

Photonic Interconnection Networks for Applications in Heterogeneous Utility Computing Systems

Cathy Chen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Science

COLUMBIA UNIVERSITY

2015

©2015
Cathy Chen
All rights reserved

ABSTRACT

Photonic Interconnection Networks for Applications in Heterogeneous Utility Computing Systems

Cathy Chen

Growing demands in heterogeneous utility computing systems in future cloud and high performance computing systems are driving the development of processor-hardware accelerator interconnects with greater performance, flexibility, and dynamism. Recent innovations in the field of utility computing have led to an emergence in the use of heterogeneous compute elements. By leveraging the computing advantages of hardware accelerators alongside typical general purpose processors, performance efficiency can be maximized. The network linking these compute nodes is increasingly becoming the bottleneck in these architectures, limiting the hardware accelerators to be restricted to localized computing.

A high-bandwidth, agile interconnect is an imperative enabler for hardware accelerator delocalization in heterogeneous utility computing. A redesign of these systems' interconnect and architecture will be essential to establishing high-bandwidth, low-latency, efficient, and dynamic heterogeneous systems that can meet the challenges of next-generation utility computing.

By leveraging an optics-based approach, this dissertation presents the design and implementation of optically-connected hardware accelerators (OCHA) that exploit the distance-independent energy dissipation and bandwidth density of photonic transceivers, in combination with the flexibility, efficiency and data parallelization offered by optical networks. By replacing the electronic buses with an optical interconnection network, architectures that delocalize hardware accelerators can be created that are otherwise infeasible.

With delocalized optically-connected hardware accelerator nodes accessible by processors at run time, the system can alleviate the network latency issues plague current heterogeneous systems. Accelerators that would otherwise sit idle, waiting

for its master CPU to feed it data, can instead operate at high utilization rates, leading to dramatic improvements in overall system performance.

This work presents a prototype optically-connect hardware accelerator module and custom optical-network-aware, dynamic hardware accelerator allocator that communicate transparently and optically across an optical interconnection network. The hardware accelerators and processor are optimized to enable hardware acceleration across an optical network using fast packet-switching. The versatility of the optical network enables additional performance benefits including optical multicasting to exploit the data parallelism found in many accelerated data sets. The integration of hardware acceleration, heterogeneous computing, and optics constitutes a critical step for both computing and optics.

The massive data parallelism, application dependent-location and function, as well as network latency, and bandwidth limitations facing networks today complement well with the strength of optical communications-based systems. Moreover, ongoing efforts focusing on development of low-cost optical components and subsystems that are suitable for computing environment may benefit from the high-volume heterogeneous computing market. This work, therefore, takes the first steps in merging the areas of hardware acceleration and optics by developing architectures, protocols, and systems to interface with the two technologies and demonstrating areas of potential benefits and areas for future work. Next-generation heterogeneous utility computing systems will indubitably benefit from the use of efficient, flexible and high-performance optically connect hardware acceleration.

Contents

List of Figures	iv
List of Tables	ix
Glossary	x
Relevant Author Publications	xii
1 Introduction	1
1.1 Large-Scale Utility Computing	2
1.1.1 High-Performance Computing	2
1.1.2 Cloud Computing	4
1.1.3 Heterogeneous Utility Computing	7
1.2 Hardware Acceleration	9
1.2.1 Location	11
1.2.2 Communication	13
1.2.3 Programming Innovations	15
1.3 The Computing-Optics Interface	17
1.3.1 Optical Interconnection Networks	18
1.3.2 Optically Connected Hardware Accelerators	19
1.4 Scope	20
2 Optical Interconnection Networks for Heterogeneous Utility Computing	23
2.1 Optical Network Design	23
2.1.1 Switching Conventions	24

2.1.2	Scalability	28
2.2	Implications for Heterogeneous Computing Systems	29
2.3	Discussion	31
3	Dynamic Data on Optical Interconnection Networks	32
3.1	Background	32
3.2	Optical Interconnection Network Interface	34
3.3	Experimental Set Up	36
3.3.1	WiMax Data Generation	36
3.3.2	VLAN and ONIC	37
3.3.3	Results	39
3.4	Discussion	41
4	A Photonic Network for Hardware Accelerator Enabled Utility Computing	43
4.1	Background	44
4.2	Lessons from Previous Work	44
4.2.1	Bandwidth Mismatches	44
4.2.2	Burst Mode Receivers	45
4.3	Optically Connected Resources Module	46
4.3.1	FPGA Hardware Design	47
4.4	Experimental Set Up	49
4.4.1	XOR Phase-Encoded Header	50
4.5	Results	53
4.6	Discussion	54
5	FPGA Implemented Bidirectional OCHA	56
5.1	Background	56
5.2	Overview of OCHA	57
5.3	Experimental Set Up and Results	59
5.3.1	PRBS Generation	59
5.3.2	Wavelength-stripped Phase-encoded Header	60
5.3.3	Error Checker and Results	62
5.4	Discussion	65

6	Summary and Conclusion	67
6.1	Overview	67
6.2	Future Work	69
6.2.1	Architectural Design	69
6.2.2	Optically Connected Memory	70
6.2.3	Silicon Photonic Integration	70
6.2.4	Burst-Mode Receivers	71
6.2.5	Runtime Allocation Integration	72
6.2.6	Commercialization	72
6.3	Summary	73
	References	75

List of Figures

1.1	Performance of Top500 HPC Systems - The Performance (in FLOPs) of the first and 500th supercomputers on the Top500 list. As well as the sum of the two together. If these trends continue, an ExaFLOP machine could be possible in the next 5-10 years. Source: Top500.org	3
1.2	Optical Interconnects in IBM Power 775 - An IBM Power 775 system drawer, with eight router multichip modules (MCMs), each with 28 transmit and receive modules, with 12x10 Gb/s bandwidth. The inset shows the underlying glass-ceramic substrate. Source: IBM	5
1.3	Acceleration in HPC - Percentage of High Performance Computing Systems (surveyed by Intersect360 Research with Accelerators from 2009 to 2013. Between systems installed in 2011 and 2012, there was a doubling of systems with accelerators, in 2013 this increased by another 3%. Source: Intersect360 Research . . .	8
1.4	Photograph of NVIDIA GRID Server - Each server contains 12 GPUs, 20 of these servers are packed into a single GRID Gaming Rack, which in turn is capable of 200 TFLOPS of computing (equivalent to 700 Xbox 360s) [1].	11
1.5	GPU-acceleration Speed Up - Time to price a 15 year cancellable range accrual on a Constant Maturity Swap Spread, using a 2-factor Heath-Jarrow-Morton model with 1 million Monte Carlo simulation paths. The model uses a full term structure for volatilities and includes calibration of correlations. [2].	12

LIST OF FIGURES

1.6	JPEGs in Facebook - Illustration of the JPEG encoding in Facebook data centers. Each picture that is uploaded is regenerated as 4 JPEGs of varying sizes, for use in various parts of the sight. Source: Facebook	13
1.7	Multicast in Spade Algorithm - Multicasting in the Spade Algorithm for bargain discovery. Source: [3]	14
1.8	Power8 Die Photo - Die photo of the IBM Power8 chip, announced in August of 2013, to be used by the OpenPOWER Foundation for use in big data and cloud computing applications. The PCIe slot on the die can be seen in bright green on the middle bottom of the chip. Source: IBM	15
1.9	Gaming latency in NVIDIA GRID - Comparison of game latency of NVIDIA GRID, Cloud Gen 1, and Console +TV game. The network (light green) takes up approximately a fourth of the overall gaming latency time. Source: NVIDIA Corp	16
1.10	Block Diagram of Optically Connected Hardware Accelerators - Block level schematic depicting how a next-generation heterogeneous system can be connected to many optically-connected hardware accelerators across an optical interconnection network.	19
2.1	Photonic Switching Node- (A) Schematic and (B) photograph of the SOA-based switching node - routing information is encoded on the header wavelengths that are decoded through an OEO conversion, where a CPLD computes the logic to control a number of SOAs. This scheme allows for the elimination of many OEO conversions that limit current switching nodes.	25
2.2	Wavelength-Striped Message Format - Low-speed header switching wavelengths (controlled by the GPIO MICTOR on the FPGA) are combined with high-speed payload wavelengths using WDM	26

LIST OF FIGURES

2.3	3×3 switching - (A) Schematic of how the switching node would be configured in a 3×3 set up, with an FPGA as the control logic and (B) photograph of the 4x4 SOA-based switching node that this could be implemented on	27
3.1	Vision - (A) With an RSSI below the threshold value, the gamer is allocated one GPU for a GaaS application, but as she nears the base station and the RSSI increase (B) her game is dynamically streamed to two GPUs now that she has a RSSI above a certain threshold	34
3.2	Optical Interconnection Network Interface - (A) Schematic of the Optical Interconnection Network Interface showing the Myri-com 10 GE card, QSFP, top level logic of on the Stratix II FPGA and (B) photographs of the components used to build the set up, including the Stratix II Development board	35
3.3	Block Diagram of Optical-WiMax Test Bed - The architecture setup. Video from the client is dynamically streamed through the WiMAX basestation, transmitted over a VLAN and through an O-NIC and WDM encoded on the optical network, and then decoded at the end node. This process is transparent to the end users.	37
3.4	Software Stack - UDP packets are streamed from the Client using VLC. The packets reach the base station and are processed by the NetServ application module. The module polls for the downstream RSSI value (distance of client from the base station) periodically and modifies the packet changing the destination IP based on the RSSI value.	38
3.5	Optical-WiMax Test Bed - The physical network setup. A Vertex IV FPGA is used as the Optical-Network Interface Card (ONIC) and modulators are used to encode the WDM striped data.	39
3.6	Optical-WiMax Results- Eyes - the output eye diagrams for CH36-CH39 of the WiMAX generated video.	40

3.7	Optical-WiMax Results - At the end node, the CPU is listening on the two IP ports. The VLC video packets can be seen switching from the lower RSSI value IP destination to the stronger RSSI value IP destination after being streamed through the WiMax base station, a VLAN and a transparent optical network.	41
4.1	Picture of OCRM - A picture of the OCRM module featuring a Altera Stratix IV FPGA, DDR3, 10/100 Mb/s ethernet port, bi-directional transceivers, expansion ports for daughter cards, and MICTOR GPIO	48
4.2	Diagram Overlaying Picture of OCRM - A diagram of the OCRM overlaying a photograph of the module identifying the locations of the Altera Stratix IV FPGA, the DDR3, 10/100 Mb/s ethernet port, bi-directional transceivers, expansion ports and the MICTOR GPIO	49
4.3	Experimental Set-up - FPGA A modulates four payload channels and four network control wavelengths over a 2x2 actively switched network test-bed. FPGA B and the BERT receive these payloads from the optical network using four PIN-TIA receivers. .	50
4.4	Message Format of XOR Phase-encoded Header - Low-speed header wavelengths are phase-encoded and sampled on the positive edge of a phase-offset sample clock. These header wavelengths are combined with high-speed payload wavelengths using WDM.	51
4.5	Photograph of Experimental Set-up - FPGA A modulates four payload channels and four network control wavelengths from the laser source trays over a 2x2 actively switched network test-bed. FPGA B and the BERT receive these payloads from the optical network using four PIN-TIA/LA receivers.	52
4.6	XOR Phase Encoded Header - XOR-Phase-encoded header network control- for (a) switch to output A (b) multicast (c) switch to output B and (d) both off	53

LIST OF FIGURES

4.7	Packet Routing - input and output Packets when (a) switch to output A (b) multicast (c) switch to output	54
5.1	Architectural Design - (A) CPU nodes with separate optically-connected hardware accelerator nodes that can be dynamically configured. The hardware modules could either be configured as a scheme where (B) the hardware accelerators are organized as a central bank or (C) each CPU has a dedicated hardware accelerator that it rents out to the system when not in use.	58
5.2	Experimental Setup - FPGA 0 acts as the CPU emulator and modulates four payloads and four network control wavelengths over a 4x4 actively switched network testbed. The packets traverse a bidirectional optical network that utilizes a wavelength-stripped Phase-encoded header (see Fig. 2.). FPGAs A, B. and C act as the hardware accelerator nodes. They receive these payloads from the optical network and send an error count on a unique wavelength. FPGA 0 reads these wavelengths and confirms error-free propagation.	60
5.3	PRBS generated using Linear Feedback Shift Registers - A 40-bit PRBS generator. The XNOR gate provides feedback to the registers as it shifts from right to left. The maximal sequence consists of every possible state.	61
5.4	Phase-encoded header design scheme - On the positive edge of the sample clock, which is time delayed, the output SOAs are controlled by the state of their control bit. This allows the system to individually control each output port and allows for a logical multicast. In these scenarios, this system would multicast, switch to A and B, and turn all the ports off, respectively.	63
5.5	Data Packet Switching - Switching of a 100 us packet through the network in a multicast, a switch to A and B, and all off. The input packet can be seen in the top box, while the optical outputs can be seen in each scenario in the lower boxes.	64

List of Tables

2.1	Wavelength-striped header logic table - When the frame wavelength is valid, the output SOAs are controlled by their control bit. In the above table, these bits are labeled A and B. The logic explains what nodes would be switched to in each scenario. This combinational logic implementation allows for simplified logic in the CPLD. In later work, this scheme was modified to allow for fast switching of longer packet.	28
3.1	Logic for wavelength-striped control - Whenever the Frame address bit is on, indicating a valid packet, the address bit controls the destination address, indicating a switch to node A or node B (1 or 0, respectively)	39
4.1	XOR Phase-encoded Header logic table - On the positive edge of the sample clock, which is time delayed, the output SOAs are controlled by an XOR of their control bit with the frame bit. In the above table, these bits are labeled A and B. The logic explains what Nodes would be switched to in each scenario. This state is held until the next positive edge of the sample clock.	51
5.1	Phase-encoded Header logic table - On the positive edge of the sample clock, which is time delayed, the output SOAs are controlled by the state of their control bit. In the above table, these bits are labeled A2, A1, and A0. The logic explains what Nodes would be switched to in each scenario. This state is held until the next positive edge of the sample clock	62

Glossary

3D	Three Dimensional	GE	Gigabit Ethernet
ASIC	Application-specific Integrated Circuit	GPIO	General Purpose Input Output
BER	Bit Error Rate	GPP	General Purpose Processing
BERT	Bit Error Rate Tester	GPU	Graphics Processing Unit
CAPEX	Capital Expenditure	HaaS	Hardware as a Service
CMOS	Complementary Metal Oxide Semiconductor	HDL	Hardware Description Language
CPLD	Complex Programmable Logic Device	HHPC	Heterogenous High Performance Computer
CPU	Central Processing Unit	HPC	High-Performance Computing
CSA	Continuous Spectrum Analyzer	I/O	Input Output
DaaS	Data as a Service	ITU	International Telecommunications Union
DIMM	Dual In-line Memory Module	LA	Limiting Amplifier
E-O	Electrical-Optical	LAN	Local Area Network
FDL	Fiber Delay Line	LiNbO3	Lithium Niobate
FFP	Floating Point Processor	MAC	Media Access Control
FLOPS	Floating Point Operation per Second	MCM	Multichip Module
FPGA	Field-Programmable Gate Array	MEMs	Microelectromechanical systems
FPP	Floating Point Processing	MICTOR	Matched Impedance Connector
GaaS	Gaming as a Service	NoC	Network on Chip
		O-E	Optical-Electrical
		OCHA	Optically Connected Hardware Accelerators
		OCRM	Optically Connected Resource Module
		OEO	Optical-Electronic-Optical
		OIN	Optical Interconnection Network
		OPEX	Operating Expenditure
		OSA	Optical Spectrum Analyzer

GLOSSARY

PCIe	Peripheral Component Interconnect Express	SerDes	Serializer-Deserializer
PD	Photo Diode	SMF	Single-mode fiber
PHY	Physical Layer	SOA	Semiconductor Optical Amplifier
PIN	P-i-n photodiode	SQL	Structured Query Language
PLL	Phase Lock Loop	TIA	Trans Impedance Amplifier
PM	Phase Modulator	TL	Tunable Laser
PPG	Pulsed Pattern Generator	TLB	Translation Lookaside Buffer
PRBS	Pseudo-random Bit Sequence	UDP	User Datagram Protocol
QSFP	Quad Small Form-factor Pluggable	VLAN	Virtual Local Area Network
RSSI	Received Signal Strength Indication	WDM	Wavelength Division Multiplexing
SaaS	Software as a Service	WiMax	Worldwide Interoperability for Microwave Access

Relevant Author Publications

- C. Chen, J. Chan, H. Wang, K. Bergman, "A Photonic Interconnection Network for Hardware Accelerator Enabled Utility Computing," Optical Interconnects Conference 2013 WA2 (May 2013).
- H. Wang, C. Chen, K. Sripanidkulchai, S. Sahu, K. Bergman, "Dynamically Reconfigurable Photonic Resources for Optically Connected Data Center Networks," Optical Fiber Communication Conference (OFC) 2012 OTu1B.2 (Mar 2012).
- A. S. Garg, H. Wang, C. Chen, K. Bergman, "Experimental Demonstration of Attenuation-Based All-Optical Time-To-Live Indicator," ECOC Technical Digest 2011 We.10.P1.42 (Jul 27, 2011).
- M. S. Wang, A. Wang, B. G. Bathula, C. P. Lai, I. Baldine, C. Chen, D. Majumder, D. Gurkan, G. Rouskas, R. Dutta, K. Bergman, "Demonstration of QoS-Aware Video Streaming over a Metro-Scale Optical Network Using a Cross-Layer Architectural Design," National Fiber Optic Engineers Conference (NFOEC) NThC4 (Mar 2011).

