may be obtained from Chapter 5.

Improving the clock distribution scheme will further raise this achievable rate to 18 Gb/s as demonstrated experimentally using a second chip that implements a point-to-point link as described in Chapter 3. This chip achieves over 18 Gb/s link rates in 0.5 μm CMOS technology at a power consumption of less than 5.5 W.

A simple analysis of ring network performance for scaling of CMOS dimensions, clock jitter and PCB microstrip interconnect dimensions is performed in Chapter 6. From scaling considerations an attempt is made to predict future achievable rates in ring networks which could potentially be in excess of 10 GB/s. Ring networks could hence continue to be a cost-effective solution for small sized networks and particularly attractive for professional workgroup multimedia applications.

Issues relating to threshold voltage mismatch effects on memory speed in CMOS chips, clock jitter at the network system level and transmission line losses and reflections in printed circuit board interconnect performance will however have to be addressed to enable these higher rates. At these higher data rates, clock jitter could however increasingly play a role in achievable network utilization. Allowable packet train length is restricted by the elasticity buffer operation to a value specified by the allowable frequency variations between adjacent ring nodes.

7.2 Suggestions for future work

Further work remains to be done in performance analysis of the physical layer with scaling, performance analysis of applications at individual nodes and in the implementation of feature enhancements to the LAC. Some suggestions are outlined in this section.

7.2.1 Scaling analysis of physical layer performance

As outlined in Chapter 6, with scaling of CMOS feature sizes, logic may be expected to scale linearly. However, memory may not quite scale in performance at the same rates. The reason for this is that threshold voltage variations become worse with smaller feature sizes, the deviation being inversely proportional to the square root of the channel length of closely-spaced transistors such as those used in a sense-amplifier. Hence, threshold voltage mismatch compensation circuitry will need to be implemented for sense-amplifiers.

An analysis of dynamic TSPC logic performance with scaling also needs to be performed. As threshold voltages decrease, the noise margins of dynamic logic further decrease. Hence, it remains to be seen whether significant gains are realized from dynamic logic as compared to static logic or whether a weak feedback from dynamic logic gate outputs will ameliorate the noise vulnerability.

At lower threshold voltages, the subthreshold leakage currents also increase. This can impact bitline design in the memory where the maximum bitline height must be restricted so as to minimize the subthreshold leakage current through access transistors lying on inactivated wordlines as compared to the saturation drain current through an access transistor lying on a wordline that is currently activated. Hence, for the above two reasons, memory is expected to scale slower in performance as compared to logic unless threshold voltages are adequately compensated for and if dual-V_t processes are adopted with higher V_t for memory transistors.

With the resolution of the above issues in scaled CMOS devices, network data rates of over 100 Gb/s can potentially be realized. As explained in Chapter 6, clock jitter can hence increasingly affect network utilization. At higher on-chip clock frequencies, substrate noise can increase potentially increasing jitter seen on the output clock. At higher data rates, as more bits are packed into the same physical length of fiber, frequency variations can build up as the bit train becomes longer. In conventional networks, to prevent overruns or underruns at the elasticity buffer, further writes are inhibited and hence packets may be dropped. A more intelligent approach may be used in a ring network where packet train length is upper-bounded by the maximum diameter of the ring

and a predetermined number of slots circulate as in a slotted-ring network. A better solution may be to fragment the packet train at a slot boundary thereby preventing any loss of packets. Hence, since the tendency of the packet train is to reduce with time, errors due to frequency variations in the ring network will reduce with ring rotation time assuming frequency variation patterns do not increase with time.

As was seen from measurements obtained from the point-to-point chip described in Chapter 3, speed is restricted first by inadequate swings on the bitlines in memory. Hence, buffer sizes are critical to determine the maximum achievable data rate in a ring network. Larger buffers may be constructed by arraying blocks of smaller buffer units. For example, the currently implemented FIFO memory unit with a height of 64 wordlines could be used as the unit sub-array size to construct larger buffers. An analysis of buffers for power, speed and area trade-offs needs to be performed to determine dependency of network data rate on buffer size. Larger buffers will be necessary to handle multiple packets headed for the same destination.

The above problem is exacerbated by a wide gap in data rates of the datapath and the external host interface (currently TTL) ports. The TxFIFO and RxFIFO interface to the host uses TTL signal levels. The reverse-clocked external interface placed a limitation of 50 MHz clock speed on this interface resulting in a data bus throughput of 200 MB/s which is a factor of 10 lower than the peak network data rate. If the receive node only receives from an identical transmit node, there will be no buffer overflow. However, if more than one node transmit to the same node at peak data rates for a sustained period of time, the receive buffer will not be able to flush out the received packets on time. Hence,

the receive buffer may be incapable of handling network received traffic. To mitigate the problem, the speed of the host interface has to be increased. Using full rail-swing TTL signal levels is not conducive towards achieving high-speeds. Hence, one will need to migrate to a reduced swing differential format such as the LVDS format for this interface to dedicated large external buffers that can store packets before they are serviced by the host.

7.2.2 Application performance analysis

A complete analysis of ring network performance requires a detailed analysis of performance of applications at individual nodes, scaling of clock jitter and hence frequency stability at lower CMOS dimensions, skew across parallel lines with scaling in technology and line encoding overheads. Statistical traffic models provide the flexibility of analyzing for traffic due to various kinds of current and anticipated multimedia applications.

Until recently, arrival processes have been mainly modeled using Poisson processes. A landmark study [108] however showed that link traffic in local area networks exhibits long-range dependence and is parsimoniously modeled using self-similar processes. The same was also shown to be true in wide-area networks [109]. The network traffic is better represented by self-similar processes when several estimated statistics exhibit power-law behavior over a wide range of time (and frequency) scales and correlation properties exhibit long-range dependence. Conventional traffic models such as Poisson processes, Markov-modulated Poisson processes, do not exhibit the power-law behavior. This then

led to a revision of the models used for source arrival processes. Slotted-ring network performance for self-similar traffic sources has not been studied previously and is a line of future work. A significant issue will relate to the necessary buffer sizes due to self-similar traffic sources.

7.2.3 Network system

The PONI network implements two link layer protocols for sharing bandwidth resources - the lower priority source release protocol and the higher priority guaranteed bandwidth protocol. Packets are released at the source which cannot reuse the slot it just released. A future implementation would include protocols that maximize usage of available bandwidth such as those based on a destination release scheme with spatial reuse. For implementing a destination removal scheme with destination reuse, the receive buffers should precede the transmit buffers in the datapath. The smoother buffer unit could also be placed earlier in the datapath where the transmit buffer contents are loaded into the datapath to utilize the smoother as an insertion buffer needed to implement buffer-insertion rings. Buffer-insertion rings could offer some advantages relative to slotted-ring networks as outlined briefly in Chapter 4 such as the ability to transmit variable size packets.

Ring reliability features have not been implemented in our prototype implementation and constitute a necessary feature of any operational network. In the case of a unidirectional ring network, optical bypass switches could be used to bypass a faulty node. Or a faulty node could be electrically bypassed within a concentrator. On dual rings,

wrapping around a faulty node keeps the ring alive. Optical bypassing is a well-established technique in serial optical links such as FDDI. Similar technologies have to be established for parallel fiber-optic rings.

Further feature enhancements are desirable on the network interface link adapter chip. Noise may in the future may limit achievable chip speed. A backside substrate contact will be useful for noise-isolation purposes. Better packaging (such as flip chip) and improved power regulation will reduce noise generated in power and ground by the digital circuitry. Advanced packaging techniques can be used to efficiently distribute and regulate power distribution. Improving substrate connection to ground will provide further noise isolation for analog circuitry. Power supply to the chip will also need to be efficiently regulated to guarantee integrity of a good, clean power supply.

A reverse clocking strategy is an effective method of distributing digital clock which when combined with using retiming latches for receiving data across the initializer, MAC, multiplexer and smoother blocks, is a simple and effective method of achieving 500 MHz digital logic operation. The frequency limitation arises from this retiming operation across datapath blocks. Further exploration is necessary to increase maximum achievable data rate by possibly introducing an additional retiming latch level within the block or by globally distributing a zero-skew clock.

Further work is also required to reduce static power consumption within the memory blocks due to paths to ground through a memory cell activated by a wordline and in the cross-coupled inverter pair in the sense-amplifier arms during the equalization operation

prior to read bitline evaluation. Besides, since writes and reads to the same memory location are not permitted, a six-transistor memory cell with common access transistors for write and read operations may suffice and should be investigated further.

References

[1] Homepage of Intel, Inc., http://www.intel.com

[2] Homepage of AMD, Inc., http://www.amd.com

[3] Homepage of IBM, Inc., http://www.ibm.com

[4] Advanced Television Systems Committee, ATSC Digital Television Standard Document A/54, 1995, http://www.atsc.org

[5] Dutton, H. J. R., and Lenhard, P.: "Asynchronous Transfer Mode," (Prentice Hall, NJ, 1995), 2nd edn.

[6] Sykas, E. D., Vlakos, K. M., and Hillyard, M. J.: "Overview of ATM networks: functions and procedures," *Computer Communications*, **14** (10), 1991, pp. 615-626

[7] International Technology Roadmap for Semiconductors, 2000 update, http://public.itrs.net/Files/2000UpdateFinal/

[8] Homepage of Rapid IO consortium, http://www.rapidio.org

[9] Homepage of Infiniband consortium, http://www.infinibandta.org

[10] IEEE Std 802.3-2000, "IEEE Standard for CSMA/CD," *Institute of Electrical and Electronics Engineers*, 2000, http://grouper.ieee.org/groups/802/3/index.html.