Acknowledgements

The adventure that is graduate school is a long and winding one, and in my case, at least, took place over some of the most defining years of my life. However, learning how to be a scientist and engineer was only a small percentage of the knowledge that I gained at Columbia. For a lack of a better phrase, I became a grown up in graduate school. I learned a great deal these past six years, and to the many people that helped me get to this point, thank you from the bottom of my heart.

First and foremost, my advisor, Dr. Keren Bergman, deserves the most credit. Your guidance and support kept me motivated throughout the highs and lows of my graduate school years. You were always available when I needed you, steered me when I was wondering off course, and most importantly, never lost faith in me, even if I did. Like the best advisors, you knew when to push me and when to let me discover on my own, and trained me to be a scientist and engineer. Thank you for all the group meetings, lab dinners, celebrations, and above all, thank you for the past six years. Without you I certainly would not have gotten to where I am today.

Secondly, to my parents, John and Sue Chen. To my hero and father, John Chen, to whom this dissertation is dedicated, thank you. Even though you missed seeing me complete this work by a mere 6 months, I know that I would not be where I am today without you. Your unyielding support, advice, and guidance were what drove me throughout the past six years and to which I owe to greatest debt. I hope you know, wherever you are, that you are still a guiding force in my life, and will forever strive to make you proud. To my mom, Sue Chen, while we are different in many ways, and at times, you didn't understand what I did in lab all day, you have always been there for me in anyway you can, and that has meant the world.

To the staff of the Electrical Engineering Department, both past and present, you have truly be the most supportive and greatest group of people I have ever encountered. Specifically, I would like to thank Elsa Sanchez, whom I affectionately refer to as my Electrical Engineering department mommy, since my first day on campus, you have been my biggest cheerleader, a great listener, and an amazing champion. I will forever be thankful to the great community in Electrical Engineering we build in my years here.

I would like to thank the dissertation committee members, Professors Debasis Mitra, Gil Zussman, Dan Kilper, and Luca Carloni, who have provided me valuable input, both directly, and through their students, over the course of my graduate studies. In addition, I thank Professor Heinz and Professor Christine Hendon (nee Flemming) for their support and encouragement of our OSA/SPIE student chapter through the years. I would like to thank Dean Ellie Bastani, Dean Jonathan Stark, Dr. Jennifer Piro, Dr. Karen Singleton, and Dean Carlos J. Alonso for their support during my years active in graduate student government at Columbia.

I am indebted to Howard Wang, Dan Brunina, Caroline Lai, Noam Ophir, and Johnnie Chan, whom I affectionately refer to as my academic older brothers and sisters, for their mentorship early in my graduate career and throughout its completion. I am grateful to Atiyah Ahsan, Jan Janak, Wenjia Zhang, Elliot Katz, and Berk Birand for their assistance in the work that comprises this dissertation, and thankful to all of the members of the Lightwave Research Laboratory, past and present, for their collaboration and friendship.

Additionally, I would like to thank Robert Margolies, Atiyah Ahsan, and Lee Zhu (The 815 crew) for being my officemates, colleagues, and friends during my studies at Columbia. From Qualifying Exams to Dissertation Defenses, we faced it all together. We have shared the same office for the better part of five years, and my experiences here would have been vastly less fulfilling without you all as part of them.

To all my friends and colleagues that made this adventure the amazing journey it has been, thank you. Thank you for coming to the parties I organize and humoring me when I tell you I can play basketball, and going to trivia with me, and dim sum trips to Chinatown, and weekends at the beach and the hundreds

of other adventures we had together. These are some of my favorite memories of graduate school (besides the research, of course) and without you all, it certainly would not have been as great as it was.

Lastly, I would like to thank and acknowledge the National Science Foundation, and the Engineering Research Center for Integrated Access Networks (CIAN) for fellowship support; the Wei Family Foundation and Neil and Mandy Grossman for scholarship support; and the Institute of Electrical and Electronics Engineers, SPIE - The International Society for Optical Engineering, OSA-The Optical Society, and GENI- Global Environment for Network Innovations for travel support throughout the course of my graduate studies.

Special thanks to Intersect360 Research which provided copies of their reports for use in this dissertation.

”Live as if you were to die tomorrow, learn as if you were to live forever.”

Thank you all.

Cathy Chen
New York, NY
Spring 2015

To...

My Inspiration and Hero

John Chen, in memorium

Best. Daddy. Ever.

Chapter 1

Introduction

The rise of high-performance computing and cloud computing models has brought with it a trend towards utility computing. These processing-as-a-service systems offer many advantages over traditional computing models, and have increasingly utilized specialized hardware acceleration to increase computation efficiencies. However, today's electronic networks have limited the performance of these heterogeneous systems due to low bandwidth densities, distance energy dissipation, and data-rate-dependent energy depletion [4]. Consequently, specialized hardware has been limited to physically close locations in the utility architecture. However, oftentimes, the location, and thereby function of the hardware acceleration is a application-specific problem [4, 5].

Many systems are now held back by the network latency [1], spending much of the compute time waiting for data to be moved around the system. The interconnection network is becoming a bottleneck in efficiency in heterogeneous utility computing systems, causing increased latency in compute times. Communication between the CPU and hardware accelerators must be high bandwidth, low latency, and energy efficient [6]. While it is feasible for electronic interconnection networks to reach per-channel data rates up to 25 Gb/s [7], the power dissipation at these high bandwidths becomes overwhelming and contributes to increase overall system cost and complexity.

In contrast with current electronic designs, the large bandwidth-distance product enabled by an optical interconnect to the hardware accelerators can provide the bandwidth, latency, and efficiency necessary to support dynamic allocation

via an interconnection network. Thusly, a photonic interconnection network for heterogeneous computing is a optimal solution to current network limitations and computing efficiencies. The design maintains the through puts required by CPU-accelerator communications, and can enable the scaling of processing capacity by allowing a CPU to dynamically use more accelerators than are accessible electronically.

1.1 Large-Scale Utility Computing

Large-Scale Computing systems like High Performance Computing and Cloud Computing models allow used access to computing resources outside the traditional Capital Expenditure (CAPEX) model of buying computing resources that depreciate over a period of time. They instead move toward an Operating Expense (OPEX) model, in which a shared infrastructure is used in a pay-per-use model similar to traditional utilities like water or electricity [5].

Consequently known as utility computing, from a hardware allocation and pricing perspective, utility systems offer new aspects to consider in computing models. Firstly is the appearance of infinite computing resources that are available on-demand, that are quick enough to follow load surges, that eliminate the need for computing users to plan far ahead for provisioning.

Secondly, there is the elimination of up-front commitments to hardware resources by users, allowing companies or individuals to start small and increase hardware resources in cases when there is an increase in need. Additionally, is the ability to pay for computing resources on a short-term basis, and release them when not in use, thereby rewarding conservation [8]. Utility computing offers many advantages over traditional General Purpose Processing (GPP) models, allowing for the parallelization of a workload of an application, as well as being faster, more efficient, and cost effective [8].

1.1.1 High-Performance Computing

High-Performance Computers (HPCs), also known as supercomputers, are computers that are designed for optimized processing capacity, typically for calculation-

intensive operations in data-intensive research fields such as particle physics, geographic information systems, and biology [9]. Supercomputers today employ tens of thousands of processors to achieve targeted performance metrics, rated in floating-point operations per second (FLOPS). The current world record holder is the Tianhe-2 machine in Guangzhou, China with a peak performance of 22.96 PetaFLOPS. And, if current trends continue (see Figure 1.1), it is feasible an exa-scale machine will exist in the near future [10].

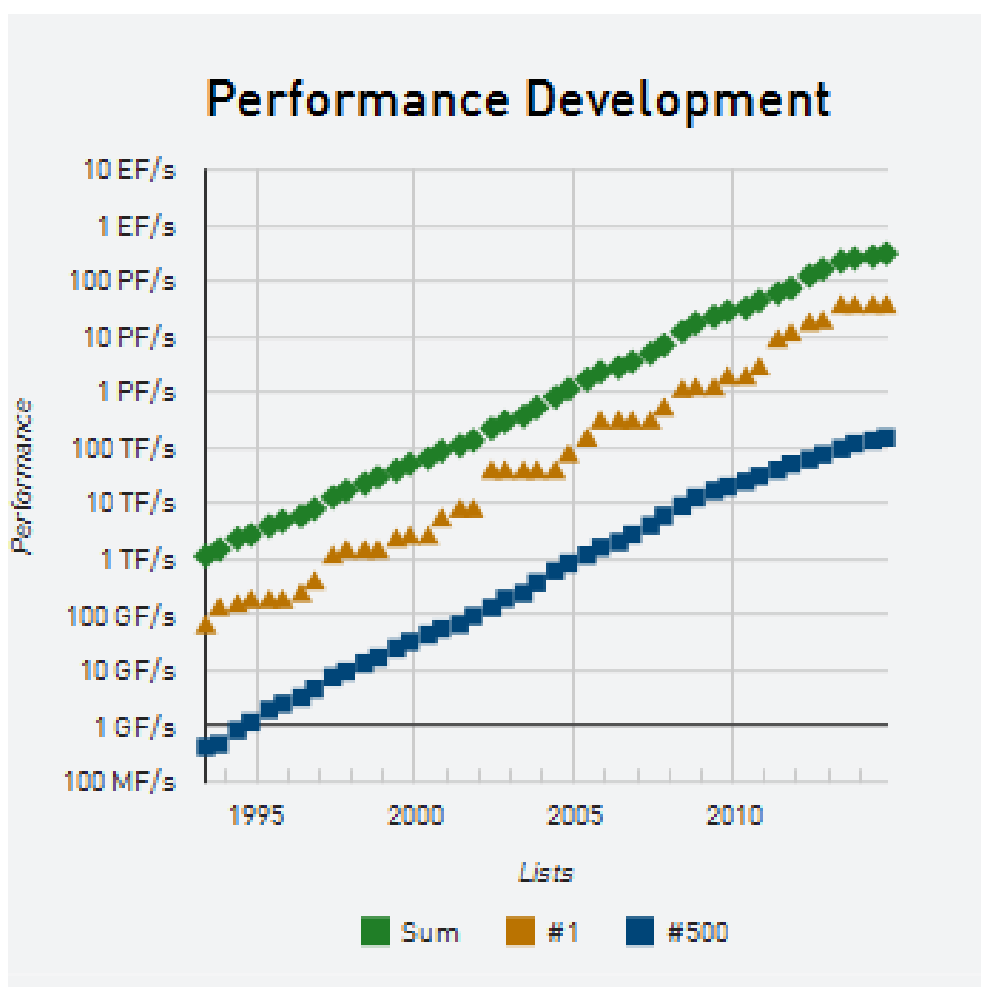


Figure 1.1: Performance of Top500 HPC Systems - The Performance (in FLOPs) of the first and 500th supercomputers on the Top500 list. As well as the sum of the two together. If these trends continue, an ExaFLOP machine could be possible in the next 5-10 years. Source: Top500.org

In many HPC systems, limitations in I/O performance are reshaping platforms, and as a result, hardware accelerators (see Section 1.1.3) and optical interconnects are propelling new architecture designs [11]. Channel data rates for off-chip interconnects have been steadily increasing in response to system needs, in order to improve cost per transported bit, power per transported bit and to meet bandwidth density requirements for wiring and area density. However, electrical interconnects become more difficult to scale past 10 Gb/s, due to frequency depend losses, frequency resonance effects, and crosstalk. Copper traces suffer from larger losses at higher frequencies due to the skin effect and dielectric losses [12].

Mitigating the issue with larger, fatter wiring exacerbates wiring density and routing problems. Optical interconnects, on the other hand, do not suffer from such strong signal degradation effects. They also provide other benefits such as smaller connector size, reducing cable bulk, and reduced electromagnetic interference. Due to the benefits of optical interconnects, the use of optics in large-scale HPC systems is increasing [12], in fact, the number of optical channels in a single HPC super computing system can be on par with the worldwide volume in parallel optical interconnects in a few years.

IBM's Power7-chip based Power 775 super computing system is an example on such system that is leveraging the advantages of optical interconnects, integrating a fiber cable optical backplane within the rack as well as for the rack-to-rack cluster fabric. Optics modules in the Power 775 are located on the same first level package as router chips, on a glass-ceramic multichip module (MCM). The MCM contains 28 transmit and receive modules with 12x10 Gb/s capabilities. These MCMs are connected electronically to the microprocessor MCMs on the same card and optically to the router MCM. Figure 1.2 show one side of the system card, with eight router chip MCMs and their associated optics [12].

1.1.2 Cloud Computing

Clouding computing is quickly emerging as a resources for many work flows due to its ability to meet the needs of numerous diverse costumers [13]. From search engines to video streaming and cloud computing applications, the cloud is capable

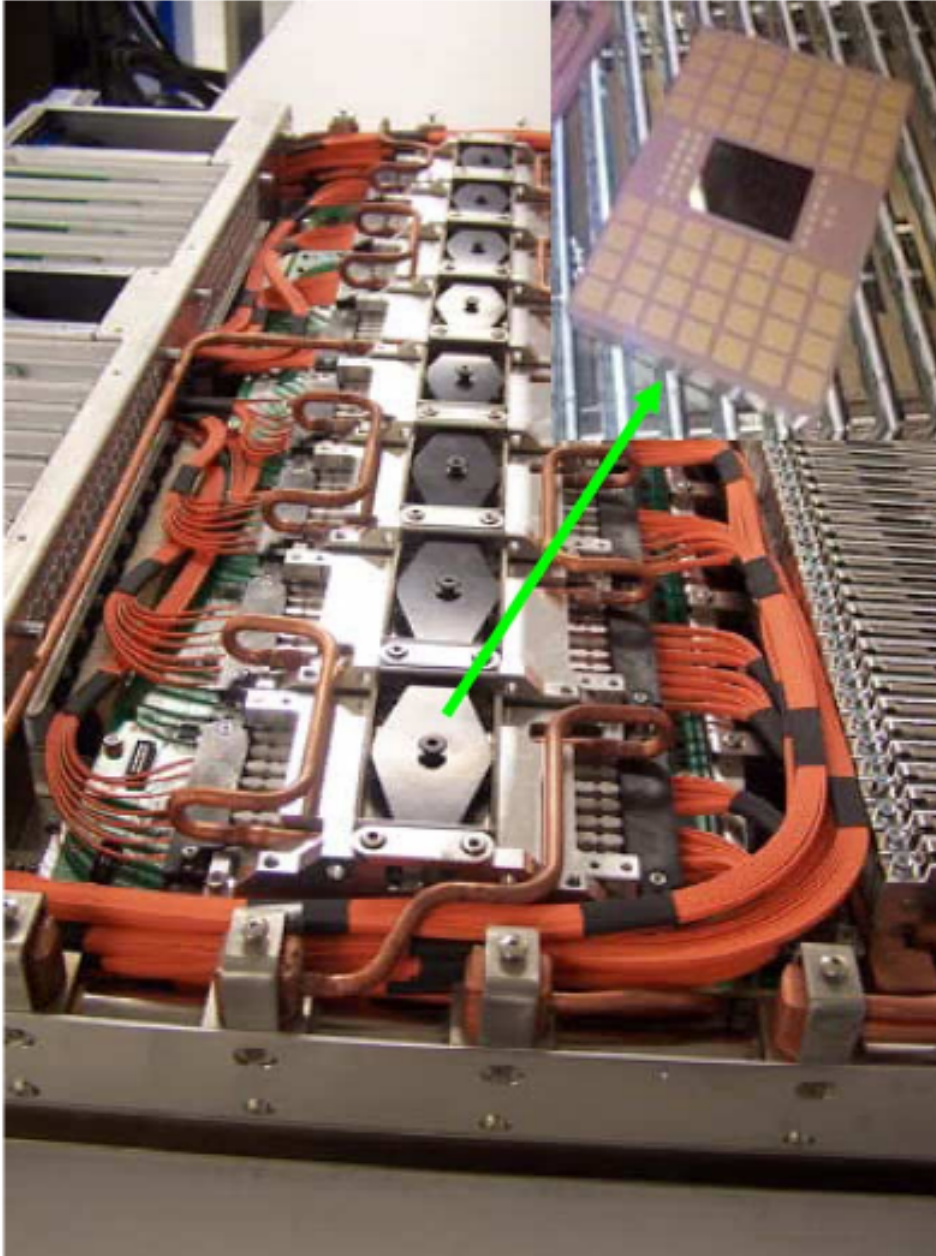


Figure 1.2: Optical Interconnects in IBM Power 775 - An IBM Power 775 system drawer, with eight router multichip modules (MCMs), each with 28 transmit and receive modules, with 12x10 Gb/s bandwidth. The inset shows the underlying glass-ceramic substrate. Source: IBM

of running every data-intensive computations that require a high level of communication [14]. This architectural model has led to a direct shift in recent years to on-demand computing services like Hardware as a Service (HaaS), Data as a Service (DaaS), Software as a Service (SaaS), and Gaming as a Service (GaaS) [1, 8].