[11] Imai, K., Ito, T., Kasahara, H., and Morita, N.: "ATMR: Asynchronous transfer mode protocol," *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 785-798

[12] Slosiar, R., Potts, M., and Beeler, R.: "MD3Q: A distributed queueing protocol with full channel capacity re-use and guarantee of bandwidth", *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 799-815

[13] Hopper, A., and Needham, R.: "The Cambridge Fast Ring Networking System," *IEEE Transactions on Computers*, **37** (10), 1988, pp. 1214-1223

[14] Adams, J. L.: “Orwell,” *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 771-784

[15] van As, H. R., Lemppenau, W. W., Schindler, H. R., and Zafiropulo, P.: “CRMA-II: A MAC protocol for ring-based Gb/s LANs and MANs,” *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 831-840

[16] IEEE Standards for Local and Metropolitan Area Networks: Distributed Queue Dual Bus (DQDB) Subnetwork of Metropolitan Area Network (MAN), 902.6-1990, *Institute of Electrical and Electronics Engineers*, 1990

[17] Hahne, E. L., Choudhury, A., and Maxemchuk, N. F.: “Improving the fairness of DQDB networks,” *Proceedings of IEEE Conference on Computer Communications Infocom ‘90*, 1990, pp. 175-184

[18] Watson, G., Banks, D., Calamvokis, C., Dalton, C., Edwards A., and Lumley, J.: "AAL5 at a gigabit for a kilobuck," *Journal of High Speed Networks*," **3** (2), 1994, pp. 127-145

[19] Ofek, Y.: "Overview of the MetaRing architecture," *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 817-830

[20] ANSI X3.148-1988, “Fiber Distributed Data Interface (FDDI) - Token Ring Physical Layer,” *American National Standards Institute*, 1988

[21] Davids, P., Meuser, T., and Spaniol, O.: "FDDI: status and perspectives," *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 657-677

[22] Hutchison, J. D., Baldwin, C., and Thompson, B. W.: “Development of the FDDI Physical Layer,” *Digital Technical Journal*, **3** (2), 1991, pp. 1-13

[23] Greaves, D. J., and Zielinski, K.: "The Cambridge Network: an overview and preliminary performance," *Computer Networks and ISDN Systems*, **25** (10), 1993, pp. 1127-1133

[24] IEEE P802.3z Gigabit Task Force, http://grouper.ieee.org/groups/802/3/z/index.html, *Institute of Electrical and Electronics Engineers*, 1998

[25] Walker, R. C., Hsieh, K.-C., Knotts, T. A., and Yen, C.-S.: "A 10 Gb/s Si-Bipolar TX/RX Chipset for Computer Data Transmission," *Proceedings of 1998 IEEE International Solid-State Circuits Conference*, 1998, (IEEE cat# 98CH36156), pp. 302-303

[26] Gigabit Ethernet Alliance, http://www.gigabit-ethernet.org

[27] Sano, B., Madhavan, B., and Levi, A. F. J.: "8 Gbps CMOS interface for parallel fiber-optic links," *Electronics Letters*, 1996, **32** (24), pp. 2262-2263

[28] Sano, B., and Levi, A. F. J.: "Networks for the professional campus environment," *Multimedia Technology for Applications*, (IEEE Press, Piscataway, NJ, 1998), pp. 413-427

[29] Blood Jr., W. R., 'MECL System Design Handbook,' (Motorola, HB205, 1988) Rev 1

[30] Kanjamala, A. P., and Levi, A. F. J.: "Subpicosecond skew in multimode fibre ribbon for synchronous data transmission," *Electronics Lett*ers, **31** (16), 1995, pp. 1376-1377

[31] IEEE Std 1596-1992, "IEEE Standard for Scalable Coherent Interface (SCI)," *Institute of Electrical and Electronics Engineers*, August 1993.

[32] Gustavson, D. B., and Li, Q.: "The scalable coherent interface (SCI)", *IEEE Communications Magazine*, **34** (8), 1996, pp. 52-63

[33] Engebretsen, D. R., Kuchta, D. M., Booth, R. C., Crow, J. D., and Nation, W. G.: "Parallel fiber-optic SCI links," *IEEE Micro*, **16** (1), 1996, pp. 20-26

[34] Cecchi, D. R., Dina, M., and Preuss, C. W.: "1 GByte/s SCI Link in 0.8 μm BiCMOS", *Proceedings of 1995 IEEE International Solid-State Circuits Conference*, 1995, pp. 326-327

[35] High-performance parallel interface - 6400 Mbit/s physical layer (HIPPI-6400-PH),' *Technical Committee of Accredited Standards*, X3T11, May 1996

[36] HIPPI-6400-OPT Working Draft T11-1, Project 1249-D, http://www.cic-5.lanl.gov, Rev 0.7, *American National Standards Institute*, 1998, http://www.hippi.org/c6400OPT.html

[37] Yang, G. M., MacDougal, M. H., and Dapkus, P. D.: "Ultra-low threshold vertical cavity surface emitting lasers obtained with selective oxidation," *Electronics Letters*, **31** (11), 1995, pp. 886-888

[38] Deppe, D. G., Huffaker, D. L., Deng, H. Y., Deng, Q., and Oh, T. H.: "Ultra-low threshold current vertical cavity surface emitting lasers for photonic integrated circuits," *IEICE Transactions on Electronics*, **E80-C** (5), 1997, pp. 664-674

[39] Hahn, K., Giboney, K. S., Wilson, R. E., Straznicky, J., Wong, E. G., Tan, M. R., Kaneshiro, K. T., Dolfi, D. W., Mueller, E. H., Plotts, A. E., Murray, D. D., Marchegiano, J. E., Booth, B. L., Sano, B. J., Madhavan, B., Raghavan, B., and Levi, A. F. J.: "Gigabyte/s Data Communications with POLO Parallel Optical Link," in *Proceedings of the 46th Electronics Components and Technology Conference*, 1996 (IEEE cat# 96CH35931), pp. 301-307

[40] Buckman, L., Yuen, A., Giboney, K., Rosenberg, P., Straznicky, J., Wu, K., and Dolfi, D.:“Parallel Optical Interconnects,” in *Hot Interconnects 6 Symposium*, Stanford University, Palo Alto, California,1998

[41] Buckman, L. A., Giboney, K. S., Straznicky, J., Simon, J., Schmit, A. J., Zhang, X. J., Corzine, S. W., Dolfi, D. W., Madhavan, B., and Kiamilev, F., “Parallel Optical Interconnects,” *Conference on Lasers and Electro-Optics, San Francisco*, 2000, pp. 535-536

[42] Drogemuller, K., Kuhl, D., Blank, J., Ehlert, M., Kraeker, T., Hohn, J., Klix, D., Plickert, V., Melchior, L., Schmale, I., Hildebrandt, P., Heinemann, M., Schiefelbein, F. P., Leininger, L., Wolf, H.-D., Wipiejewski, T., and Ebberg, A.:“Current progress of advanced high-speed parallel optical links for computer clusters and switching systems,” *2000 Electronic Components and Technology Conference*, 2000, pp. 1227-1235

[43] Siemens Semiconductor Group: “Fiber optics data book 1998-1999”, p. 153, Berlin, Germany

[44] Drogemuller, K., Kuhl, D., Blank, J., Ehlert, M., Kraeker, T., Hohn, J., Klix, D., Plickert, V., Melchior, L., Schmale, I., Hildebrandt, P., Heinemann, M., Schiefelbein, F. P., Leininger, L., Wolf, H.-D., Wipiejewski, T., and Ebberg, A.: “Current progress of advanced high speed parallel optical links for computer clusters and switching systems,” *Proceedings of 2000 Electronic Components and Technology Conference*, 2000, pp. 1227-1235

[45] Kuchta, D. M., Crow, J., Pepeljugoski, P., Stawiasz, K., Trewhella, J., Booth, D., Nation, W., DeCusatis, C., and Muszynski, A.: "Low cost 10 Gigabit/s optical interconnects for parallel processing," *Proceedings of 5th international conference on massively parallel processing*, 1998, pp. 210-215

[46] Weigandt, T. C., Kim, B., and Gray, P. R., "Analysis of timing jitter in CMOS ring oscillators," *Proceedings of 1994 IEEE International Symposium on Circuits and Systems ISCAS'94*, 1994, pp. 27-30

[47] McNeill, J. A., "Jitter in ring oscillators," *IEEE Journal of Solid-State Circuits*, **32** (6), pp. 870-879

[48] Madhavan, B.: *Ph.D. Thesis*, University of Southern California, June 2000

[49] Madhavan, B., and Levi, A. F. J.: "Link components for a 2.5 Gb/s/channel 12-wide parallel optical interface in 0.5 μm CMOS," *Proceedings of Conference on Lasers and Electro-Optics*, 2000, Paper CThT1, p. 533

[50] Lee, T. H., and Hajimiri, A.: "Oscillator phase noise: a tutorial," *IEEE Journal of Solid-State Circuits*, **35** (3), 2000, pp. 326-336

[51] Razavi, B.: "A study of phase noise in CMOS oscillators," *IEEE Journal of Solid-State Circuits*, **31** (3), 1996, pp. 331-343

[52] Herzel, F., and Razavi, B.: "Oscillator jitter due to supply and substrate noise," *Proceedings of 1998 IEEE Custom Integrated Circuits Conference CICC'98*, 1998, pp. 489-492

[53] Razavi (Ed.), B.: "Monolithic phase locked loops and clock recovery circuits," (IEEE Press, Piscataway, NJ), 1996

[54] Homepage of 10 Gigabit Ethernet Alliance, http://www.10gea.org

[55] Aloisio, A., Cevenini, F., and Fiore, D. J.: "Bus in a new light", *IEEE Transactions on Nuclear Science*, **47** (2), 2000, pp. 309-312

[56] XILINX, Inc.: "The programmable logic data book 1998", (San Jose, CA, 1998)

[57] Jonsson, M.: Ph. D. Thesis, Department of Computer Engineering, Chalmers University of Technology, Goteborg, Sweden, 1999, pp. 143-222

[58] Schwartz, D. B., Chun, K. Y., Choi, N., Diaz, D., Planer, S., Raskin, G. and Shook, S. G.: "OPTOBUS I: performance of a 4 Gb/s optical interconnect," *Proceedings of Massively Parallel Processing using Optical Interconnections (MPPOI '96)*, 1996, pp. 256-263

[59] Shrikhande, K. V., White, I. M., Wonglumsom, D., Gemelos, S. M., Rogge, M. S., Fukashiro, Y., Avenarius, M., and Kazovsky, L. G., "HORNET: a packet-over-WDM multiple access metropolitan area network," *IEEE Journal on selected areas in communications*, **18** (10), 2000, pp. 2004-2016

[60] White, I. M., Rogge, M. S., Shrikhande, K., Fukashiro, Y., Wonglumsom, D., An, F.-T., and Kazovsky, L. G., "Experimental demonstration of a novel media access protocol for HORNET: a packet-over-WDM multiple-access MAN ring," *IEEE Photonics Technology Letters*, **12** (9), 2000, pp. 1264-1266

[61] Voo, T., and Toumazou, C.: 'High-speed current mirror resistive compensation technique,' *Electronics Letters*, **31** (4), 1995, pp. 248-250

[62] Avant Hspice User manual, 1998

[63] Homepage of Cadence, Inc., http://www.cadence.com

[64] Yuan, J., and Svensson, C.: "High-Speed CMOS circuit technique," *IEEE Journal of Solid State Circuits*, **24** (1), 1989, pp. 62-71

[65] Homepage of Kyocera, Inc., http://www.kyocera.com

[66] Applied Micro Circuits Corporation homepage, http://www.amcc.com, "S5933 PCI controller data book," (San Diego, CA, 1996)

[67] Shanley, T., and Anderson, D.: "PCI System Architecture," (Addison-Wesley Publishing Company, 1995), 3rd edn.

[68] BlueWater Systems, Inc., http://www.bluewatersystems.com, "WinDK User's Manual," (Edmonds, WA, 1996)

[69] The Tolly Group, http://www.tolly.com, "MultiSwitch 900/VNswitch 900 Fast Ethernet and Multi-topology switching performance," Report #7303, 1997, pp. 1-6

[70] Madhavan, B., and Levi, A. F. J.: '55 Gbps/cm data bandwidth density interface in 0.5 μm CMOS for advanced parallel optical interconnects,' *Electronics Letters*, **34** (19), 1998, pp. 1846-1847

[71] Brustoloni, J., and Bershad, B.: http://reports-archive.adm.cs.cmu.edu/cs.html, "Simple Protocol Processing for High-Bandwidth low-latency networking," CMU-CS-93-132, School of Computer Science, CMU, March 1992

[72] Clark, D., Jacobson, V., Romkey, J., and Salwen, H.: "An Analysis of TCP processing overhead," *IEEE Communications Magazine*, **27** (6), 1989, pp. 23-29

[73] Kay, J., and Pasquale, J.: "The Importance of Non-Data Touching Processing Overheads in TCP/IP," in *Proceedings of ACM SIGCOMM '93 Computer Communication Review*, **23** (4), 1993, pp. 259-268

[74] Draft Standard for Low-Voltage Differential Signals (LVDS) for Scalable Coherent Interface (SCI), (IEEE, New York, 1992), Draft 1.26 IEEE P1596.3-1995

[75] Raghavan, B., Kim, Y.-G., Chuang, T.-Y., Madhavan, B., and Levi, A. F. J.: "A Gbyte/s parallel fiber-optic network interface for mutimedia applications," *IEEE Network Magazine*, **13** (1), 1999, pp. 20-28

[76] Sidman, S., Spaderna, D., Miller, J., and Jenkins, D.: "FIFOs - innovation through architecture," *IEEE Electro International Conference Record*, 1991, pp. 142-143

[77] Wadell, B. C., "Transmission Line Design Handbook," (Artech House, Norwood, MA), 1991, pp. 11-17, 47-51, 93-101, 125-128

[78] Hoffman, R. K.: "Handbook of Microwave Integrated Circuits," (Artech House, Norwood, MA, 1987)

[79] Matick, R. E.: "Transmission lines for digital and communication networks," (McGraw-Hill, New York, 1969)

[80] Pucell, R. A., Masse, D. J., and Hartwig, C. P.: "Losses in Microstrip", *IEEE Tran. on Microwave Theory and Techniques*, **MTT-16** (6), 1968, pp. 342 - 350

[81] Pucell, R. A., Masse, D. J., and Hartwig, C. P.: "Corrections to losses in microstrip", *IEEE Tran. on Microwave Theory and Techniques*, **MTT-16** (12), 1968, p. 1064

[82] Wheeler, H. A.: "Formulas for the skin effect", *Proc. IRE*, **30**, 1942, pp. 412-424

[83] Bakoglu, H. B.: “Circuits, Interconnections and Packaging for VLSI,” (Addison Wesley Publishing Company, Reading, MA, 1990)

[84] Yuan, J., and Svensson, C.: “New Single-Clock CMOS Latches and Flip-flops with Improved Speed and Power Savings,” *IEEE Journal of Solid State Circuits*, **32** (1), 1997, pp. 62-69

[85] Afghahi, M., and Svensson, C.: "Performance of Synchronous and Asynchronous Schemes for VLSI Systems," *IEEE Transactions on Computers*, **41** (7), 1992, pp. 858-872

[86] Lemppenau, W. W., van As, H. R., and Schindler, H. R.: "A 2.4 Gbit/s ATM implementation of the CRMA-II dual-ring LAN and MAN," *Proceedings of 11th Annual European Conference on Fibre Optic Communiations and Networks, EFOC/LAN’93*, 1993, pp. 274-281

[87] Zurfluh, E. A., Cideciyan, R. D., Dill, P., Heller, R., Lemppenau, W., Mueller, P., Schindler, H. R., and Zafiropulo, P.: “The IBM Zurich Laboratory’s 1.13 Gb/s LAN/MAN prototype,” *Computer Networks and ISDN Systems*, **26** (2), 1993, pp. 163-183

[88] Kleinrock, L.: “The latency/bandwidth tradeoff in gigabit networks,” *IEEE Communications Magazine*, **30** (4), 1992, pp. 36-40

[89] Friedman, A. D., and Menon, P. R.: “Theory and design of switching circuits,” (Computer Science Press, Inc., Woodland Hills, CA, 1975)

[90] Mealy, G. H.: “A method for synthesizing sequential circuits,” *Bell System Technical Journal*, **34**, 1955, pp. 1054-1079

[91] Moore, E. F.: “Gedanken experiments on sequential machines,” in Automata Studies, C. E. Shannon and J. McCarthy (Eds.), (Princeton University Press, Princeton, New Jersey, 1956), pp. 129-153

[92] Fair, H., and Bailey, D.: “Clocking design and analysis for a 600 MHz Alpha microprocessor,” *Proceedings of 1998 IEEE International Solid-State Circuits Conference*, 1998 (IEEE cat# 98CH36156), pp. 398-399, p. 473

[93] Nass, R.: “Ring architecture connects up to 128 PCI buses,” Electronic Design, Nov. 3, 1997

[94] Hewlett Packard Company, Convex Division: “Exemplar Architecture,” (Richardson, Texas, January 1997), 1st edn.

[95] Castaneda, R., Zhang, X., and Hoover Jr., J. M.: “A comparative evaluation of hierarchical network architecture of the HP-Convex Exemplar,” *Proceedings of IEEE International Conference on Computer Design, ICCD ‘97*, 1997, pp. 258-266

[96] HP 9000 V-Class Information Library, “http://www.unixservers.hp.com/highend/vclass/infolibrary/index.html”.

[97] Scott, S.: “The Gigaring Channel,” *IEEE Micro*, **16** (1), 1996, pp. 27-34

[98] Homepage of IEEE 802.17 Resilient Packet Ring Working Group, http://www.ieee802.org/rprsg/index.html.