HaaS emerged as a result of advances in hardware virtualization, usage meter, pricing, and IT automation, and provides customers with scalable and manageable hardware as needed, examples include Amazon's EC2, IBM's Blue Cloud project, Nimbus, Eucalyptus, and Enomalism [15]. Nvidia's GaaS platform, NVIDIA Grid provides high-quality multi-device gaming with less hassle, and click-to-play simplicity, with no need to purchase new gaming hardware, game patches, or digital downloads needed [1, 16]. In SaaS, applications that are traditionally desktop based, like word processing and spreadsheets, and move them to the cloud, alleviating the customer of the burden of software maintenance, and simplifying testing and development for the provider [17].

According to research from IBM, 85 percent of new software today is being built for the cloud. One-quarter of the world's applications will be available on the cloud by 2016, and almost three-quarters of developers say that they are using the cloud in applications they are developing now [18]. Recently, several cases of cloud computing being used to solve mobile computing issues have been seen.

Current mobile computing applications are demanding compute intensive capabilities like natural language processing, computer vision, augmented reality, and speech recognition. These demands and computations are not being performed in the mobile devices themselves, but rather in the cloud. Amazon's Silk browser is an example of mobile applications that leverage the cloud. Silk is a "cloud accelerated" Web browser, where the software resides both on the Kindle as well as Amazon's EC2 cloud. Silk divided the labor of a page request between the mobil hardware and the Amazon E2, looking at factors like network conditions and location of content. As a whole these applications are known as mobile cloud computing applications (mCloud)[19].

Due to the diverse and data-intensive nature of cloud applications, high interaction is needed between servers. This requirement poses a significant challenge to the networking in data centers, needing to create interconnection networks

with high bandwidth and low latency [14]. Networks in cloud computing systems must improve to sustain increasing network traffic nodes, in addition to keeping the total power consumption inside the rack almost the same, due to thermal constraints [14].

Recent developments in cloud architectures have focused on improving the ratio of performance to cost, mainly by addressing the physical-layer technology within the network. High-radix microelectromechanical systems (MEMs) based optical circuit switches have been proposed for use in data centers in cloud computing [20, 21]. MEMs are attractive due to their energy efficiency and bandwidth density of optics when compared to electronic switches. However, MEMs switches have a relatively high latency, leading to inflexibility in network topologies. Moreover, MEMs based approaches are inadequate for diver and unpredictable traffic in cloud computing applications and thusly, additional physical-layer advances are necessary for next-generation cloud computing systems.

1.1.3 Heterogeneous Utility Computing

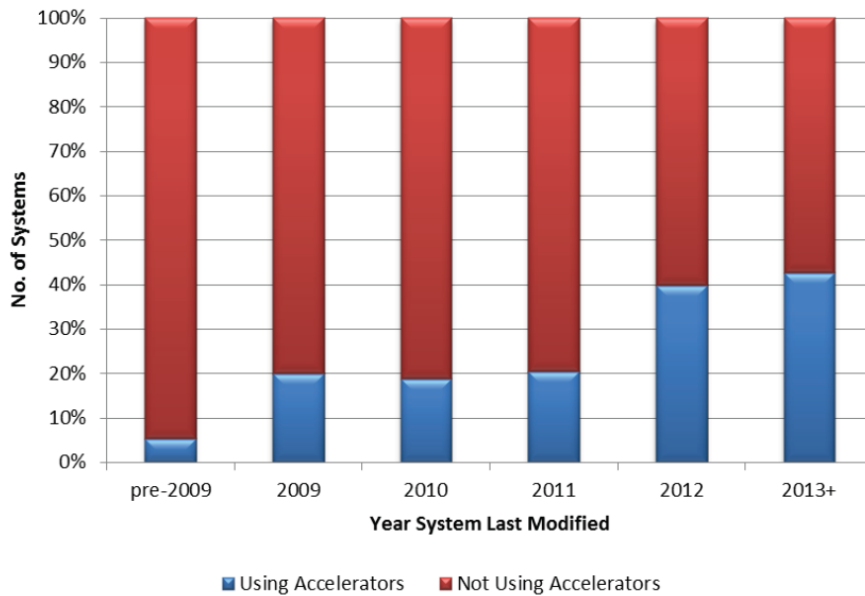
These large systems are quickly becoming heterogeneous in design, and next-generation large-scale computing systems must leverage novel physical-layer technologies to close the network latency gap, enabling Large-Scale Systems to reach their full potential. By leveraging economies of scale and optimizing resource utilization, utility computing is beneficial to both operators as well as the end users, allowing for lower costs, higher efficiencies, greater flexibility and scalability in the utilization of hardware resources [22]. Computational and data centers are often limited by power density, efficiency, and computer density, and while GPP microprocessors are working toward improving power efficiencies, heterogeneous processing (in the form of hardware accelerators) can provide an order of magnitude improvement in these metrics[23].

Beyond the ability to offer massive parallelization, specialized computation hardware can be used to accelerate tasks such as regular expressions evaluation, linear algebra solving, or digital signal processing (all common computations in these utility machines). As a result, the use of hardware acceleration (specialized hardware for specific computational tasks) has emerged as a critical architectural

1.1 Large-Scale Utility Computing

entity [4]. Heterogeneous systems show a lot of promise by combining the benefits of conventional architectures with those of specialized accelerators [24]. Not only are these cloud computing systems able to parallelize the workload of an application across multiple processors, they can also offer specialized hardware to off-load and accelerate programming execution [6].

Following the trend of acceleration adoption in the past several years, this year, more than half of newly installed systems will incorporate accelerators as HPC system operators move from the initial "heat seeker" phase of technological innovation toward the early adopter phase. The large numbers seen in this deployment phase are an indication of the high user expectations for performance gains relative to CPUs [11].



Source: Intersect360 Research, 2014

Figure 1.3: Acceleration in HPC - Percentage of High Performance Computing Systems (surveyed by Intersect360 Research with Accelerators from 2009 to 2013. Between systems installed in 2011 and 2012, there was a doubling of systems with accelerators, in 2013 this increased by another 3%. Source: Intersect360 Research

Graphics Processing Units (GPUs), Field Programmable Gate Arrays (FPGAs), Floating Point Processing (FPP), Regular Expressions hardware, and other forms of hardware acceleration are already being used commercially in industry.

Known as heterogeneous computing, these systems offer users a unique platform for computation that traditional Personal Computers can't offer. Using different kinds of processors - x86, CPUs, GPUs, FPGAs, Encryption engines - in cooperation on a computing task in a pay-per-use systems creates an ecosystem for utility computing operators to invited in specialized hardware that would be inefficient in traditional non-utility computing models. As seen in Figure 1.6 between 2012 and 2013, Intersect360 Research's High Performance Computing (HPC) User Site Census found that the percentage of High Performance Computing Systems with Accelerators increased from 24 percent to 44 percent [2, 25].

Additionally, in HPCs, the number of accelerators per systems is increasing. This increase is indicative of HPC sites moving from a phase of testing the concept of acceleration, but rather putting accelerators into real problems, at times large problem sets. From 2013 to 2014, of the systems surveyed by Intersect360, the average number of accelerators per systems was 121, as compared to 58 in 2013 reports (excluding outlier systems of more than 2,000 compute nodes and single systems with more than 1,000 accelerators per system [25]).

The Department of Defense's Heterogeneous High Performance Computer (HHPC) combines 48 Pentium4 Xeon nodes with a 12 million gate Annapolis Microsystems Wildstar FPGA, using it to accelerate the parsing for Joint Battlespace Infosphere pub-sub brokering from 2 ms to 14 us using the FPGA [26]. The HHPC sustained a rate of 34 trillion operations per second on 48 nodes and one FPGA, comparable to the top HPC at the time, the Earth Simulator in Japan, which had a peak rate of 36 trillion operations per second, on 640 nodes.

1.2 Hardware Acceleration

Hardware Accelerators are compute nodes that are faster at performing specific calculations than general purpose processors. Computation kernels that are often found in cloud computing algorithms, such as pattern matching and digital signal processing, can greatly benefit from hardware acceleration. Utility computing systems offer a unique platform for specialized hardware. Known as hardware acceleration, these units are build not to compute any task, as General Purpose

1.2 Hardware Acceleration

Processing is, but rather to compute specific tasks or to work on specific data sets.

The GPU is an example of such a hardware system. GPUs are specially designed to manipulate and create data sets specific to graphics processing. Resultantly, these hardware accelerators offer large parallel computing capabilities. Many compute tasks and applications with highly parallel data manipulation, like those algorithms found in Monte Carlo simulations in financial analysis and video game graphics rendering in Gaming as a Service (GaaS), are perfectly aligned for hardware acceleration on GPUs. GPUs, Floating Point Processors, and Regular Expressions Hardware are all forms of Application Specific Integrated Circuits (ASIC) that are specially designed for computationally intensive software code. Hardware accelerators are much faster than software, at the expense of taking up more space and specialization to a specific task.

In the financial services industry, as firms push towards executing trades at faster speeds, many are turning to GPUs to get an extra edge [6]. There is a 200-300 percent increase in performance in GPU/CPU systems when compared to a single x86 core. These systems are lower latency, consume less power and deliver higher performance for the same power. As seen in 1.5, Murex has been able to achieve 150x speedup using Nvidia GPUs in when compared to a single core Xenon processor [2].

The NVIDIA Corporation recently announced their GPU-accelerated Gaming as a Service (GaaS) system (Cloud gaming), NVIDIA GRID. GaaS allows for the game to be rendered on a cloud system, and design scheme known as any-device gaming. GRID allows for high-quality, low latency, multi device gaming on any PC, Tablet, smartphone or television. Additionally, anytime accessibility to a library of gaming titles allows the game to be saved on the cloud, so playing and continuing to play games is device agnostic. These systems also eliminate hardware setups, game discs, digital downloads game installations and game patches. This system has been proven to achieve a 30 ms reduction in latency on the NVIDIA GRID platform [1]. In Grid, each server (pictured in figure 1.4) contains 12 GPUs, 20 servers are packed into a GRID rack, equaling the computing power of 700 Xbox 360s.

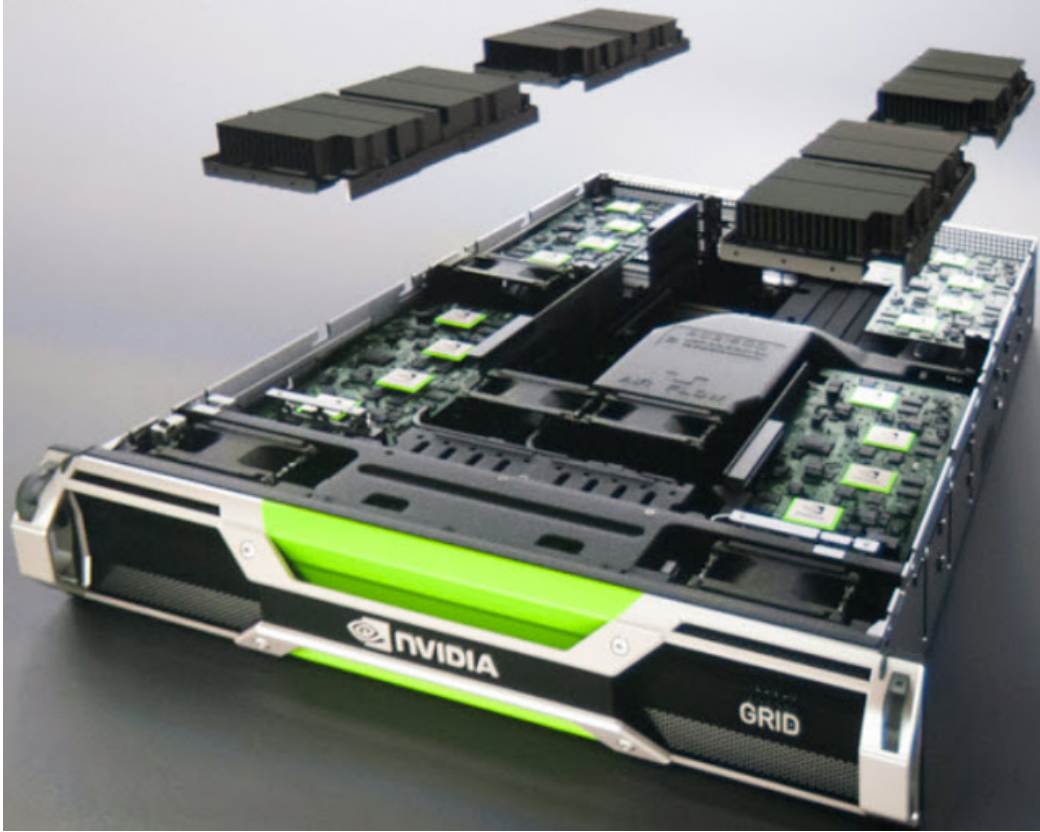


Figure 1.4: Photograph of NVIDIA GRID Server - Each server contains 12 GPUs, 20 of these servers are packed into a single GRID Gaming Rack, which in turn is capable of 200 TFLOPS of computing (equivalent to 700 Xbox 360s) [1].

The Field Programmable Gate Array (FPGA) can also be used in heterogeneous systems to provide direct implementation of an algorithm in hardware. The FPGA provides this ability with custom logic arrays and programmable logic devices, though this comes at the expense of logic speed and density. Though, modern FPGAs provide very-large logic arrays, at reasonable clock speeds [27, 28].