[99] Burd, T.: “General processor information,” http://bwrc.eecs.berkeley.edu/CIC/summary/, 1994

[100] Amrutur, B. S., and Horowitz, M. A.: “Speed and power scaling of SRAMs”, *IEEE Transactions on solid-state circuits*, **35** (2), 2000, pp. 175-185

[101] Dobberpuhl, D., Witek, R. T., Allmon, R., Anglin, R., Bertucci, D., Britton, S., Chao, L., Conrad, R., Dever, D., Gieseke, B., Hassoun, S. M. N., Hoeppner, G., Kuchler, K., Ladd, M., Leary, M., Madden, L., McLellan, E., Meyer, D., Montanaro, J., Priore, D., Rajagopalan, V., Samudrala, S., and Santhanam, S.: “A 200-MHz 64-b Dual-issue CMOS microprocessor,” *IEEE Journal of Solid-State Circuits*, **27** (11), 1992, pp. 1555-1567

[102] Sullivan, S., Johnson, B., Reid, D., and Taylor, S.: “A 1.8 V, 2.0 ns cycle, 32 KB embedded memory with interleaved castout/reload,” *Proceedings of the 1999 IEEE Custom Integrated Circuits Confrence*, 1999, pp. 235-238

[103] De, V., and Borkar, S.: “Technology and design challenges for low power and high performance,” *Proceedings of International symposium on low-power electronics and design*, 1999, pp. 163-168

[104] Burnett, D., Erington, K., Subramanian, C., and Baker, K.:“Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits,” *Proceedings of 1994 Symposium on VLSI Technology*, 1994, pp. 15-16

[105] Mizuno, T., Iwase, M., Niiyama, H., Shibata, T., Fujisaki, K., Nakasugi, T., Toriumi, A., and Ushiku, Y.: “Performance fluctuations of 0.10 μm MOSFETs- limitation of 0.1 μm ULSIs”, *Proceedings of 1994 Symposium on VLSI Technology*, 1994, pp. 13-14

[106] Meindl, J. D., De, V. K., Wills, D. S., Eble, J. C., Tang, X., Davis, J. A., Austin, B., and Bhavnagarwala, A. J.: "The impact of stochastic dopant and interconnect distributions on gigascale integration," *Proceedings of 1997 International Solid-State circuits conference*, 1997, pp. 232-233, 463

[107] Hamzaoglu, F., Ye, Y., Keshavarzi, A., Zhang, K., Narendra, S., Borkar, S., Stan, M., and De, V.: "Dual-V_T SRAM cells with full-swing single-ended bit line sensing for high-performance on-chip cache in 0.13 μm technology generation," *Proceedings of International symposium on low-power electronics and design*, *ISLPED'00*, 2000, pp. 15-19

[108] Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V.,: "On the self-similarity of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, **2** (1), 1994, pp. 1-15

[109] Paxson, V., and Floyd, S.: "Wide-area traffic: the failure of Poisson modeling", *IEEE/ACM Transactions on Networking,* **3** (3), 1995, pp. 226-244

[110] Jain, R., and Routhier, S. A.: "Packet trains: measurements and a new model for computer network traffic," *IEEE Journal on Selected Areas in Communications*, **4** (6), 1986, pp. 986-995

[111] Gusella, R.: "A characterization of the variability of packet arrival processes in workstation networks," Ph.D. dissertation, University of California, Berkeley, 1990.

[112] Gusella, R.: "A measurement study of diskless workstation traffic on an Ethernet," *IEEE Transactions on Communications*, **38** (9), 1990, pp. 1557-1568

[113] Gusella, R.: "Characterizing the variability of arrival processes with indexes of dispersion," *IEEE Journal on Selected Areas in Communications*, **9** (2), 1991, pp. 203-211

[114] Willinger, W., Taqqu, M. S., Sherman, R., and Wilson, D. V.: "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, **5** (1), 1997, pp. 71-86

[115] Brichet, F., Roberts, J., Simonian, A., and Veitch, D.: "Heavy traffic analysis of a fluid queue fed by on/off sources with long-range dependence," *Queueing Systems*, **23 (**1-4), 1996, pp. 197-215

[116] Pruthi, P., and Erramilli, A.: “Heavy-tailed on/off source behavior and self-similar traffic,” *Proceedings of 1995 IEEE International Conference on Communications ICC’95*, Seattle, June 1995

[117] Taqqu, M. S.: “Self-similar processes,” In S. Kotz and N. Johnson, editors, Encyclopedia of Statistical Sciences, **8**, Wiley, New York, 1987.

[118] Tanabe, A., Umetani, M., Fujiwara, I., Ogura, T., Kataoka, K., Okihara, M., Sakuraba, H., Endoh, T., and Masuoka, F.: “A 10 Gb/s demultiplexer IC in 0.18 μm CMOS using current mode logic with tolerance to the threshold voltage fluctuation,” *Proceedings of 2000 IEEE International Solid-State Circuits Conference*, 2000, pp. 62- 63

[119] Seno, K., Knorpp, K., Shu, L.-L., Teshima, N., Kihara, H., Sato, H., Miyaji, F., Takeda, M., Sasaki, M., Tomo, Y., Chuang, P. T., and Kobayashi, K.: “A 9-ns 16-Mb CMOS SRAM with offset-compensated current sense amplifier,” *IEEE Journal of Solid-State Circuits*, **28** (11), 1993, pp. 1119-1124

[120] Taur, Y.: “The incredible shrinking transistor,” *IEEE Spectrum*, **36** (7), 1999, pp. 25-29

[121] Hajimiri, A., and Heald, R.: “Design issues in cross-coupled inverter pair sense amplifier,” *Proceedings of the 1998 IEEE International Symposium on Circuits and Systems, ISCAS ‘98*, **2**, 1998, pp. 149-152

[122] Dally, W. J., and Poulton, J., “Transmitter equalization for 4-Gbps signalling,” *IEEE Micro*, **17** (1), 1997, pp. 48-56

[123] Dally, W. J., Poulton, J., and Tell, S., “A tracking clock recovery scheme for 4- Gbps signaling,” *IEEE Micro*, **18** (1), 1998, pp. 25-27

[124] Gedney, R. W., McElroy, J. B., and Winkler, P. E., “The implications of roadmapping on university research,” *Proceedings of 48th IEEE International conference on electronic components and technology*, 1998, pp. 638-642

[125] X3T9.3 Task Group of ANSI: Fibre Channel Physical and Signaling Interface (FC-PH), Rev. 4.2, 1993.

[126] Veendrick, H.: “The behavior of flip-flops used as synchronizers and prediction of their failure rate,” *IEEE Journal of Solid State Circuits*, **SC-15** (2), 1980, pp. 169-176

[127] Flanagan, S. T.: “Synchronization reliability in CMOS technology,” *IEEE Journal of Solid State Circuits*, **SC-20** (4), 1985, pp. 880-882

[128] Kacprzak, T., and Albicki, A.: “Analysis of metastable operation in RS CMOS flip-flops,” *IEEE Journal of Solid-State circuits*, **SC-22** (1), 1987, pp. 57-64

[129] Portmann, C. L., and Meng, T. H. Y.: “Metastability in CMOS library elements in reduced supply and technology scaled applications,” *IEEE Journal of Solid-State circuits*, **30** (1), 1995, pp. 39-46

[130] Wellheuser, C.: “Metastability performance of clocked FIFOs,” Texas Instruments Application Note.

[131] Dike, C., and Burton, E.: “Miller and noise effects in a synchronizing flip-flop,” *IEEE Journal of Solid-state circuits*, **34** (6), 1999, pp. 849-855

[132] Taur, Y., Mii, Y.-J., Frank, D. J., Wong, H.-S., Buchanan, D. A., Wind, S. J., Rishton, S. A., Sai-Halasz, G. A., and Nowak, E. J.: “CMOS scaling into the 21st century: 0.1 μm and beyond,” *IBM Journal of Research and development*, **39** (1-2), 1995, pp. 245-260

[133] Keyes, R. W.: “The effect of randomness in the distribution of impurity atoms on FET thresholds,” *Journal of Applied Physics*, **8** (3), 1975, pp. 251-259

[134] Saito, M., Ogawa, J., Gotoh, K., Kawashima, S., and Tamura, H.: “Technique for controlling effective V_{th} in multi-Gbit DRAM sense amplifier,” *Proceedings of 1996 symposium on VLSI circuits*, 1996, pp. 106-107

[135] Optical Internetworking Forum, Implementation Agreement OIS-VSR4-01.0, 2000, http://www.oiforum.com

[136] Optical Internetworking Forum, Implementation Agreement OIS-VSR4-03.0, 2000, http://www.oiforum.com

[137] Walker, R. C., Stout, C. L., Wu, J.-T., Lai, B., Yen, C.-S., Hornak, T., and Petruno, P. T.: “A two-chip 1.5-GBd serial link interface,” *IEEE Journal of Solid-state circuits*, **27** (12), 1992, pp. 1805-1811

Bibliography

Adams, J. L.: “Orwell,” *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 771-784

Afghahi, M., and Svensson, C.: "Performance of Synchronous and Asynchronous Schemes for VLSI Systems," *IEEE Transactions on Computers*, **41** (7), 1992, pp. 858-872

Aloisio, A., Cevenini, F., and Fiore, D. J.: “Bus in a new light”, *IEEE Transactions on Nuclear Science*, **47** (2), 2000, pp. 309-312

Amrutur, B. S., and Horowitz, M. A.: “Speed and power scaling of SRAMs”, *IEEE Transactions on solid-state circuits*, **35** (2), 2000, pp. 175-185

Bakoglu, H. B.: “Circuits, Interconnections and Packaging for VLSI,” (Addison Wesley Publishing Company, Reading, MA, 1990)

Blood Jr., W. R., ‘MECL System Design Handbook,’ (Motorola, HB205, 1988) Rev. 1

Brichet, F., Roberts, J., Simonian, A., and Veitch, D.: “Heavy traffic analysis of a fluid queue fed by on/off sources with long-range dependence,” *Queueing Systems*, **23 (**1-4), 1996, pp. 197-215

Brustoloni, J., and Bershad, B.: http://reports-archive.adm.cs.cmu.edu/cs.html, "Simple Protocol Processing for High-Bandwidth low-latency networking," CMU-CS-93-132, School of Computer Science, CMU, March 1992

Buckman, L., Yuen, A., Giboney, K., Rosenberg, P., Straznicky, J., Wu, K., and Dolfi, D.:“Parallel Optical Interconnects,” in *Hot Interconnects 6 Symposium*, Stanford University, Palo Alto, California,1998

Buckman, L. A., Giboney, K. S., Straznicky, J., Simon, J., Schmit, A. J., Zhang, X. J., Corzine, S. W., Dolfi, D. W., Madhavan, B., and Kiamilev, F., “Parallel Optical Interconnects,” *Conference on Lasers and Electro-Optics, San Francisco*, 2000, pp. 535-536

Burd, T.: “General processor information,” http://bwrc.eecs.berkeley.edu/CIC/summary/, 1994

Burnett, D., Erington, K., Subramanian, C., and Baker, K.:“Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits,” *Proceedings of 1994 Symposium on VLSI Technology*, 1994, pp. 15-16

Castaneda, R., Zhang, X., and Hoover Jr., J. M.: “A comparative evaluation of hierarchical network architecture of the HP-Convex Exemplar,” *Proceedings of IEEE International Conference on Computer Design, ICCD ‘97*, 1997, pp. 258-266

Cecchi, D. R., Dina, M., and Preuss, C. W.: “1 GByte/s SCI Link in 0.8 μm BiCMOS”, *Proceedings of 1995 IEEE International Solid-State Circuits Conference*, 1995, pp. 326-327

Clark, D., Jacobson, V., Romkey, J., and Salwen, H.: “An Analysis of TCP processing overhead," *IEEE Communications Magazine*, **27** (6), 1989, pp. 23-29

Dally, W. J., and Poulton, J., “Transmitter equalization for 4-Gbps signalling,” *IEEE Micro*, **17** (1), 1997, pp. 48-56

Dally, W. J., Poulton, J., and Tell, S., “A tracking clock recovery scheme for 4- Gbps signaling,” *IEEE Micro*, **18** (1), 1998, pp. 25-27

Davids, P., Meuser, T., and Spaniol, O.: "FDDI: status and perspectives," *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 657-677

De, V., and Borkar, S.: “Technology and design challenges for low power and high performance,” *Proceedings of International symposium on low-power electronics and design*, 1999, pp. 163-168

Deppe, D. G., Huffaker, D. L., Deng, H. Y., Deng, Q., and Oh, T. H.: "Ultra-low threshold current vertical cavity surface emitting lasers for photonic integrated circuits," *IEICE Transactions on Electronics*, **E80-C** (5), 1997, pp. 664-674

Dike, C., and Burton, E.: “Miller and noise effects in a synchronizing flip-flop,” *IEEE Journal of Solid-state circuits*, **34** (6), 1999, pp. 849-855

Dobberpuhl, D., Witek, R. T., Allmon, R., Anglin, R., Bertucci, D., Britton, S., Chao, L., Conrad, R., Dever, D., Gieseke, B., Hassoun, S. M. N., Hoeppner, G., Kuchler, K., Ladd, M., Leary, M., Madden, L., McLellan, E., Meyer, D., Montanaro, J., Priore, D., Rajagopalan, V., Samudrala, S., and Santhanam, S.: "A 200-MHz 64-b Dual-issue CMOS microprocessor," *IEEE Journal of Solid-State Circuits*, **27** (11), 1992, pp. 1555-1567

Drogemuller, K., Kuhl, D., Blank, J., Ehlert, M., Kraeker, T., Hohn, J., Klix, D., Plickert, V., Melchior, L., Schmale, I., Hildebrandt, P., Heinemann, M., Schiefelbein, F. P., Leininger, L., Wolf, H.-D., Wipiejewski, T., and Ebberg, A.:"Current progress of advanced high-speed parallel optical links for computer clusters and switching systems," *2000 Electronic Components and Technology Conference*, 2000, pp. 1227-1235

Drogemuller, K., Kuhl, D., Blank, J., Ehlert, M., Kraeker, T., Hohn, J., Klix, D., Plickert, V., Melchior, L., Schmale, I., Hildebrandt, P., Heinemann, M., Schiefelbein, F. P., Leininger, L., Wolf, H.-D., Wipiejewski, T., and Ebberg, A.: "Current progress of advanced high speed parallel optical links for computer clusters and switching systems," *Proceedings of 2000 Electronic Components and Technology Conference*, 2000, pp. 1227-1235

Dutton, H. J. R., and Lenhard, P.: "Asynchronous Transfer Mode," (Prentice Hall, NJ, 1995), 2nd edn.

Engebretsen, D. R., Kuchta, D. M., Booth, R. C., Crow, J. D., and Nation, W. G.: "Parallel fiber-optic SCI links," *IEEE Micro*, **16** (1), 1996, pp. 20-26

Fair, H., and Bailey, D.: "Clocking design and analysis for a 600 MHz Alpha microprocessor," *Proceedings of 1998 IEEE International Solid-State Circuits Conference*, 1998 (IEEE cat# 98CH36156), pp. 398-399, p. 473

Flanagan, S. T.: "Synchronization reliability in CMOS technology," *IEEE Journal of Solid State Circuits*, **SC-20** (4), 1985, pp. 880-882

Friedman, A. D., and Menon, P. R.: "Theory and design of switching circuits," (Computer Science Press, Inc., Woodland Hills, CA, 1975)

Gedney, R. W., McElroy, J. B., and Winkler, P. E., "The implications of roadmapping on university research," *Proceedings of 48th IEEE International conference on electronic components and technology*, 1998, pp. 638-642

Greaves, D. J., and Zielinski, K.: "The Cambridge Network: an overview and preliminary performance," *Computer Networks and ISDN Systems*, **25** (10), 1993, pp. 1127-1133

Gusella, R.: “A characterization of the variability of packet arrival processes in workstation networks,” Ph.D. dissertation, University of California, Berkeley, 1990

Gusella, R.: “A measurement study of diskless workstation traffic on an Ethernet,” *IEEE Transactions on Communications*, **38** (9), 1990, pp. 1557-1568

Gusella, R.: “Characterizing the variability of arrival processes with indexes of dispersion,” *IEEE Journal on Selected Areas in Communications*, **9** (2), 1991, pp. 203-211

Gustavson, D. B., and Li, Q.: “The scalable coherent interface (SCI)”, *IEEE Communications Magazine*, **34** (8), 1996, pp. 52-63

Hahn, K., Giboney, K. S., Wilson, R. E., Straznicky, J., Wong, E. G., Tan, M. R., Kaneshiro, K. T., Dolfi, D. W., Mueller, E. H., Plotts, A. E., Murray, D. D., Marchegiano, J. E., Booth, B. L., Sano, B. J., Madhavan, B., Raghavan, B., and Levi, A. F. J.: "Gigabyte/s Data Communications with POLO Parallel Optical Link," in *Proceedings of the 46th Electronics Components and Technology Conference*, 1996 (IEEE cat# 96CH35931), pp. 301-307

Hahne, E. L., Choudhury, A., and Maxemchuk, N. F.: “Improving the fairness of DQDB networks,” *Proceedings of IEEE Conference on Computer Communications Infocom ‘90*, 1990, pp. 175-184

Hajimiri, A., and Heald, R.: “Design issues in cross-coupled inverter pair sense amplifier,” *Proceedings of the 1998 IEEE International Symposium on Circuits and Systems, ISCAS ‘98*, **2**, 1998, pp. 149-152

Hamzaoglu, F., Ye, Y., Keshavarzi, A., Zhang, K., Narendra, S., Borkar, S., Stan, M., and De, V.: “Dual-V_T SRAM cells with full-swing single-ended bit line sensing for high-performance on-chip cache in 0.13 μm technology generation,” *Proceedings of International symposium on low-power electronics and design, ISLPED’00*, 2000, pp. 15-19

Herzel, F., and Razavi, B.: “Oscillator jitter due to supply and substrate noise,” *Proceedings of 1998 IEEE Custom Integrated Circuits Conference CICC’98*, 1998, pp. 489-492

Hoffman, R. K.: “Handbook of Microwave Integrated Circuits,” (Artech House, Norwood, MA, 1987)

Hopper, A., and Needham, R.: “The Cambridge Fast Ring Networking System,” *IEEE Transactions on Computers*, **37** (10), 1988, pp. 1214-1223

Hutchison, J. D., Baldwin, C., and Thompson, B. W.: “Development of the FDDI Physical Layer,” *Digital Technical Journal*, **3** (2), 1991, pp. 1-13

Imai, K., Ito, T., Kasahara, H., and Morita, N.: “ATMR: Asynchronous transfer mode protocol,” *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 785-798

Kanjamala, A. P., and Levi, A. F. J.: "Subpicosecond skew in multimode fibre ribbon for synchronous data transmission," *Electronics Lett*ers, **31** (16), 1995, pp. 1376-1377

Kay, J., and Pasquale, J.: “The Importance of Non-Data Touching Processing Overheads in TCP/IP," in *Proceedings of ACM SIGCOMM ’93 Computer Communication Review*, **23** (4), 1993, pp. 259-268

Keyes, R. W.: “The effect of randomness in the distribution of impurity atoms on FET thresholds,” *Journal of Applied Physics*, **8** (3), 1975, pp. 251-259

Kleinrock, L.: “The latency/bandwidth tradeoff in gigabit networks,” *IEEE Communications Magazine*, **30** (4), 1992, pp. 36-40