1.2.1 Location

Oftentimes, the location and function of the hardware accelerator is task-dependent. Where you place the accelerator and what you have it do is heavily influenced by the situation. One such example is database acceleration - the SQL domain-by aggregation can be DRAM-limited, cache-bound, TLB-prefetch-bound, or

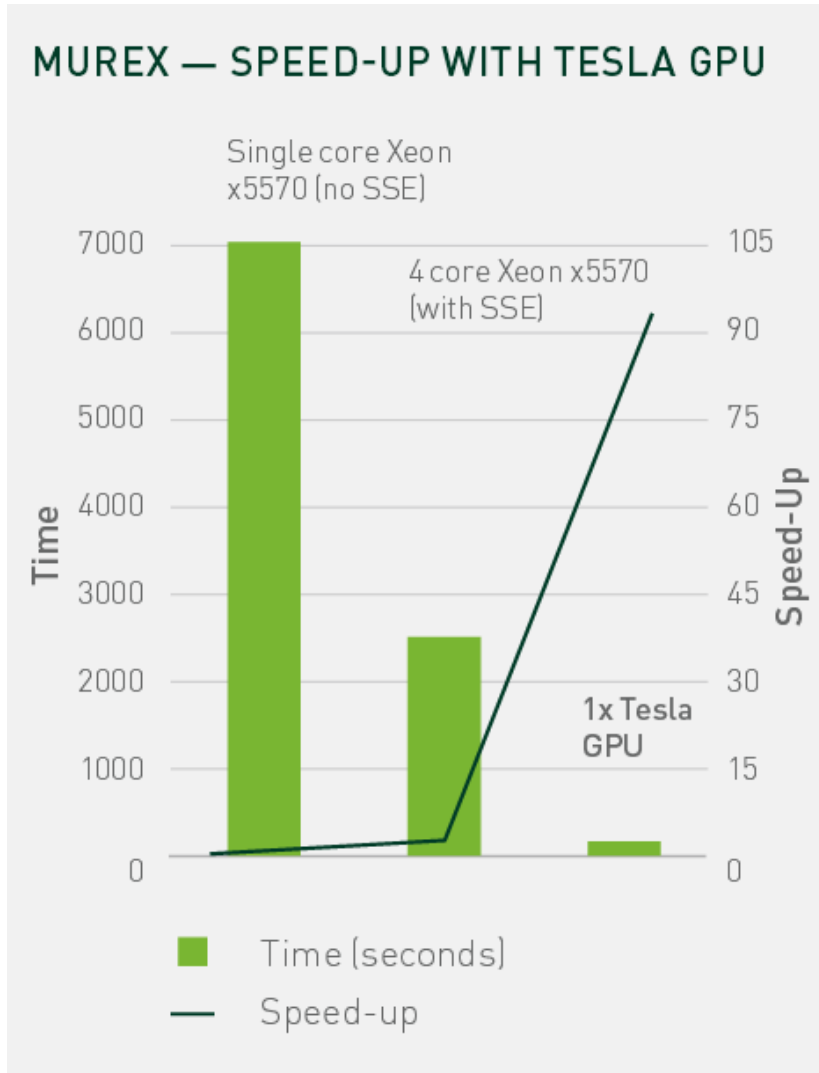


Figure 1.5: GPU-acceleration Speed Up - Time to price a 15 year cancellable range accrual on a Constant Maturity Swap Spread, using a 2-factor Heath-Jarrow-Morton model with 1 million Monte Carlo simulation paths. The model uses a full term structure for volatilities and includes calibration of correlations. [2].

instruction-bound depending on the cardinality, the particular data set, and the operations you are performing [4]. An aggregation operation spanning many disk drives could benefit from accelerators that decompress, formate and select relevant field as a flow-through process, but another type of database operation could require the entire data set to be streamed through caches, and thus need

accelerators to process as CPUs and FPGAs. As data sets and tasks change, one may want to rethink both location and function of accelerators [4].

Communication between the Central Processing Unit (CPU) and these hardware accelerators must be high bandwidth, low latency, and energy efficient [3]. Due to these demands and the power limitations associated with high-speed electronic communications over long distances, accelerators must be placed physically close to the CPU (localized on the motherboard). This architectural limitation severely constrains the number of accelerators each CPU can directly access (only those local to it), and can lead to the under-utilization of these accelerators (cant access more accelerators than those local to it) [3].

1.2.2 Communication

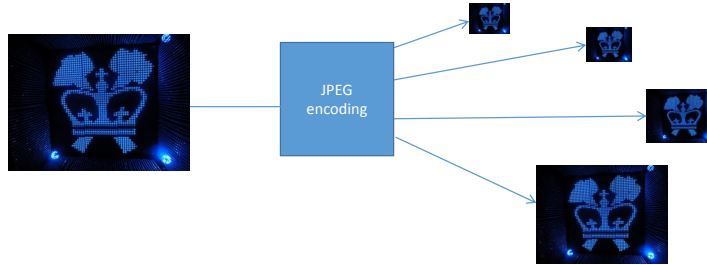


Figure 1.6: JPEGs in Facebook - Illustration of the JPEG encoding in Facebook data centers. Each picture that is uploaded is regenerated as 4 JPEGs of varying sizes, for use in various parts of the sight. Source: Facebook

As these systems grow, the interconnect is becoming a bottleneck that limits the speed of computation [8]. Communication between the Central Processing Unit (CPU) and these hardware accelerators must be high bandwidth, low latency, and energy efficient. The Peripheral Component Interconnect Express (PCIe) [29] protocol used in many GPU platforms has a peak capacity of 2 GB/s per lane (in each direction). Typical GPU systems require 16-32 lanes (32-64 GB/s) to accommodate their bandwidth needs. In order to strive towards zero latency, market data is distributed uncompressed, driving applications toward terabit networking [30].

1.2 Hardware Acceleration

The input/output (I/O) interface (ie. PCIe) of a processor chip is a well suited interface point for architectures are paired with general-purpose processing cores, and allow for standard server models to be augmented with application-specific accelerators. However, traditional I/O attachment protocols introduce significant device driver and operating system software latencies [31]. And, while electronically interconnects can theoretically reach per-channel data rates of 25 Gb/s, the power dissipation at these bandwidths become overwhelming [32]. In NVIDIA GRID, a third of the game latency time is caused by the network in the system [2].

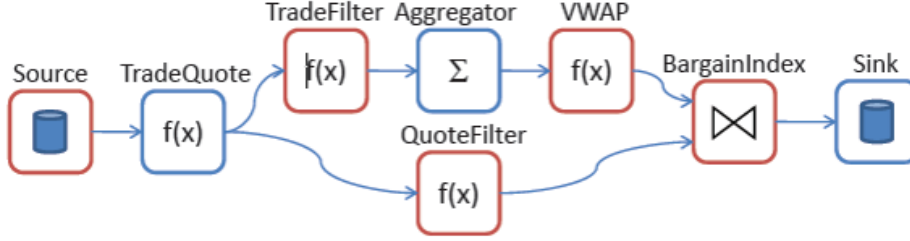


Figure 1.7: Multicast in Spade Algorithm - Multicasting in the Spade Algorithm for bargain discovery. Source: [3]

Additionally, current network models do not exploit the data parallelism found in many utility computing applications. Multicasting data is an integral part of many hardware accelerator architectures. As seen in Figure 1.7, in the SPADE application for bargain discovery for example, Trade Quotes are multicast to a Trade Filter and Quote Filter to help determine if the current asking price for a stock is less than the volume-weighted average price [3]. In Facebook data centers, every uploaded picture is encoded and saved as four jpegs of differing size for use on various parts of the site [33]. Over 220 million new photos are uploaded to Facebook per week, resulting in 25TB of additional data being generated a week in these data centers.

Resultantly, in many of these heterogeneous systems, a large portion of the processing time is now due to networking. As seen in Fig. 1.9, in the NVIDIA GRID system, approximately a fourth of the overall gaming latency time is caused by the networking within the architecture. Due to these demands and the power

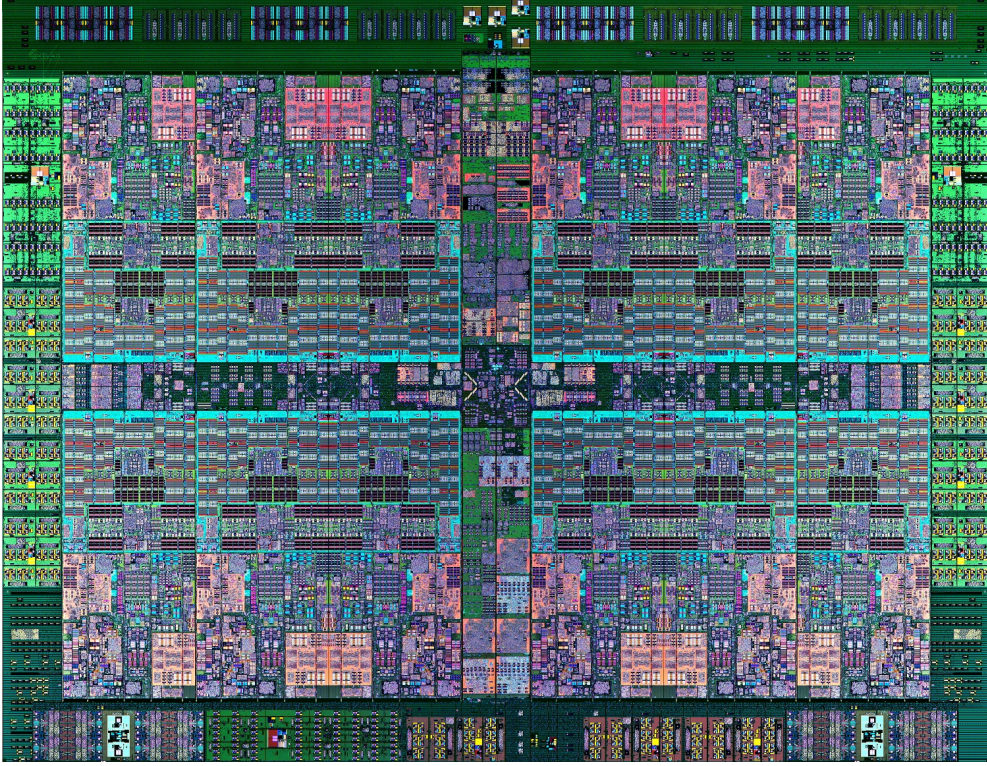


Figure 1.8: Power8 Die Photo - Die photo of the IBM Power8 chip, announced in August of 2013, to be used by the OpenPOWER Foundation for use in big data and cloud computing applications. The PCIe slot on the die can be seen in bright green on the middle bottom of the chip. Source: IBM

limitations associated with high-speed electronic signaling over long distances, the placement of such accelerators have been necessarily constrained to close physical proximities to the CPU.

1.2.3 Programming Innovations

Various programming languages exist for CPUs, GPUs, FPGAs and various other accelerators in isolation. OpenCL [34], CUDA [35], and OpenMP [36] are all languages used in programming GPUs that are extensions of the C programming language [37, 38]. In order to extract the high-performance benefits of these systems a programmer must program in different languages and models. This makes it hard for the programmer to work equally well on all aspects of an

COMPARISON OF GAME LATENCY

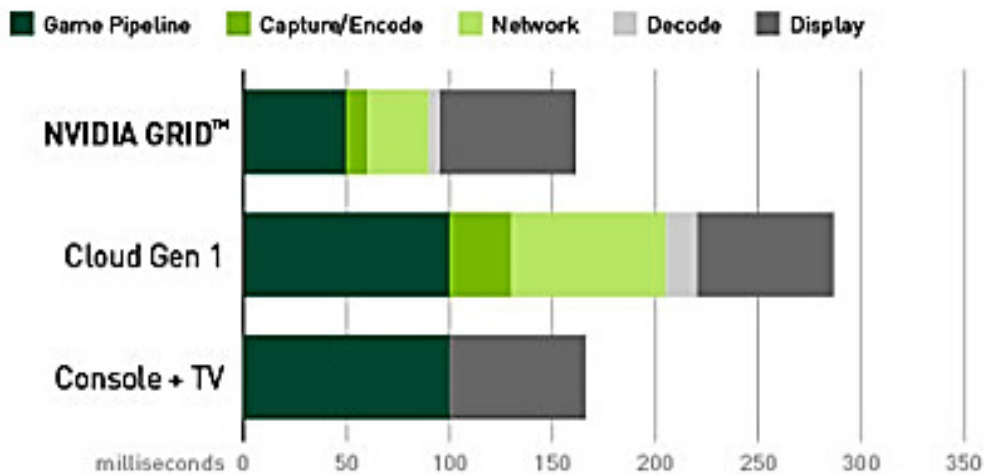


Figure 1.9: Gaming latency in NVIDIA GRID - Comparison of game latency of NVIDIA GRID, Cloud Gen 1, and Console +TV game. The network (light green) takes up approximately a fourth of the overall gaming latency time. Source: NVIDIA Corp

application. Additionally, at current, very little attention is paid to the idea of co-execution - the problem of arranging programming execution using multiple distinct computing elements that work seamlessly together [37].

Co-execution requires that a programmer be able to do a number of things in different languages. Namely, partition a program into tasks that are mapped to specific processing nodes, schedule tasks on these computational elements, and handle the communications between the computational elements, which requires serializing data and preparing it for transmissions, routing data between processing elements and receiving and deserializing data. This complexity is further aggravated by the fact that some accelerators require very specific programming to run efficiently. This places a large burden on programmers and has thus far, in part limited the abilities and full potential of heterogeneous systems [37].

However, recent innovations in this realm seek to improve these limitations. One such compiler and runtime are IBM’s Liquid Metal, a compiler and runtime for a programming language called Lime . Liquid Metal allows for the co-execution of the resulting programming on CPUs and accelerators that include GPUs and FPGAs, and allows for the use of a single programming language for heterogeneous computing platforms [37]. When using Liquid Metal, a program is presented as a representation that describes the computation as independent but interconnected computational nodes. The end result after compilation is a collection of artifacts for different architectures, labeled with the computational node that is implemented.

Consequently, the runtime can then choose a number of functionally-equivalent configurations depending on what nodes are available. With Liquid Metal, one is no longer bound to static, premature partitioning of a problem, and runtime-partitioning is not permanent, is adaptable to program workloads, phase changes, availability of resources, and other dynamic features [37, 38]. Another tool in development is LegUp, an open source high-level synthesis tool that aims to improve the C to Verilog synthesis, making FPGA programming easier for software programmers [39]. LegUp accepts standard C as an input and compiles the program to a hybrid architecture containing an FPGA-based MIPS soft processor and custom hardware accelerators [40].

1.3 The Computing-Optics Interface

The use of photonic technologies holds the potential to enable high-bandwidth links with novel functionalities to reduce off-chip data access latency and power dissipation [41]. Optical interconnects can not only achieve high per-channel data rates, they can also significantly improve communication bandwidths through wavelength-division multiplexing (WDM), and can support terabits-per-second of optical band switching using single optical fiber [42].

While active optical cables have started the shift towards optics in computing, these implementations rely on traditional, inefficient electrical transceivers, providing only moderate energy and performance advancements. Concurrently,

the development of optical switching can significantly improve the overall performance and energy efficiency of utility computing systems [43, 44, 45, 46].

By combining these distinct technologies of optical networks and computing, maximal benefits can be achieved in future optically-enabled utility computing systems. It is important to note that electronic technology in current microprocessors has been optimized for short-distance, bursty communication, thereby relying heavily on point-to-point links in which data is frequently buffered and retransmitted. Contrastingly, a lack of optical buffering technology has resulted in optical network technology development that utilize unique communication protocols. As a result, there is no directly mapping of existing electronic systems to optical networking technologies. Therefore, it is necessarily to develop a computing-optics interface that will be able to enable future processors to both leverage the benefits of optical interconnects, while simultaneously minimizing significant modifications to the surrounding processing technologies.

1.3.1 Optical Interconnection Networks

Optical Interconnection Networks (OINs) are an attractive solution to the communication bottleneck (See: 1.2.1) within future large-scale computing systems [44, 46, 47, 48]. Currently microelectromechanical system (MEMs) switches are being considered for integration into data center architectures. However, due to the inherent mechanical nature of MEMs based switches, and consequently their high switching latency, these switches are unsuitable for most networks.

Semiconductor Optical Amplifiers (SOAs) on the other hand, have been demonstrated to provide high-bandwidth, low-latency switching for optical switches [45, 49]. Silicon photonic devices show great promise in their ability to provide high-bandwidth, low-latency, and energy-efficient switches, and could then be used to create large-scale optical networks [50]. However, silicon photonic technologies are still in its infancy and due to this immature state are not yet suitable for creating large-scale networks. As a result, SOA-based OINs serve as the basis for this dissertation. It should be noted that the developed protocols in this work remain compatible with both SOA-based and silicon photonic switches.

1.3.2 Optically Connected Hardware Accelerators

Heterogeneous Utility computing architectures are well suited for the deployment of optical interconnects, especially optical networks, due to performance and energy requirement of hardware accelerated computing, as well as the necessary flexibility needed in the network to support needed to support the inherent data parallelism in traffic patterns, and the application-dependent location and function of hardware accelerators. By leveraging the bandwidth-density and distance-immunity of optics, optically connected hardware accelerators can alleviate the electronic-interconnect constraints facing current heterogeneous systems. Latency in optical links is purely a function of distance, therefore allowing efficient, transparent optical networks low access latency [51]. With these OINs, delocalized, dynamically allocated hardware acceleration can be realized.