Kuchta, D. M., Crow, J., Pepeljugoski, P., Stawiasz, K., Trewhella, J., Booth, D., Nation, W., DeCusatis, C., and Muszynski, A.: “Low cost 10 Gigabit/s optical interconnects for parallel processing,” *Proceedings of 5th international conference on massively parallel processing*, 1998, pp. 210-215

Jain, R., and Routhier, S. A.: “Packet trains: measurements and a new model for computer network traffic,” *IEEE Journal on Selected Areas in Communications*, **4** (6), 1986, pp. 986-995

Jonsson, M.: Ph. D. Thesis, Department of Computer Engineering, Chalmers University of Technology, Goteborg, Sweden, 1999, pp. 143-222

Kacprzak, T., and Albicki, A.: “Analysis of metastable operation in RS CMOS flip-flops,” *IEEE Journal of Solid-State circuits*, **SC-22** (1), 1987, pp. 57-64

Lee, T. H., and Hajimiri, A.: “Oscillator phase noise: a tutorial,” *IEEE Journal of Solid-State Circuits*, **35** (3), 2000, pp. 326-336

Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V.,: “On the self-similarity of Ethernet traffic (extended version),” *IEEE/ACM Transactions on Networking*, **2** (1), 1994, pp. 1-15

Lemppenau, W. W., van As, H. R., and Schindler, H. R.: "A 2.4 Gbit/s ATM implementation of the CRMA-II dual-ring LAN and MAN," *Proceedings of 11th Annual European Conference on Fibre Optic Communiations and Networks, EFOC/LAN’93*, 1993, pp. 274-281

Madhavan, B.: *Ph.D. Thesis*, University of Southern California, June 2000

Madhavan, B., and Levi, A. F. J.: “Link components for a 2.5 Gb/s/channel 12-wide parallel optical interface in 0.5 μm CMOS,” *Proceedings of Conference on Lasers and Electro-Optics*, 2000, Paper CThT1, p. 533

Madhavan, B., and Levi, A. F. J.: ‘55 Gbps/cm data bandwidth density interface in 0.5 μm CMOS for advanced parallel optical interconnects,’ *Electronics Letters*, **34** (19), 1998, pp. 1846-1847

Matick, R. E.: “Transmission lines for digital and communication networks,” (McGraw-Hill, New York, 1969)

McNeill, J. A., “Jitter in ring oscillators,” *IEEE Journal of Solid-State Circuits*, **32** (6), pp. 870-879

Mealy, G. H.: “A method for synthesizing sequential circuits,” *Bell System Technical Journal*, **34**, 1955, pp. 1054-1079

Meindl, J. D., De, V. K., Wills, D. S., Eble, J. C., Tang, X., Davis, J. A., Austin, B., and Bhavnagarwala, A. J.: “The impact of stochastic dopant and interconnect distributions on gigascale integration,” *Proceedings of 1997 International Solid-State circuits conference*, 1997, pp. 232-233, 463

Mizuno, T., Iwase, M., Niiyama, H., Shibata, T., Fujisaki, K., Nakasugi, T., Toriumi, A., and Ushiku, Y.: “Performance fluctuations of 0.10 μm MOSFETs- limitation of 0.1 μm ULSIs”, *Proceedings of 1994 Symposium on VLSI Technology*, 1994, pp. 13-14

Moore, E. F.: “Gedanken experiments on sequential machines,” in Automata Studies, C. E. Shannon and J. McCarthy (Eds.), (Princeton University Press, Princeton, New Jersey, 1956), pp. 129-153

Nass, R.: “Ring architecture connects up to 128 PCI buses,” Electronic Design, Nov. 3, 1997

Ofek, Y.: "Overview of the MetaRing architecture," *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 817-830

Paxson, V., and Floyd, S.: “Wide-area traffic: the failure of Poisson modeling”, *IEEE/ACM Transactions on Networking*, **3** (3), 1995, pp. 226-244

Portmann, C. L., and Meng, T. H. Y.: “Metastability in CMOS library elements in reduced supply and technology scaled applications,” *IEEE Journal of Solid-State circuits*, **30** (1), 1995, pp. 39-46

Pruthi, P., and Erramilli, A.: “Heavy-tailed on/off source behavior and self-similar traffic,” *Proceedings of 1995 IEEE International Conference on Communications ICC’95*, Seattle, June 1995

Pucell, R. A., Masse, D. J., and Hartwig, C. P.: “Losses in Microstrip”, *IEEE Tran. on Microwave Theory and Techniques*, **MTT-16** (6), 1968, pp. 342 - 350

Pucell, R. A., Masse, D. J., and Hartwig, C. P.: “Corrections to losses in microstrip”, *IEEE Tran. on Microwave Theory and Techniques*, **MTT-16** (12), 1968, p. 1064

Raghavan, B., Kim, Y.-G., Chuang, T.-Y., Madhavan, B., and Levi, A. F. J.: “A Gbyte/s parallel fiber-optic network interface for mutimedia applications,” *IEEE Network Magazine*, **13** (1), 1999, pp. 20-28

Razavi, B.: “A study of phase noise in CMOS oscillators,” *IEEE Journal of Solid-State Circuits*, **31** (3), 1996, pp. 331-343

Razavi (Ed.), B.: “Monolithic phase locked loops and clock recovery circuits,” (IEEE Press, Piscataway, NJ), 1996

Saito, M., Ogawa, J., Gotoh, K., Kawashima, S., and Tamura, H.: “Technique for controlling effective V_{th} in multi-Gbit DRAM sense amplifier,” *Proceedings of 1996 symposium on VLSI circuits*, 1996, pp. 106-107

Sano, B., Madhavan, B., and Levi, A. F. J.: "8 Gbps CMOS interface for parallel fiber-optic links," *Electronics Letters*, 1996, **32** (24), pp. 2262-2263

Sano, B., and Levi, A. F. J.: “Networks for the professional campus environment,” *Multimedia Technology for Applications*, (IEEE Press, Piscataway, NJ, 1998), pp. 413-427

Schwartz, D. B., Chun, K. Y., Choi, N., Diaz, D., Planer, S., Raskin, G. and Shook, S. G.: “OPTOBUS I: performance of a 4 Gb/s optical interconnect,” *Proceedings of Massively Parallel Processing using Optical Interconnections (MPPOI ‘96)*, 1996, pp. 256-263

Scott, S.: “The Gigaring Channel,” *IEEE Micro*, **16** (1), 1996, pp. 27-34

Seno, K., Knorpp, K., Shu, L.-L., Teshima, N., Kihara, H., Sato, H., Miyaji, F., Takeda, M., Sasaki, M., Tomo, Y., Chuang, P. T., and Kobayashi, K.: “A 9-ns 16-Mb CMOS SRAM with offset-compensated current sense amplifier,” *IEEE Journal of Solid-State Circuits*, **28** (11), 1993, pp. 1119-1124

Shanley, T., and Anderson, D.: "PCI System Architecture," (Addison-Wesley Publishing Company, 1995), 3rd edn.

Shrikhande, K. V., White, I. M., Wonglumsom, D., Gemelos, S. M., Rogge, M. S., Fukashiro, Y., Avenarius, M., and Kazovsky, L. G., “HORNET: a packet-over-WDM multiple access metropolitan area network,” *IEEE Journal on selected areas in communications*, **18** (10), 2000, pp. 2004-2016

Sidman, S., Spaderna, D., Miller, J., and Jenkins, D.: “FIFOs - innovation through architecture,” *IEEE Electro International Conference Record*, 1991, pp. 142-143

Slosiar, R., Potts, M., and Beeler, R.: “MD3Q: A distributed queueing protocol with full channel capacity re-use and guarantee of bandwidth”, *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 799-815

Sullivan, S., Johnson, B., Reid, D., and Taylor, S.: “A 1.8 V, 2.0 ns cycle, 32 KB embedded memory with interleaved castout/reload,” *Proceedings of the 1999 IEEE Custom Integrated Circuits Confrence*, 1999, pp. 235-238

Sykas, E. D., Vlakos, K. M., and Hillyard, M. J.: “Overview of ATM networks: functions and procedures,” *Computer Communications*, **14** (10), 1991, pp. 615-626

Tanabe, A., Umetani, M., Fujiwara, I., Ogura, T., Kataoka, K., Okihara, M., Sakuraba, H., Endoh, T., and Masuoka, F.: “A 10 Gb/s demultiplexer IC in 0.18 μm CMOS using current mode logic with tolerance to the threshold voltage fluctuation,” *Proceedings of 2000 IEEE International Solid-State Circuits Conference*, 2000, pp. 62- 63

Taqqu, M. S.: “Self-similar processes,” In S. Kotz and N. Johnson, editors, Encyclopedia of Statistical Sciences, **8**, Wiley, New York, 1987

Taur, Y.: “The incredible shrinking transistor,” *IEEE Spectrum*, **36** (7), 1999, pp. 25-29

Taur, Y., Mii, Y.-J., Frank, D. J., Wong, H.-S., Buchanan, D. A., Wind, S. J., Rishton, S. A., Sai-Halasz, G. A., and Nowak, E. J.: “CMOS scaling into the 21st century: 0.1 μm and beyond,” *IBM Journal of Research and development*, **39** (1-2), 1995, pp. 245-260

van As, H. R., Lemppenau, W. W., Schindler, H. R., and Zafiropulo, P.: “CRMA-II: A MAC protocol for ring-based Gb/s LANs and MANs,” *Computer Networks and ISDN Systems*, **26** (6-8), 1994, pp. 831-840

Veendrick, H.: “The behavior of flip-flops used as synchronizers and prediction of their failure rate,” *IEEE Journal of Solid State Circuits*, **SC-15** (2), 1980, pp. 169-176

Voo, T., and Toumazou, C.: ‘High-speed current mirror resistive compensation technique,’ *Electronics Letters*, **31** (4), 1995, pp. 248-250

Wadell, B. C., “Transmission Line Design Handbook,” (Artech House, Norwood, MA), 1991, pp. 11-17, 47-51, 93-101, 125-128

Walker, R. C., Hsieh, K.-C., Knotts, T. A., and Yen, C.-S.: "A 10 Gb/s Si-Bipolar TX/RX Chipset for Computer Data Transmission," *Proceedings of 1998 IEEE International Solid-State Circuits Conference*, 1998, (IEEE cat# 98CH36156), pp. 302-303

Walker, R. C., Stout, C. L., Wu, J.-T., Lai, B., Yen, C.-S., Hornak, T., and Petruno, P. T.: “A two-chip 1.5-GBd serial link interface,” *IEEE Journal of Solid-state circuits*, **27** (12), 1992, pp. 1805-1811

Watson, G., Banks, D., Calamvokis, C., Dalton, C., Edwards A., and Lumley, J.: "AAL5 at a gigabit for a kilobuck," *Journal of High Speed Networks*," **3** (2), 1994, pp. 127-145

Weigandt, T. C., Kim, B., and Gray, P. R., “Analysis of timing jitter in CMOS ring oscillators,” *Proceedings of 1994 IEEE International Symposium on Circuits and Systems ISCAS’94*, 1994, pp. 27-30

Wellheuser, C.: “Metastability performance of clocked FIFOs,” Texas Instruments Application Note.

Wheeler, H. A.: “Formulas for the skin effect”, *Proc. IRE*, **30**, 1942, pp. 412-424

White, I. M., Rogge, M. S., Shrikhande, K., Fukashiro, Y., Wonglumsom, D., An, F.-T., and Kazovsky, L. G., “Experimental demonstration of a novel media access protocol for HORNET: a packet-over-WDM multiple-access MAN ring,” *IEEE Photonics Technology Letters*, **12** (9), 2000, pp. 1264-1266

Willinger, W., Taqqu, M. S., Sherman, R., and Wilson, D. V.: “Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level,” *IEEE/ACM Transactions on Networking*, **5** (1), 1997, pp. 71-86

Yang, G. M., MacDougal, M. H., and Dapkus, P. D.: "Ultra-low threshold vertical cavity surface emitting lasers obtained with selective oxidation," *Electronics Letters*, **31** (11), 1995, pp. 886-888

Yuan, J., and Svensson, C.: “High-Speed CMOS circuit technique,” *IEEE Journal of Solid State Circuits*, **24** (1), 1989, pp. 62-71

Yuan, J., and Svensson, C.: “New Single-Clock CMOS Latches and Flipflops with Improved Speed and Power Savings,” *IEEE Journal of Solid State Circuits*, **32** (1), 1997, pp. 62-69

Zurfluh, E. A., Cideciyan, R. D., Dill, P., Heller, R., Lemppenau, W., Mueller, P., Schindler, H. R., and Zafiropulo, P.: “The IBM Zurich Laboratory’s 1.13 Gb/s LAN/MAN prototype,” *Computer Networks and ISDN Systems*, **26** (2), 1993, pp. 163-183

Appendix A Metastability

The elasticity buffer used in the LAC enables a distributed clocking scheme with plesiochronous clocking between adjacent nodes (small frequency variations between adjacent ring nodes). The operation of the elasticity buffer as explained in Section 5.2.2 on page 102 is briefly as follows. When a packet (in a slot) is received by the elasticity buffer from the network, write operation commences. Read operations on the elasticity buffer commence after a certain number of words have been written into the buffer. A handshake signal generated by the write clock domain is synchronized by the read clock domain to specify that read operations can commence.

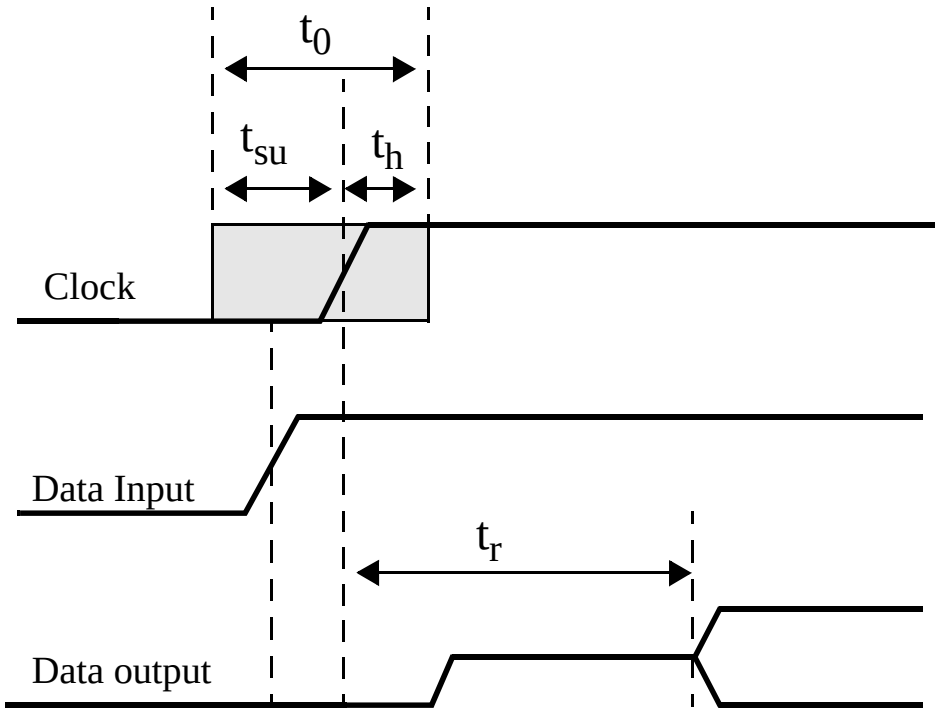


Figure A-1: Diagram showing parameters influencing synchronization error due to metastability. Setup time for data is given by t_{su}, hold time is t_r, metastability error window duration is given by t_0, resolution time for output is given by t_r.

Synchronization across clock boundaries can typically lead to metastability related issues. Metastability occurs across two asynchronous clock domains when a signal generated in the first domain is not resolved properly by the second and hence the latched signal is indeterminate (neither a logic high nor logic low but with a signal amplitude that is somewhere in between those two values).

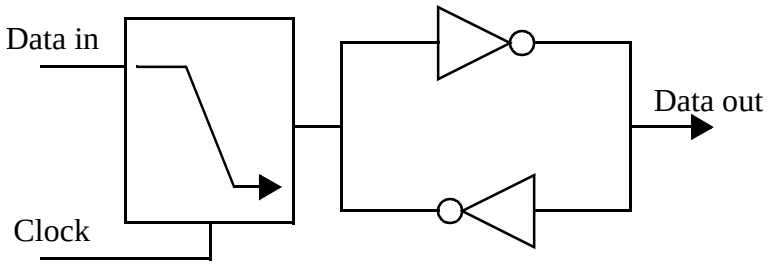


Figure A-2: shows a simple static CMOS latch with cross-coupled inverter pair output.

The schematic in Figure A-2 shows a simple CMOS latch with cross-coupled pair inverter output. A data transition that occurs within a window of duration t_0 as shown in Figure A-1 may lead to metastability problems wherein the output may not resolve completely but change to an intermediate level where it remains for an indefinite period of time. Calculation of the failure rate due to metastability is studied in [126][127][128][129][130][131].

The probability of an output remaining in a metastable state for an infinite amount of time is zero. The probability p(r) of resolving within a time t_r is given in [126] by the equation

$$p(r) = e^{-(t_r/\tau)}$$

where τ is a circuit time constant and has been shown to be inversely proportional to the gain-bandwidth product of the circuit.

For a single-stage synchronizer and an asynchronous data edge with a uniform probability density within the latching stage clock period, the failure rate due to metastability errors is given by the relation,

$$\frac{1}{failurerate} = MTBF_1 = \frac{e^{(t_r/\tau)}}{t_0 f_c f_d}$$

where f_c = clock frequency, f_d = asynchronous data edge frequency, t_0 is the metastability error window as shown in Figure A-1.

For a two-stage synchronizer, the failure rate is given by the equation,

$$\frac{1}{failurerate} = MTBF_2 = \frac{e^{(t_{r1}/\tau)}}{t_0 f_c f_d} e^{(t_{r2}/\tau)}$$

where t_{r1} is the resolve time for the first stage of the synchronizer and t_{r2} is the resolve time allowed for the second stage of the synchronizer in excess of the propagation delay.

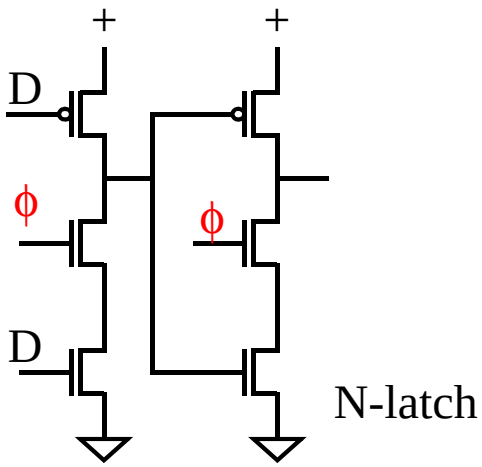


Figure A-3: Schematic of a true single phase clocked latch used for synchronizing in the elasticity buffer across the write and read clock domains.