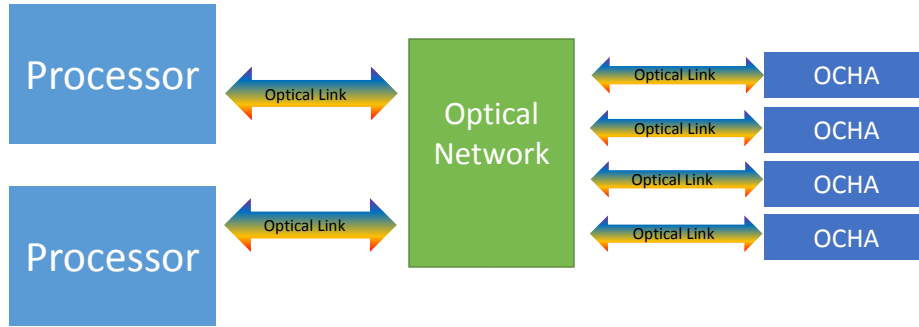


Figure 1.10: Block Diagram of Optically Connected Hardware Accelerators - Block level schematic depicting how a next-generation heterogeneous system can be connected to many optically-connected hardware accelerators across an optical interconnection network.

There has been continuing work into alleviating the network latency in heterogeneous systems. In [31] the authors present CAPI, a Coherent Accelerator Processor Interface that attaches the accelerator as a coherent CPU peer over the I/O physical interface. Designed for the POWER8 platform, the CAPI provides the capability for off-chip accelerators to be plugged into PCIe slots, participating in the system memory coherence protocols and enabling the use of effective addresses to reference data structures in the same manner as applications running

on the cores. This bypasses the costly driver and software I/O stacks used in most systems impose a high overhead and a cumbersome communication model, decreasing the speedup with heterogeneous systems could have.

Ongoing work into 3D heterogeneous computing integration at the chip level shows promise. In [52], a scalable heterogeneous multi-core processor is presented as 3D heterogeneous chip stacking of a CPU and reconfigurable multi-core accelerators in massive parallel computing. In this chip, the interconnect is a scalable 3D Network on Chip (NoC). In this work, a change in the number of stacked accelerator chips can scale processor parallelism through an inductive-coupling ThruChip Interface. This chip was fabricated at 65nm CMOS.

1.4 Scope

The primary contribution of this dissertation is the development and implementation of an OCHA system. In this system, the electronic bus between processor and accelerators is replaced by an optical interconnection network, thereby allowing the delocalization of hardware accelerator resources. To achieve this, processors and accelerators interface with local photonic transceivers. Moreover, this work encompasses the creation of a novel optical-network-aware hardware accelerator allocator that functions as the optic-computing interface, as well as a novel switching protocol for a wavelength-stripped phase encoded header. A series of experiments is presented that characterize the OCHA systems across three key metrics that must be addressed in next generation heterogeneous systems.

- **Bandwidth** - The low bandwidth-density of electrical interconnects [53] limit hardware accelerator bandwidth. OCHA must allow a road map towards larger, delocalized heterogeneous system architectures.
- **Latency** - The network latency in heterogeneous systems is a large part of the overall system latency. Optical networks must be designed to minimized additional latency when compared to a traditional electronic link.
- **Efficiency** - In many systems, the efficiency of the systems is directly correlated to profit margins [4]. Depending on the industry and the application,

efficiency can be defined differently. The speedups provided for optical interconnected heterogeneous systems must compare in efficiency to electronic systems.

This dissertation is organized as follows.

- **Chapter 2** - This chapter details the optical network architecture utilized throughout this dissertation. Optical networks are paramount to enabling technology for future optically connected hardware accelerators (OCHA) in utility computing systems. Here, the protocols and network functionalities are analyzed with respect to their abilities and impact on future utility computing systems.
- **Chapter 3** - This chapter details initial work done in dynamic optical interconnection networks. Next-generation mobile computing cloud applications will need to be user aware and be able to dynamically route data and allocate resources as a function of location and network conditions of the mobile client. In this chapter, a Wimax/Optical network testbed is presented, utilizing the a Optical Interconnection Network Interface in conjunction with a WiMax antenna and a VLAN connection to a transparent WDM optical network.
- **Chapter 4** - This chapter details initial work done to explore the multicasting abilities of OCHA systems in the areas of bandwidth and latency. The OCRM used in this work and the work presented in Chapter 5 is outlined and detailed in this chapter. This is the first demonstration of a optical test bed that is able to leverage the phase encoded header as a fast-switching control of large hardware accelerator packets.
- **Chapter 5** - In the work presented here, the OCHA network is demonstrated as a bidirectional dynamically reconfigurable functional system. This critical step demonstrates that the flexibility provided by optical interconnects provides improved bandwidth, latency, and efficiency. In this experiment we validate our proposed architecture with a FPGA-based bidirectional emulation test bed. The optical packets generated by the FPGA

are sent through a semiconductor optical amplifier (SOA)-based, wavelength stripped, optical network, and utilizes a phase-encoded header for routing. Additionally, it demonstrates the ability of optical interconnects to provide novel architectures in heterogeneous computing.

- **Chapter 6** - This chapter summarizes the contributions of this dissertation and describes ongoing and future work towards developing OCHAs in heterogeneous utility computing systems.

Chapter 2

Optical Interconnection Networks for Heterogeneous Utility Computing

This chapter details the optical network architecture utilized throughout this dissertation. Optical networks are paramount to enabling technology for future, dynamically allocated, reconfigurable, and optically connected hardware accelerators (OCHA) in utility computing systems. Here, the protocols and network functionalities are analyzed with respect to their abilities and impact on future utility computing systems.

2.1 Optical Network Design

The optical networks used throughout this dissertation makes use of SOA-based optical switching nodes [54] to implement 2×2 or 4×4 switching fabric test beds. The modular switching nodes can be linked together to create larger, multi-staged networks. Utilizing a wavelength-stripped format for messages, this these switching node minimize latency by eliminating many inefficiencies in current switching technologies, namely those caused by Optical-Electronic-Optical conversions.

Figure 2.1 shows the wavelength-stripped format that enables messages to be transparently routed through the switching node. By using wavelength-stripped routing, routing is simplified and switching latency is minimized. In this scheme,

routing information is encoded on header wavelengths that are combined with the hardware accelerator data using WDM, as seen in Figure 2.2. Each header utilizes a phase-encoded header (see section 4.4.1 or section 5.3.2 for control logic). This system allows for dynamic switching to each destination node independently.

The switching node utilizes a broadcast-and-select architecture, the wavelength-striped message enters the input port, where passive optical elements(eg. couplers, filters, fiber) direct the appropriate header wavelength to low-speed (155-Mb/s)photo detectors (PD) on the switch boards on the 30 side of the splitter. The now-electrical header is then passed to simple, high-speed control logic that in turn gates an SOA on or off. The SOAs act as a broadband optical amplifier at each output port; the gain provided by the SOA restores power equivalent to any optical losses incurred throughout the switching node.

Concurrently with this process, the payload data packet (70 side of splitter) passes through a fiber delay line (FDL) that matches the time required to filter, receive, and process the aforementioned header information (approximately 10 ns). The FDL routes the payload data into the SOAs, and should the logic indicate that the node is to be switched to, the SOA is enabled just-in-time to allow transparent, low-latency routing. The control logic is implemented using a complex programmable logic device (CPLD) [55] or an FPGA, depending on the board used. The FPGA allows for more diverse and advanced network functionality. Figure 2.3 shows an illustration of how the switching node could be configured in a 3×3 network.

2.1.1 Switching Conventions

The photonic switching nodes outlined in 2.1 have been demonstrated to operate as a packet switch [49], circuit switch [56], or hybrid packet and circuit switches[57]. Wavelength-striped optical packet switching is achieved by modulating the header wavelength such that they are consistent for the duration of each packet, with some guard time at the beginning and end of the packet. In this scheme, optical packets can range from tens of nanoseconds to milliseconds in length, with a payload ranging from 10 Gb/s to 8×40 Gb/s WDM channels per packet [58]. Table 2.1 indicates the combinational logic used in the CPLD to

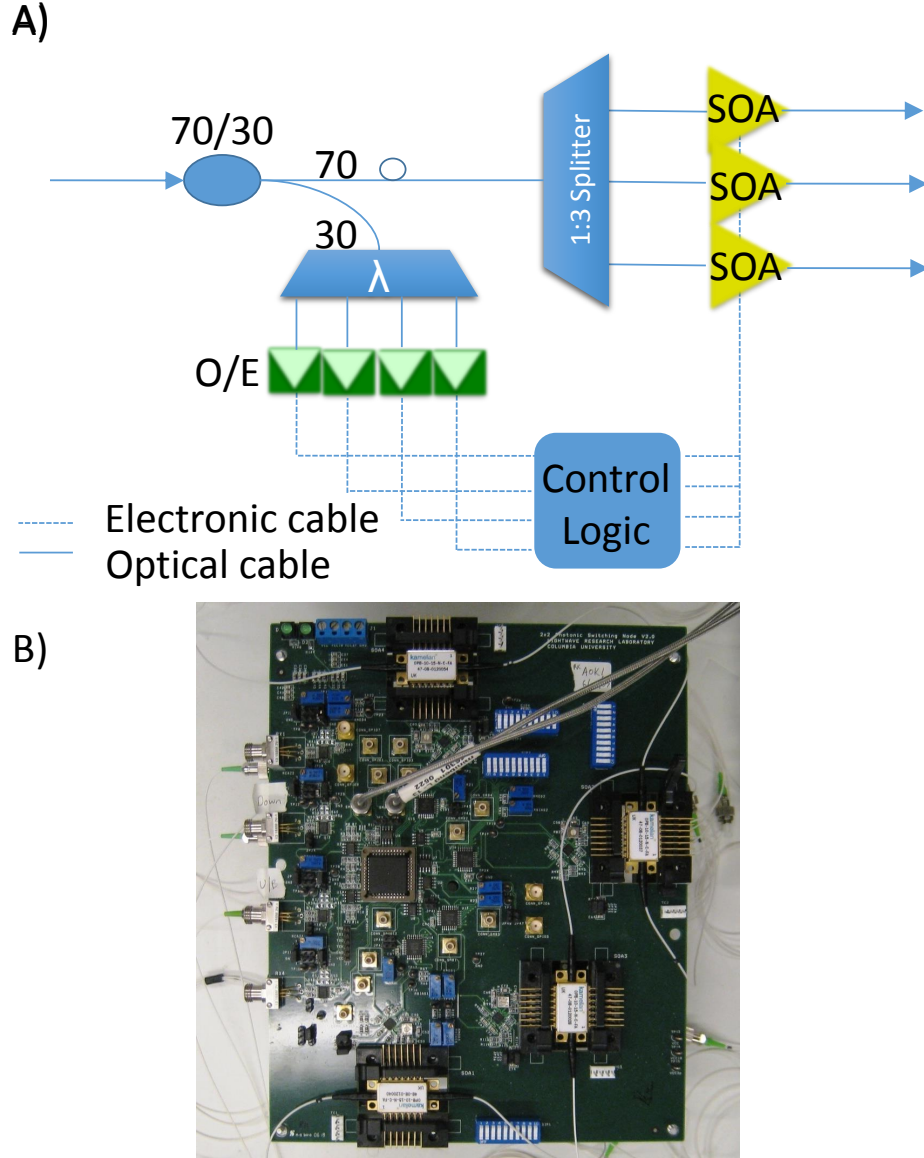


Figure 2.1: Photonic Switching Node- (A) Schematic and (B) photograph of the SOA-based switching node - routing information is encoded on the header wavelengths that are decoded through an OEO conversion, where a CPLD computes the logic to control a number of SOAs. This scheme allows for the elimination of many OEO conversions that limit current switching nodes.

control the SOAs. This protocol was modified in later work to create the phase

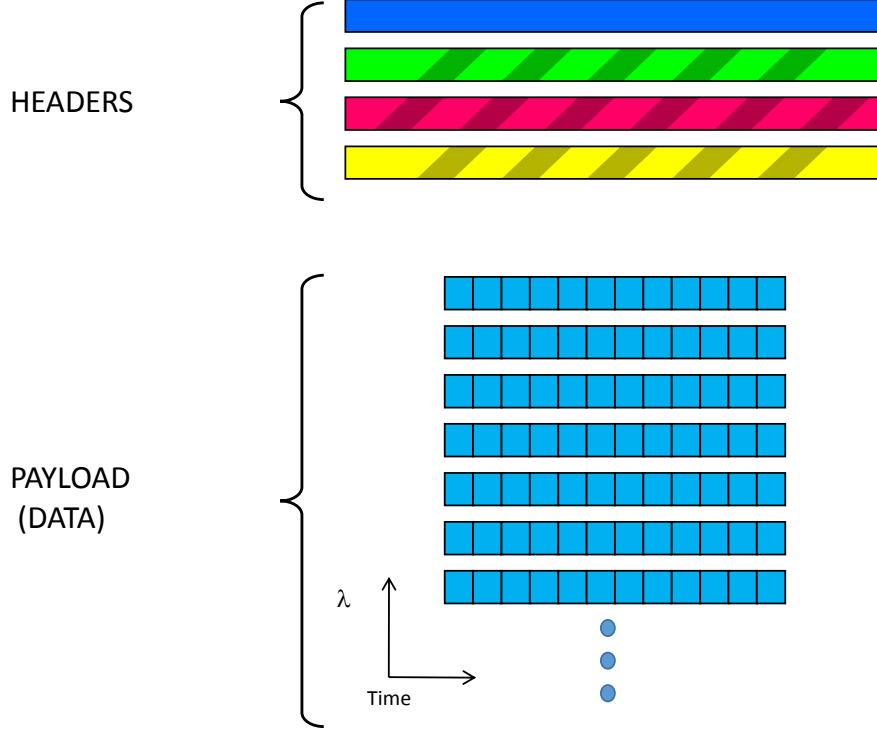


Figure 2.2: Wavelength-Striped Message Format - Low-speed header switching wavelengths (controlled by the GPIO MICTOR on the FPGA) are combined with high-speed payload wavelengths using WDM

encoded header schemes detailed in Sections 4.4.1 and 5.3.2. An attractive inherent feature of optical interconnection networks, as illustrated in Section 2.1, is the inherent ability to perform a multicast in the optical domain using passive devices that can split the power of an input signal into several outputs [59].

OCHA systems may exhibit unpredictable communication patterns and benefit greatly from quick switching abilities. Communication patterns also involve messages of varying lengths. It is therefore desirable to make runtime decisions regarding resource allocation and switching implementations.

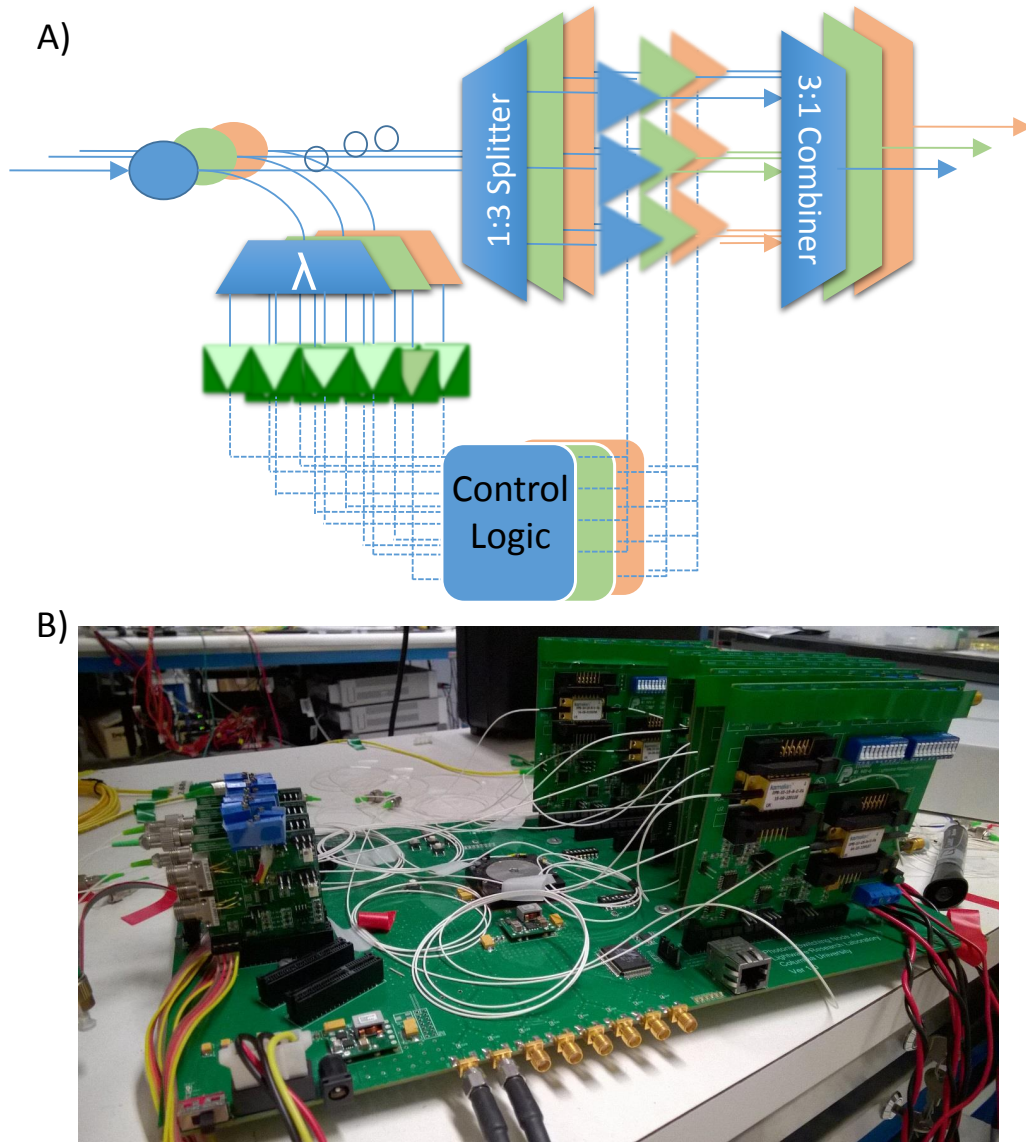


Figure 2.3: 3×3 switching - (A) Schematic of how the switching node would be configured in a 3×3 set up, with an FPGA as the control logic and (B) photograph of the 4×4 SOA-based switching node that this could be implemented on

C	B	A	Frame	Output
X	X	X	0	No Packet
0	0	0	1	Valid Packet, ALL off
0	0	1	1	Switch to Node A
0	1	0	1	Switch to Node B
0	1	1	1	Switch to Node A and B
1	0	0	1	Switch to Node C
1	0	1	1	Switch to Node A and C
1	1	0	1	Switch to Node B and C
1	1	1	1	Switch to Node A, B, and C

Table 2.1: Wavelength-striped header logic table - When the frame wavelength is valid, the output SOAs are controlled by their control bit. In the above table, these bits are labeled A and B. The logic explains what nodes would be switched to in each scenario. This combinational logic implementation allows for simplified logic in the CPLD. In later work, this scheme was modified to allow for fast switching of longer packet.

2.1.2 Scalability

In order for optical interconnection network to enable future computing systems to achieve greater performance and scalability, high data rates and advanced modulation formats must be adapted. Though the traditional non-return-to-zero on-off keying (NRZ-OOK) is popular for its simplicity, differential-binary-phase-shift keying (DPSK) has been seen as a potential improvement, due to its 3-dB improved receiver sensitivity (with balanced detection) as compared to OOK. These benefits become more apparent as data rates exceed 40 Gb/s and even 100 Gb/s, and the optical network elements begin to exceed the abilities of driver and receiver electronic circuitry.

Next-generation large-scale systems will require optical interconnection networks that utilize high per-channel data rates and advanced modulation formats that improved resilience and spectral efficiency. The photonic switching nodes outline in 2.1 have been demonstrated to transmit 8x40Gb/s WDM packet across

a 4x4 optical interconnection network test bed using OOK and DPSK [60]. These experiments correctly routed 8x40 Gb/s data with error free operation (BER 10⁻¹²) with best case power penalties of 1 dB for OOK and .52dB for DPSK streams.

2.2 Implications for Heterogeneous Computing Systems

The latency characteristics of an optical interconnect approach to hardware acceleration is a critical issue to exam. Current networking latencies in heterogeneous systems are already growing, additional latencies are not desirable. The transparency of the photonic switching nodes reduces the overall latency to the time-of-flight between a processor and accelerator. Each additional meter of single-mode fiber (SMF) adds approximately 5 ns [51] of latency to the communications path.

And, while this may be negligible at a single rack, it may become problematic for links that span large-scale computing systems. A goal of this OCHA design is the minimize the accesses to accelerators that are more than a few meters away (similar to the case of today's electronic networks). One of the main advantages of the OCHA systems is that it's sole limitation is the latency caused by time-of-flight. On the other hand, in electronic interconnects, communications distances are limited by latency, power, bandwidth, and real estate (space on die/board). In this dissertation, local hardware accelerators are the ones with the shortest optical path.

The OIN described here enables the latency and bandwidth performance of the OCHA system to meet the demands of heavily loaded heterogeneous utility computing systems with hundreds to thousands of processors and accelerators. Moreover, high-speed transceivers operate at high per-channel data rates, demonstrated by 10 Gb/s, 25 Gb/s and the recently amended 40 and 100 Gb/s Ethernet standards [61], and recently announced 400 Gb/s Ethernet [62]. Using multiple of these high-speed transceivers with WDM creates the bandwidth density necessary for OCHA nodes with bandwidth in the 100's or 1000's of gigabits per seconds, on a single fiber. Many OCHA nodes could then be combined further using an optical interconnection network with petabit system bandwidths.

2.2 Implications for Heterogeneous Computing Systems

The efficiency of optical interconnects supports diverse applications and effects both compute and power efficiencies. The reconfigurability of the proposed OCHA network supports many workloads and computational situations. Take for example an application of cloud gaming, in which a user begins playing a game on a small mobile platform with low connectivity on her way home, the application could allocate her game stream to a GPU accelerator for graphics and physics processing.

If sometime later the same cloud game is restarted on a Fibre-to-the-home connected 4K display medium, the game could be reconfigured to 8 GPU accelerators for processing. The system can be configured to allocate accelerators based on availability and application specifics, including but not limited to connectivity, calculation size, etc. Additionally, a web search application with less predictable communications could offload processing to the cloud when a connection is strong, but do more localize processing when the user jumps to a less strong connection.

The application specific functionality and location can also be address with reconfiguration in OCHAs. An ASIC like a FPP could be configured in the optical network to appear in different parts of a pipeline chain by reorganizing the optical network. With a combination of FPGAs and optical networks, one could imagine a system wherein the location and function of an accelerator was a runtime, or close to runtime, allocation. The flexibility and computational possibilities of this platform are enormous, opening heterogeneous computing systems a new applications that were previously unattainable.

The network nodes would be configured to exploit the data parallelism in many heterogeneous utility computing processes. In many applications, where a multicast of the same data is needed, as mentioned in Section 1.2.2, a processor can simultaneously multicast data to multiple accelerators to perform different calculations on with a single accelerator access. Alternatively, the multicast-capable OCHA system could be configures for resilience to tolerate a hardware failure of an entire accelerator, or even accelerator node, while maintaining efficiency, the use of multicasting can also be used to transmit along serval redundant paths to the same destination node to ensure coherence.

2.3 Discussion

Optical interconnection networks are critical components of future optically-connect heterogeneous systems that allow high-performance, energy-efficient integrated optical links to leverage low-latency, transparent routing through the large scale networks in next-generation heterogeneous utility computing systems. This in turn, enables the deployment of OCHA systems with the necessary computing capacity and bandwidth to mitigate the increasing network latency issues.

The optical network test bed components described in this chapter can provide low-latency optical network between processors and hardware accelerators. For the remainder of this dissertation, the above optical network is used. Additional comments that were utilized are described in subsequent sections.

Chapter 3

Dynamic Data on Optical Interconnection Networks

This chapter details initial work done in dynamic optical interconnection networks, and preliminary work involving the wavelength-striped header packet format. Next-generation mobile computing cloud applications will need to be user aware and be able to dynamically route data and allocate resources as a function of location and network conditions of the mobile client. In this chapter, a Wimax/Optical network testbed is presented, utilizing the a Optical Interconnection Network Interface in conjunction with a WiMax antenna and a VLAN connection to a transparent WDM optical network.

This test bed dynamically changed the destination IP address of the packets by looking at the received signal strength of the mobile user. In this work a system was demonstrated that dynamically streams video data from a wireless client through the WiMAX base station. The packets are processed in real time by a netserv module, sent through a transparent WDM optical network, and received by an end node that receives the pack and decodes the data into a video.

3.1 Background

Current mobile computing applications are demanding compute intensive capabilities like natural language processing, computer vision, augmented reality, and

speech recognition [19]. These demands and computations are not being performed in the mobile devices themselves, but rather in the cloud. According to research from IBM, 85 percent of new software today is being built for the cloud. One-quarter of the world's applications will be available on the cloud by 2016, and almost three-quarters of developers say that they are using the cloud in applications they are developing now [18].

Recently, several cases of cloud computing being used to solve mobile computing issues have been seen. Amazon's Silk browser is an example of a mobile application that leverages the cloud. Silk is a "cloud accelerated" Web browser, where the software resides both on the Kindle (wireless device) as well as Amazon's EC2 cloud. Silk divides the labor of a page request between the mobile hardware and the Amazon E2 cloud computer, looking at factors like network conditions and location of content to make decision on how to divide the workload. As a whole these applications are known as mobile cloud computing applications (mCloud)[19].

As these trends in mCloud continue, the cloud network will need to become application and user aware. One way in which this can be achieved would be to route data as a function of the end user's signal strength. A simple way to measure the location of a user is to look at the Received Signal Strength Indication (RSSI) of a the received mobile radio signal. The higher the RSSI number, the stronger the user's signal, the lower the RSSI number, the weaker, and thus further away.

In this dynamic network, with applications in mobile gaming and augmented reality, the RSSI value would be a factor in hardware accelerator allocation. As illustrated in Figure 3.1, take for example, a system of GaaS, where a mobile gamer with a weak RSSI (and thus poor video quality) was allocated a single GPU for physics engine computations in their game, while another gamer, with a must stronger RSSI (and thus better video quality) was allocated two or three GPUs for physics engine computations. This would maximize resource allocation in the cloud gaming system.

3.2 Optical Interconnection Network Interface

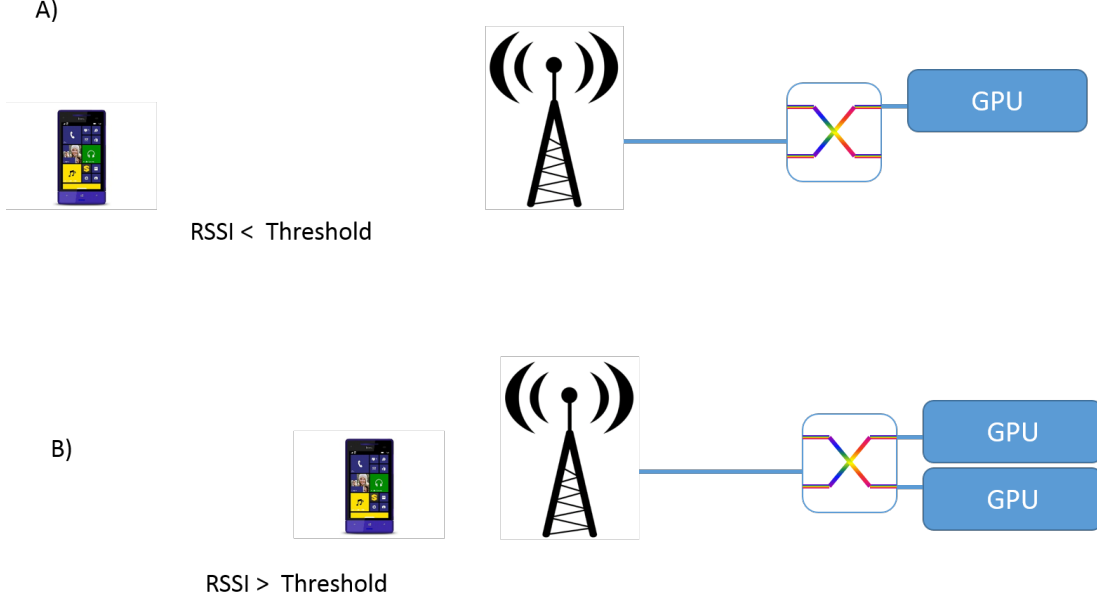


Figure 3.1: Vision - (A) With an RSSI below the threshold value, the gamer is allocated one GPU for a GaaS application, but as she nears the base station and the RSSI increase (B) her game is dynamically streamed to two GPUs now that she has a RSSI above a certain threshold

3.2 Optical Interconnection Network Interface

In order for heterogeneous systems to truly capitalize on the benefits of an optical network, an efficient interface needs to support transparent interconnects between electronic processors and the OIN. This interface must be high-bandwidth and low-latency as well as represent itself as a switch that is compatible with standard network protocols[63].

The Optical Interconnection Network Interface serves as a bridge between the electronic protocols and the optical network protocols. The Network Interface Card (NIC) is a specialized hardware component that connects to a CPU and a OIN and provides transparent optical network communications, in that the optical network is unseen to the electronic end nodes.

Ethernet (IEEE 802.3) [64], (abbreviated as 10GE for 10 Gigabit Ethernet) originally a class of Local Area Network (LAN), has evolved into a large scale computing networking standard. Given it's prevalence in large scale computing,

3.2 Optical Interconnection Network Interface

as well as its compatibility with TCP/IP, the Optical Interconnection Network Interface contains a 10 GE NIC in an end host, a FPGA-based development board, connect through a Quad Small Form-factor Pluggable (QSFP) cables that set up a 4x3.125 Gb/s transparent connection. The Network interface takes Transport Control Protocol data from a CPU running Linux. The data packet is re-encoded into the wavelength-stripped packet outlined in section 2.1.

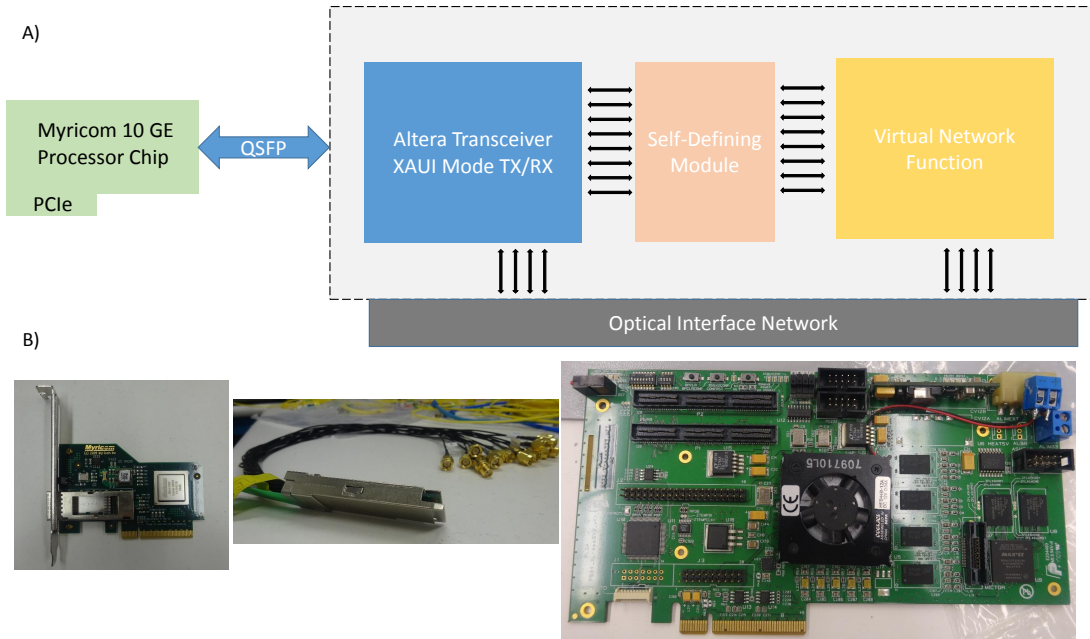


Figure 3.2: Optical Interconnection Network Interface - (A) Schematic of the Optical Interconnection Network Interface showing the Myricom 10 GE card, QSFP, top level logic of on the Stratix II FPGA and (B) photographs of the components used to build the set up, including the Stratix II Development board

The Ethernet link originates in the Network Interconnection Card of a 64-bit host computer, through a 10 Gb/s Myri-NIC interface. The Optical Network Interface Card (ONIC) is implemented on an Altera Stratix II GX FPGA [65] development board. Using a 10 Gigabit Media Independent Interface (between the MAC and PHY layers of 10 GbE -XAUI Mode) transceiver where data is de-serialized, aligned, 8b/10b decoded, and passed to self-defining modules.

The self-defining modules parse the Ethernet header information, transfer clock domains and buffer the data packets. The parsed information is delivered

to a virtual network function module and control of the optical switch is generated through optical headers. Align/Sync/Idle sequences (K28.0, K28.3, and K28.5 in 8b/10b encoding control signals), are sent during channel idle time.