The latch used for synchronizing across elasticity buffer clock domains in the ring network interface chip is shown in Figure A-3. To find τ, the clock was permanently enabled high and the gain bandwidth product of a single-stage latch was evaluated by applying a small-signal input voltage at D. From simulations, unity gain resulted at an input signal frequency of 1800 MHz, resulting in a τ of 0.56 ns. From simulations, the parameter t_0 is approximately 0.35 ns.

In the elasticity buffer, the signals retimed using the TSPC latches are to enable and disable read operations. To enable read operations, a flag is generated within the write clock domain after a certain number of words (in the current design, at least three words) have been written into the elasticity buffer. To disable read operations, the gray-coded write counters and read counters are compared. When the incoming packet has been completely written in, the write pointer is disabled and the read pointer is allowed to catch

up. If the write and read clock frequencies are perfectly matched, the number of cycles available to perform this computation is equal to the initial spacing between write and read pointers (in this case three). If accumulated over an extended period of time, however frequency mismatches may cause this spacing to vary and potentially reduce the number of clock cycles thus available. For a slot size of 1 kB used in the current design and clock frequency variations of under 0.05 % as is typical for crystal clock frequencies used in this design, this initial separation will not change and hence at least three clock cycles will be available over which the write counter value is stable prior to allowing the read pointer to catch up with the write pointer. The commencement of read operations can however be delayed by a cycle due to the synchronization and this leads to a variation in the idle gap between slots. If left unchecked the idle gap can eventually disappear or corrupt slots. The smoother module maintains a minimum idle spacing between slots and thus prevents failure. Synchronization errors in the elasticity buffer are more likely due to overruns or underruns due to accumulated frequency variations over extended slot times.

Appendix B LAC Package and Pin-Out

The pin out of the LA Chip is shown in Figure B-1. The LA Chip has been designed for direct connection to the PONI module and to the host interface glue logic board without signal line crossover. This allows the use of a single signal line layer for both high-speed impedance matched lines and the lower speed TTL signal lines to the host interface.

The LA chip separates the analog and digital power supplies to suppress noise on the digital power lines from the constant bias currents on the analog power supplies. PWR1 is used for analog power while PWR2 is used for digital power. Termination voltage (i.e. Vtt = 2 volts) is supplied by signal pins and capacitively decoupled inside the carrier package. Ground lines are dedicated pins on the QFP 248 pin package, manufactured by Kyocera Inc. under USC specifications for this project, along with PWR1 and PWR2 meaning they can not be reassigned.

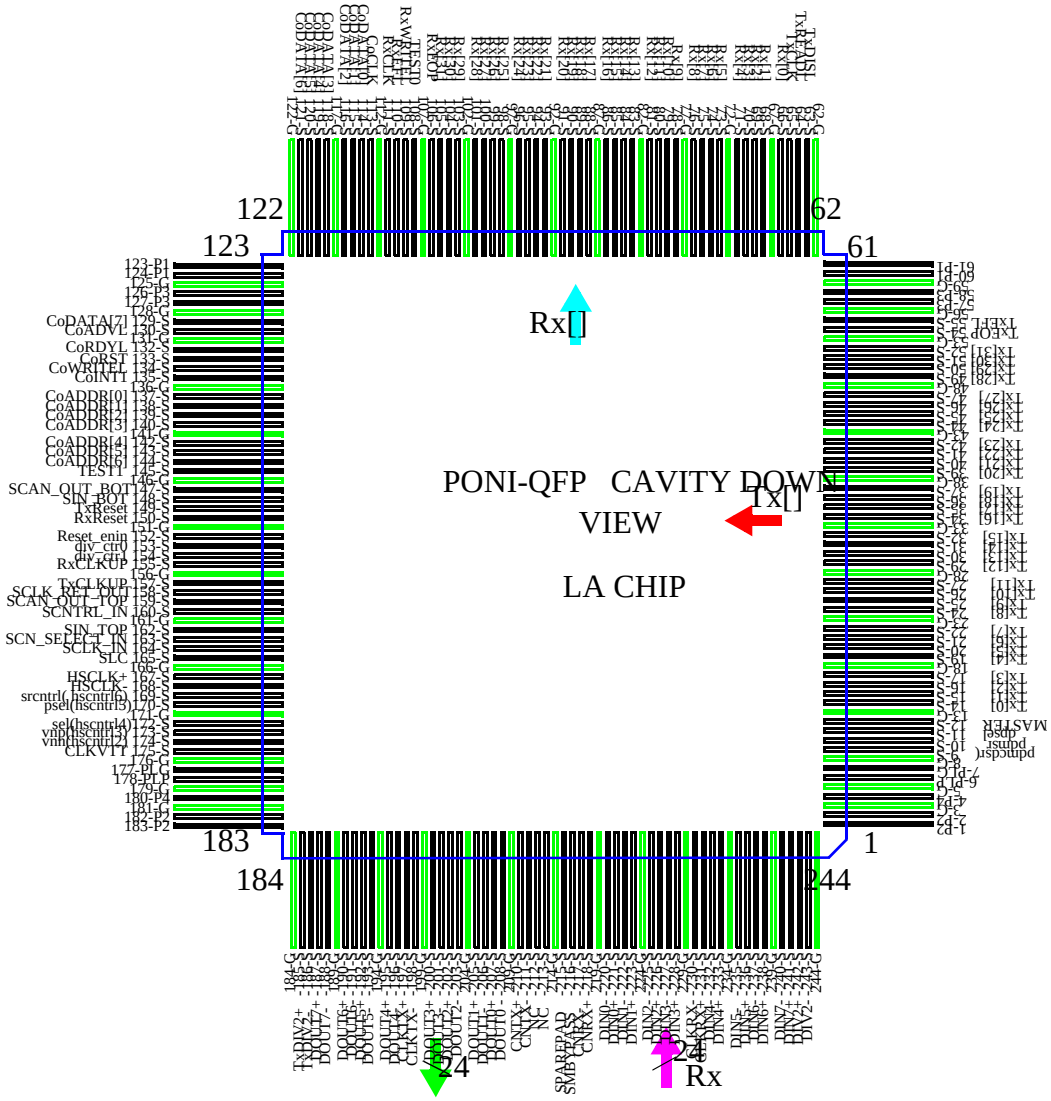


Figure B-1: LAC Package Pin-Out. The LAC uses a 244-pin QFP manufactured by Kyocera Inc. It features separate power shelves for analog, digital and TTL VDD.

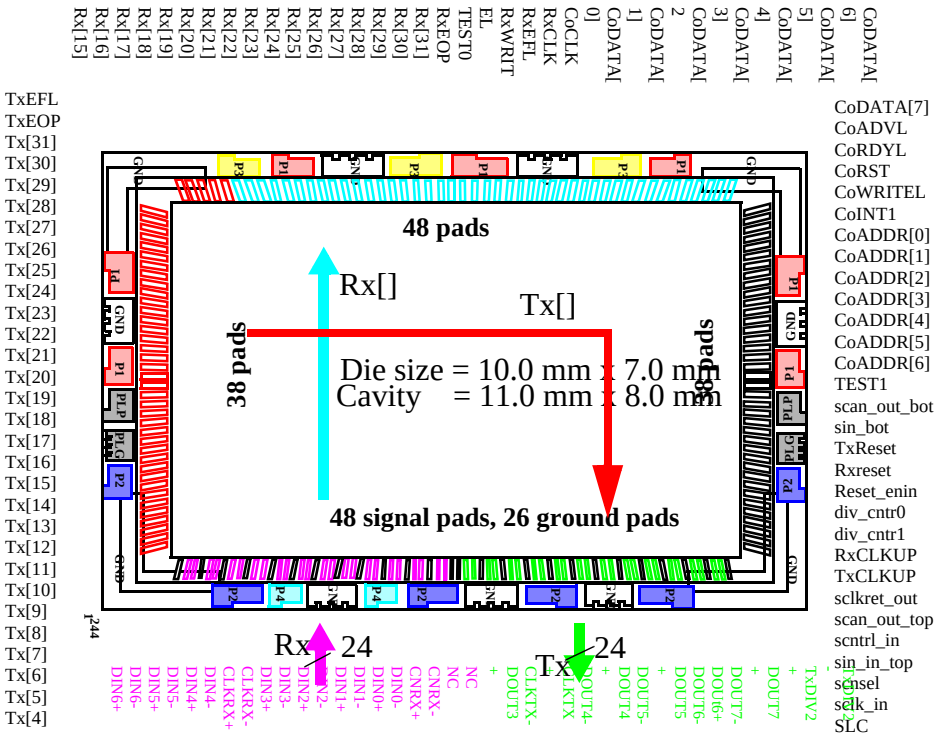


Figure B-2: LAC cavity up view -a view of package for the PONI LAC IC. Signal pads at the bottom (names with prefix B and prefix B) interface to the bipolar circuitry. The CLK signal at the LVDS interface is the seventh signal pair from the outside. LVDS signal pads (names with a prefix of L) arrive at the top and sides. Signal pads in red are control, test and PLL signals. Signal pads at the bottom are to be designed on a 50 Ω differential strip-line basis: every pair is separated by a ground line. The signals on the other three sides are to be designed as 50 Ω single-ended strip lines. Ground pads are in black. On the upper terrace, cyan is Power 4, yellow is Power 3, blue is Power 2, red is Power 1, black is PLL-power and PLL-ground.