3.3 Experimental Set Up

In this experimental demonstration, Worldwide Interoperability for Microwave Access (WiMax) wireless video packets were generated on a mobile device. As the mobile user maneuvered around, the User Datagram Protocol (UDP) packets were sent to the WiMax base station from the Client using the VideoLAN Client (VLC)[66], a open source software multimedia player product developed by the VideoLAN Project. The packets reached the base station and were processed by a NetServ [67] application module. This module polled for the downstream RSSI value (or the distance of the client from the base station) periodically. Based on this value, the module would then modify the destination IP address of the packet and send the packet on a VLAN through an Optical Network Interface Card to a transparent WDM optical network.

3.3.1 WiMax Data Generation

WiMax refers to interoperable implementations of the IEEE 802.16 [68] family of wireless-network standards. A WiMax-card enabled laptop computer, using the VLC video client was used to live stream video data. The mobile node was allowed to wonder within the range of WiMax antenna and base station, thus varying the RSSI value of the video packets. Video captured by the camera of the mobile node was encoded and immediately streamed though a WiMax card. Video packets that were wirelessly streamed to a WiMax base station.

As seen in Figure 3.4 the software stack consisted of a Netserv module running on top of a Linux Kernal. The Netserv networking module [67]is a node architecture for deploying in-network services in the next generation Internet. Netserv allows for network nodes to implement network services as modules. In this scheme, UDP packets are streamed from the client using VLC. The packets reach the base station and are processed by the NetServ application module. The

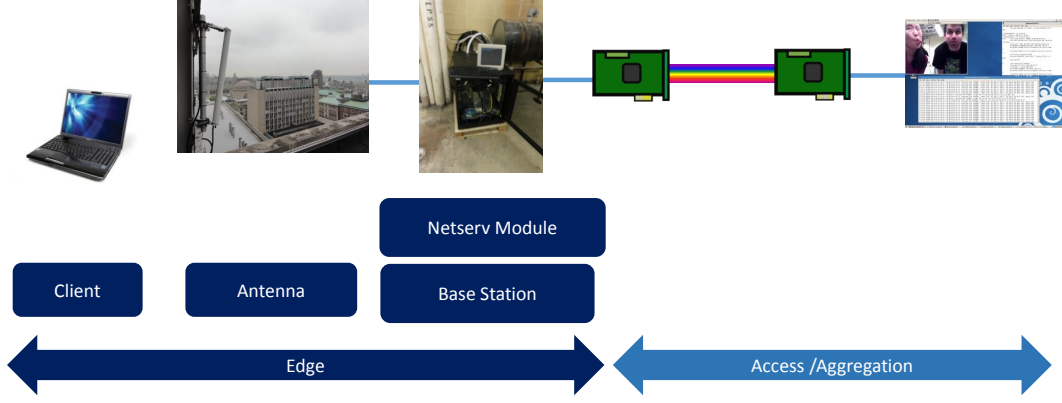


Figure 3.3: Block Diagram of Optical-WiMax Test Bed - The architecture setup. Video from the client is dynamically streamed through the WiMAX base-station, transmitted over a VLAN and through an O-NIC and WDM encoded on the optical network, and then decoded at the end node. This process is transparent to the end users.

module polls for the downstream RSSI value (distance of client from the base station) periodically and modifies the packet changing the destination IP based on the RSSI value, actively changing the IP destination as the RSSI value changes. From the base station, the packets were sent of a Virtual Local Area Network (VLAN) to a computer host.

3.3.2 VLAN and ONIC

A VLAN is set up by technology services between the base station and host computer. The host computer acts as a node in a cloud and relays the incoming packets to the transparent optical network. Attached to this host computer is a Myricom 10 GE card connected to the ONIC through a Quad Small Form-factor Pluggable (QSFP). In the ONIC, the electronic signal was repackaged into a optical network protocol package (as detailed in Section 2.1.1. With the IP address being interpreted as an address signal and the valid packet being identified by a framing wavelength. In this scheme, when the RSSI value is above

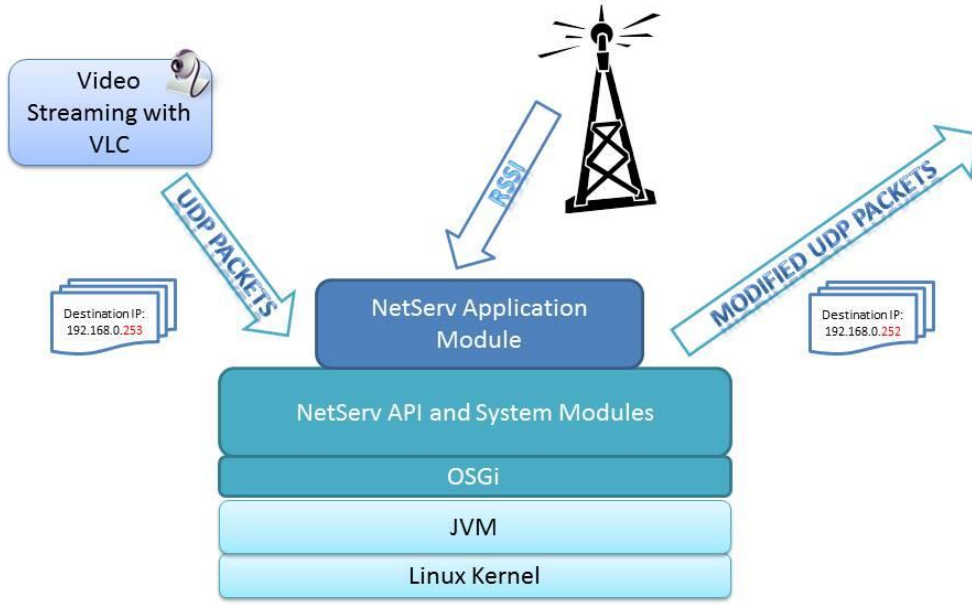


Figure 3.4: Software Stack - UDP packets are streamed from the Client using VLC. The packets reach the base station and are processed by the NetServ application module. The module polls for the downstream RSSI value (distance of client from the base station) periodically and modifies the packet changing the destination IP based on the RSSI value.

a predefined threshold, and the packet is being sent, the framing wavelength is on, and the IP address is changed to that of node A, which is interpreted by the optical network as a 1.

On the contrary, if the RSSI value is below a predefined threshold, the IP address is changed to node B, and interpreted by the optical network as a 0. Table 3.1 illustrates the logic of this scheme. In this architecture, the number of wavelengths needed would be $N+1$, wherein N is the number of nodes, a nominal cost to the system. The wavelengths are modulated by Lithium Niobate modulators (LiNbO₃) to be translated into the optical domain. These routing wavelengths

3.3 Experimental Set Up

are combined with the WDM data/payload wavelengths produced by the ONIC through a passive optical multiplexer. The WDM packets use Ch36-Ch39 ITU [69] spacing with the Frame and Address using Ch27 and Ch53. Figure 3.5 illustrates the test bed set up.

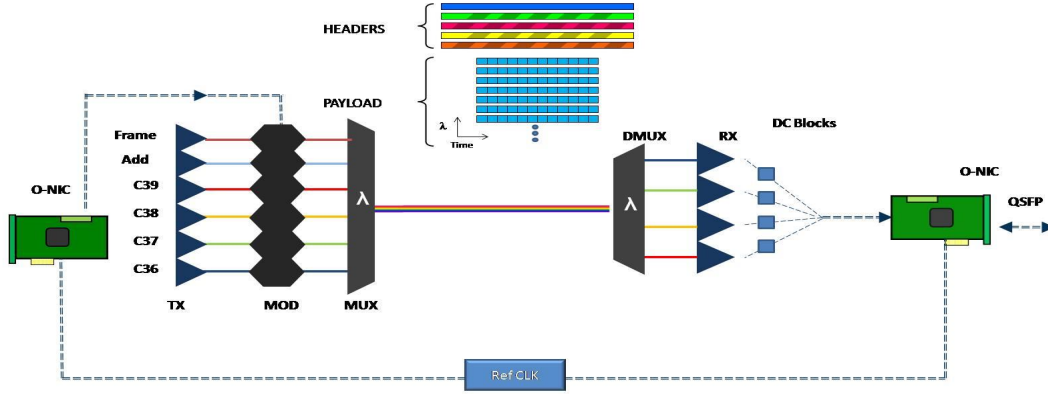


Figure 3.5: Optical-WiMax Test Bed - The physical network setup. A Vertex IV FPGA is used as the Optical-Network Interface Card (ONIC) and modulators are used to encode the WDM striped data.

A	Frame	Output
0	0	No Packet
0	1	Packet Valid and switch to B
1	0	No Packet
1	1	Packet Valid and Switch to A

Table 3.1: Logic for wavelength-striped control - Whenever the Frame address bit is on, indicating a valid packet, the address bit controls the destination address, indicating a switch to node A or node B (1 or 0, respectively)

3.3.3 Results

VLC generated video was streamed wirelessly to a WiMax basestation. The Netserve module actively modified the packet destination based on the RSSI

3.3 Experimental Set Up

value of the incoming packets. This can be seen in Figure 3.7, which shows a screenshot of the two ports that the receive node is listening on. The video could be seen switching from one port to another as the mobile client's RSSI changed to above the determined switching threshold. Figure 3.6 shows the output eyes for all ITU [69] spaced channels, 36-39. The WDM optical packets had header wavelengths that changed depending on the destination IP of the packet.

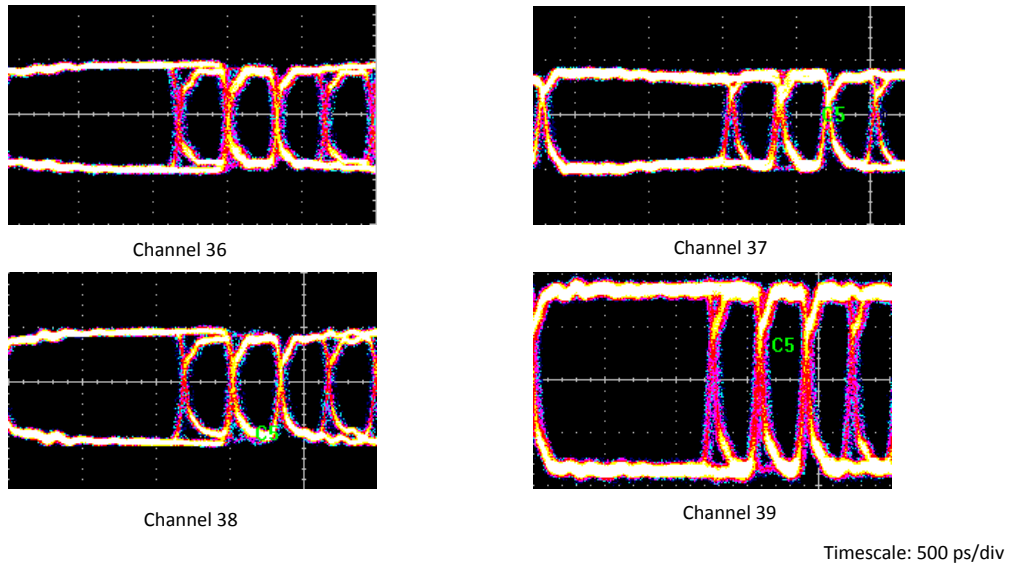


Figure 3.6: Optical-WiMax Results- Eyes - the output eye diagrams for CH36-CH39 of the WiMAX generated video.

The end node client actively listened on the two IP ports for incoming data streams. As the mobile client's RSSI value changed, the video stream could be seen switching from one IP port to the other and back. This experiment was a first step toward dynamic transparent optically connected resources for heterogeneous systems. With applications in mCloud technologies and next-generation cloud computing services, this work is a first step in mobile-wireless and optical network integration and cross-domain awareness.

Switching of data similar to the data presented here was shown in later work, please see Section 4.2.1 for more details on those lessons and how they were addressed in subsequent experimental work.



Figure 3.7: Optical-WiMax Results - At the end node, the CPU is listening on the two IP ports. The VLC video packets can be seen switching from the lower RSSI value IP destination to the stronger RSSI value IP destination after being streamed through the WiMax base station, a VLAN and a transparent optical network.

3.4 Discussion

The goal of the work presented in this chapter is to move beyond the current mobile cloud computing models that are user-unaware, by demonstrating a user-signal-strength-aware dynamic transparent optical system. The user-aware Net-Serv application module and Optical Interconnection Network Interface presented here abstracts away the optical network, which allows the commercial ethernet card to leverage the advantages of optical interconnects and the WDM data systems without changing the underlying architectures that have been developed over the last several decades.

In this experiment, video from the client is dynamically streamed through the WiMAX base station, transmitted over a VLAN and through an O-NIC and WDM encoded on the optical network, and then decoded at the end node. This

process is transparent to the end users. In this work, we explore a photonic network for dynamic switching of WiMax generated video packets transparently through an optical network.

A Wimax/Optical network test bed is presented, utilizing the ONIC in conjunction with a WiMax antenna and a VLAN connection to a transparent WDM optical network. This test bed dynamically changed the destination address of the packets by looking at the RSSI value of the mobile user. In this work a system was demonstrated that dynamically streams video data from a wireless client through the WiMAX base station. The packets are processed in real time by a netserv module, sent through a transparent WDM optical network, and received by the endnode.

This work demonstrates the ability of optical interconnects to be user-aware, and for the physical and software stack layers to work together to provide transparent, high-bandwidth, dynamic optical systems for next-generation mobile cloud computing applications. These systems must be high performance, with high bandwidth, improved latency, and improved energy and computing efficiency.

Chapter 4

A Photonic Network for Hardware Accelerator Enabled Utility Computing

By leveraging the inherent data parallelism optical interconnection can provide, hardware acceleration efficiencies can be improved. In this chapter, the issue of dynamic provisioning of resources is explored. The challenge of network functionality in Optically Connected Hardware Accelerators (OCHA) are addressed by expanding on existing protocols and developing hardware accelerator access protocols that leverage the optical network architecture described in Chapter 2. In order for this architecture to succeed, the data patterns found in heterogeneous computing systems must be understood and properly addressed.

Here, a high bandwidth, reconfigurable OCHA is presented and experimentally characterized that can dynamically allocated hardware accelerators, and allows for the delocalization of hardware acceleration in heterogeneous utility computing. In this experiment we validate our proposed architecture with a FPGA-based emulation test bed. The optical packets generated by the FPGA are sent through a semiconductor optical amplifier (SOA)-based, wave-length stripped, optical network, and utilizes a XOR phase-encoded header for routing.

The two characteristics addressed in this work are the multicasting abilities of optical networks and the lack of reconfigurability in current heterogeneous systems. Large amounts of data are often offloaded to the accelerator, and the

location of accelerator, as they appear to the processor is often Current networks for heterogeneous systems do not leverage the data parallelism inherent in many hardware accelerated applications, nor can the electronics do much to mitigate the high bandwidth demands.

4.1 Background

Multicasting data is an integral part of many hardware accelerator architectures. An ideal architecture for hardware accelerators would be massively data parallel. An optical network is capable of this data parallelism and has previously been shown to interface well with memory [56]. In order for this architecture to succeed, the network must be able to actively switch and multicast. In this work, we experimentally demonstrate a dynamic, optically switched and multicasted network that uniquely exploits the parallelism of wavelength-division multiplexing (WDM) in order to serve as an initial validation for our proposed architecture.

4.2 Lessons from Previous Work

OCHA and the Optically Connected Resources Module 4.3, build on preliminary work on optical interconnection network interfaces, as well as preliminary work in the realm of optically connected memory, seeking to optimize technologies found in these systems for hardware accelerator systems. Throughout these earlier experiments, the overall performance of the systems were limited by a number factors associated with using off-the-self FPGA development boards and commercial optical components.

4.2.1 Bandwidth Mismatches

In the initial work into dynamic optical interconnects, the aim was to achieve fast switching times of FPGA generated packets that well-emulated the data patterns found in typical heterogeneous utility computing systems. At the same time, an all optical solution was pursued, and thus devised the switching protocol outlined in Chapter 2.

The initial switching implementation was to have the header wave lengths function as a packet switch, where in the signal stayed high (logic 1) or low (logic 0) for the duration of the packet. However, upon incorporating this switching/multicasting technology into our test bed, we discovered an AC coupling issue in the PIN-TIA receivers on the optical switching boards.

In order to achieve the desired switching times, we needed to use specific PIN-TIA receivers, which unfortunately needed to see a positive edge within a specific frequency range. This maximum time between positive edges of the receivers was less than the minimum length of the phase-lock loops (PLLs) on the FPGA to lock onto a data signal.

Thus, in this work, a Phase-Encoded Header was designed and implemented to achieve the fast switch of all-optical data while also allowing the architecture to send and receive large packet as they do in heterogeneous utility systems.

4.2.2 Burst Mode Receivers

Early optical memory work, as well as the Optical Network Interface work suffered from clock and data recovery overhead. Clock recovery is the process in which high speed serial links that operate without a shared clock [70], require a phase-lock-loop to align based on a series of incoming zeros and ones. This process is known as clock recovery and adds latency to the system. In traditional electronic links, once the link is established during a reset or system power-on, it is maintained indefinitely using idle data, as it did in Section 3.2. Consequently, there has been little focus on fast clock recovery methods.

However, as recent needs in energy efficiency in computing have lead many systems to turn towards standards in which idle links are turned off to increase energy efficiency. And, while this may be effective in lowering energy rates, the process of repeatedly turning off links makes clock recovery a more pressing matter then a simple start up overhead. Additionally, the move to optical interconnects with switching will likely prevent the transmission of idle data [20, 58, 71, 72], further increasing recover overhead. Due to the large size of the data packets in many heterogeneous systems (monte carlo simulations, market data, and graphics data) a certain amount of latency is tolerable.

4.3 Optically Connected Resources Module

However, in order to leverage and maximize the performance and efficiency benefits of optical interconnection networks as a whole without suffering the frequent recovery overhead, each of which could be several microseconds, focus has recently shifted to high-speed burst-mode receivers [73, 74]. Burst-mode receivers are designed to operate with unpredictable traffic patterns in the absence of "idle data", reducing the recovery overhead significantly. In [74], a locking time of 31 ns was achieved on 25 Gb/s data, and in [75] a locking time of less than 10 ns was achieved on 10 Gb/s data.

FPGAs, processors, accelerators and commercial FPGA development boards do not contain fast burst-mode receivers, and therefore integrating any existing processors and/or FPGAs with optical switching will suffer from high clock recovery overhead. To address this issue, the OCRM was designed with high-bandwidth expansion slots to enable future integration of burst-mode receivers like the ones in [74, 75]. In this scheme, a separate daughter card could be created that implements burst-mode receiving and serializer-deserializer (SerDes) functionality. This card would process the serial data and deliver equivalent parallel data and clock to the FPGA. This process would completely bypass the serial links of the FPGA used in this dissertation and would avoid the high recovery overhead inherent in the FPGA's serial transceivers. Using a daughter card in the expansion port allows for various burst-mode receiver implementations, which is a necessity due to the nature of ongoing burst-mode receiver research, and the resulting lack of commercial options at current.

4.3 Optically Connected Resources Module

The Optically Connected Resources Module (OCRM) is a custom FPGA-based design and test board fabricated for and used in this dissertation in experiments detailed in Chapters 4 and 5. The OCRM is a custom FPGA-based board that enables the implementation and characterization of diverse OCHA architectures. The main advantage of using the OCHA is the fast prototyping abilities, reconfigurability and the close integration with high-speed serial transceivers by way of a high-performance FPGA. The FPGA provides inexpensive and flexible prototyping functionalities that are not possible using application-specific integrated

4.3 Optically Connected Resources Module

circuits, or off-the-shelf processors. This in turn creates a low-latency interface between the electrical domain within the hardware and the optical domain within the network

The OCRM may be located physically distant from the associated processor, and thus leverage the distance-immunity of single mod optical fibers. Relocating the accelerator devices to be physically distant, but logically near, the processor frees up board space near the processor without significant impact on the system. This in turn will give system and architecture designers flexibility. Additionally, more accelerators than accessible electronically can be accessed by all the processors in the system.

The OCRM consists of an Altera Stratix IV FPGA [76] with twelve bi-directional transceivers, each of which is capable of 11.3 Gb/s operation, replaceable DDR3 Dual Line Memory Module (DIMM) , a 10/100-MB/s ethernet port, expansion ports capable of supporting peripherals with over 135 Gb/s aggregate bandwidth, and banks of general purpose input/output (GPIO) pins, one of which is a Matched Impedance Connector (MICTOR). Figure 4.1 shows a photograph of the OCRM while Figure 4.2 shows the physical mapping of each component over the OCRM photograph.

The OCRM is configured such that the FPGA can function as a processor emulator and a hardware accelerator emulator, that can communicate with a local photonic transceiver chip, which serves as the interface between the OIN and the compute elements. The photonic transceiver chip is a combination of the SerDes logic and high-speed transceivers on the FPGA, and discrete optical components. The FPGA was chosen for its relative inexpensiveness when compared to building custom ASICs, its flexibility, and the ability to do fast prototyping of different compute functionalities.

4.3.1 FPGA Hardware Design

The hardware structures implemented on the FPGA on the OCRM are created using the Verilog Hardware Description Language (HDL) [77]. The top-level entities are the Logic (CPU Emulator or Accelerator Emulator), and the SerDes (Figure 4.1). Also of note is the use of the MICTOR GPIO port for header

4.3 Optically Connected Resources Module

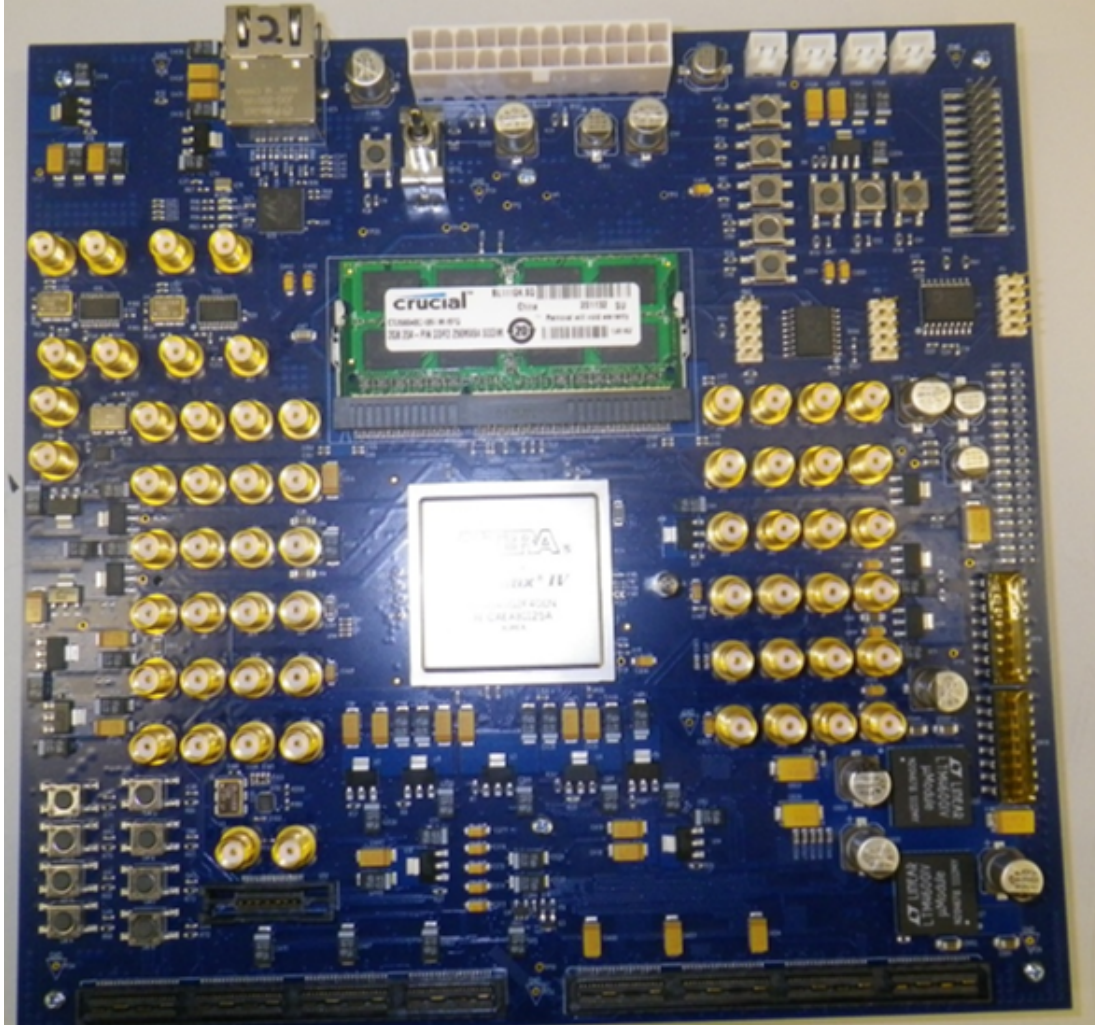


Figure 4.1: Picture of OCRM - A picture of the OCRM module featuring a Altera Stratix IV FPGA, DDR3, 10/100 Mb/s ethernet port, bi-directional transceivers, expansion ports for daughter cards, and MICTOR GPIO

wavelength control (please see ??). The SerDes module receives serial data to be de-serialized and relayed to the logic module (in these experiments 40 bit words), or serialized logic module data to relay to the photonic transceivers, with minimal latency.

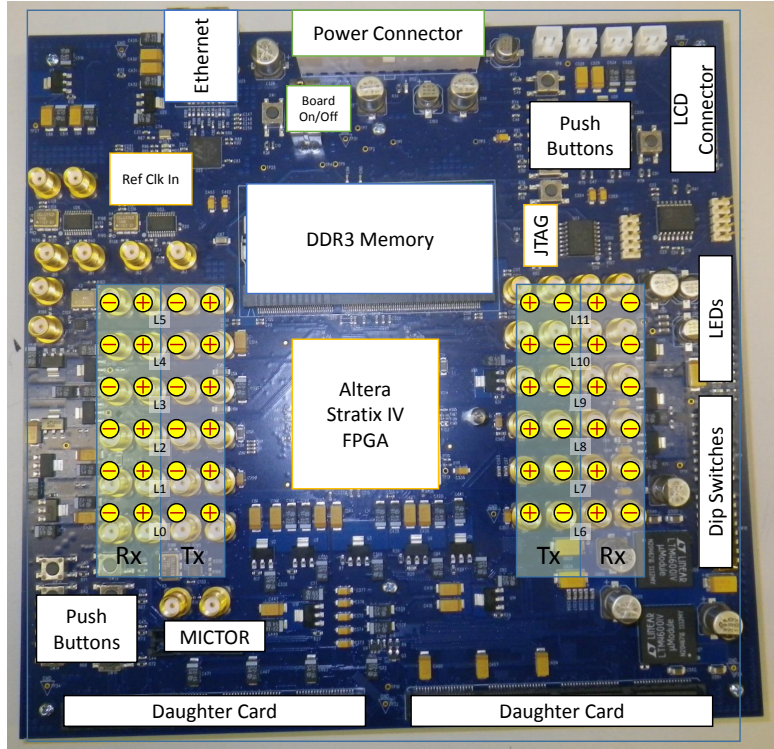


Figure 4.2: Diagram Overlaying Picture of OCRM - A diagram of the OCRM overlaying a photograph of the module identifying the locations of the Altera Stratix IV FPGA, the DDR3, 10/100 Mb/s ethernet port, bi-directional transceivers, expansion ports and the MICTOR GPIO

4.4 Experimental Set Up

This experiment demonstrates the feasibility of the system as well as the efficiency of optically connected networks for utility computing. The system uses two Altera Stratix IV FPGA boards to emulate the CPU and hardware accelerator nodes. A third node is attached to a Bit Error Rate Tester (BERT) to confirm error free operation. The network control signals utilize low-speed general purpose input/output (GPIO) pins on the board to drive four SOAs to modulate the control bits for the network.

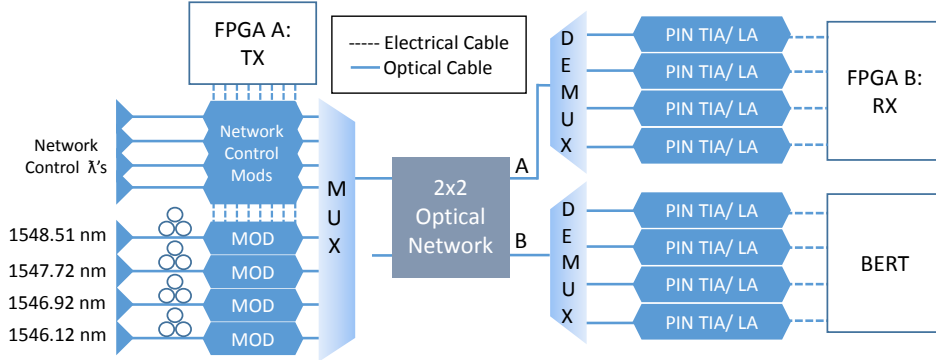


Figure 4.3: Experimental Set-up - FPGA A modulates four payload channels and four network control wavelengths over a 2x2 actively switched network test-bed. FPGA B and the BERT receive these payloads from the optical network using four PIN-TIA receivers.

4.4.1 XOR Phase-Encoded Header

The header wavelengths are driven by GPIO pins on the OCRM, and internally on the logic array by an instantiation of the AltPLL I/O MegaFunction wizard. For the purposes of this experiment, the following parameters were set. Though, it should be noted that for even faster switching times, a faster phase encoded header can be used. the input clock was 10.089 MHz signal, in which the PLL was selected automatically. The PLL used the feedback path inside the PLL in normal mode, with no additional inputs or outputs with autobandwidth settings. The clock frequency parameters were a multiplication factor of 1, a division factor of 2017 to create the 800 ns pulsed signal, with a clock duty cycle of 50 %. These signals then drove low-speed optical amplifiers on control wavelength in the ITU [69] grid dedicated as Frame, A and B (1555.73, 1535.04, and 1533.47 nm, respectively) in table 4.1.

At the input to the switch, the packet goes through a 70/30 splitter. The 70 side of the split, containing both the header wavelengths and payload wavelengths continues on as the data packet. Meanwhile the 30 side is decoded for the header wavelengths. The header wavelengths then go through an Optical-Electronic conversion at the photodiode Trans Impedance Amplifier / Limiting Amplifier (PIN-TIA/LA) input port of the wavelength-stripped switch, while the rest of

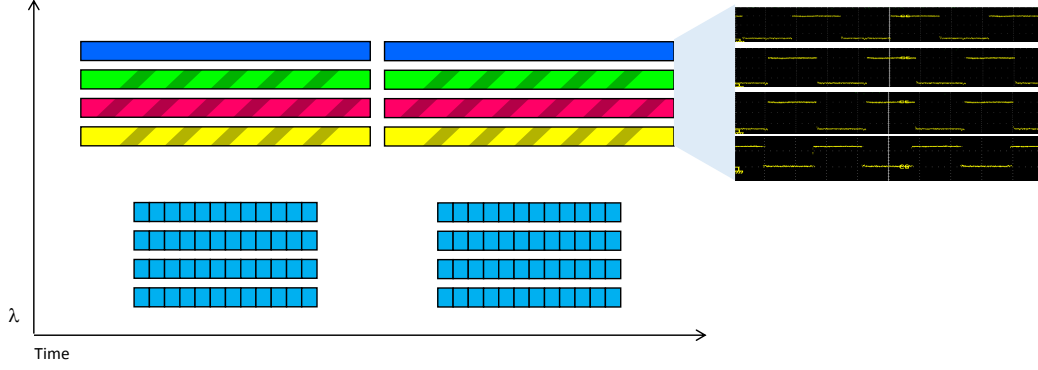


Figure 4.4: Message Format of XOR Phase-encoded Header - Low-speed header wavelengths are phase-encoded and sampled on the positive edge of a phase-offset sample clock. These header wavelengths are combined with high-speed payload wavelengths using WDM.

B	A	Frame	Output
0	0	0	Switch to Node A and B
0	0	1	Both off
0	1	0	Switch to Node B
0	1	1	Switch to Node A
1	0	0	Switch to Node A
1	0	1	Switch to Node B
1	1	0	Both off
1	1	1	Switch to Node A and B

Table 4.1: XOR Phase-encoded Header logic table - On the positive edge of the sample clock, which is time delayed, the output SOAs are controlled by an XOR of their control bit with the frame bit. In the above table, these bits are labeled A and B. The logic explains what Nodes would be switched to in each scenario. This state is held until the next positive edge of the sample clock.

the packet is kept on a FDL.

By eliminating the typical OEO conversions that you would see in a switching node, we save both time and energy, while speeding up performance (Please see

2.1 for an schematic and picture of this switching hardware). These elements are timed such that the delay through the fiber is the same as the electronic delay through the switching node (approx 10 ns).

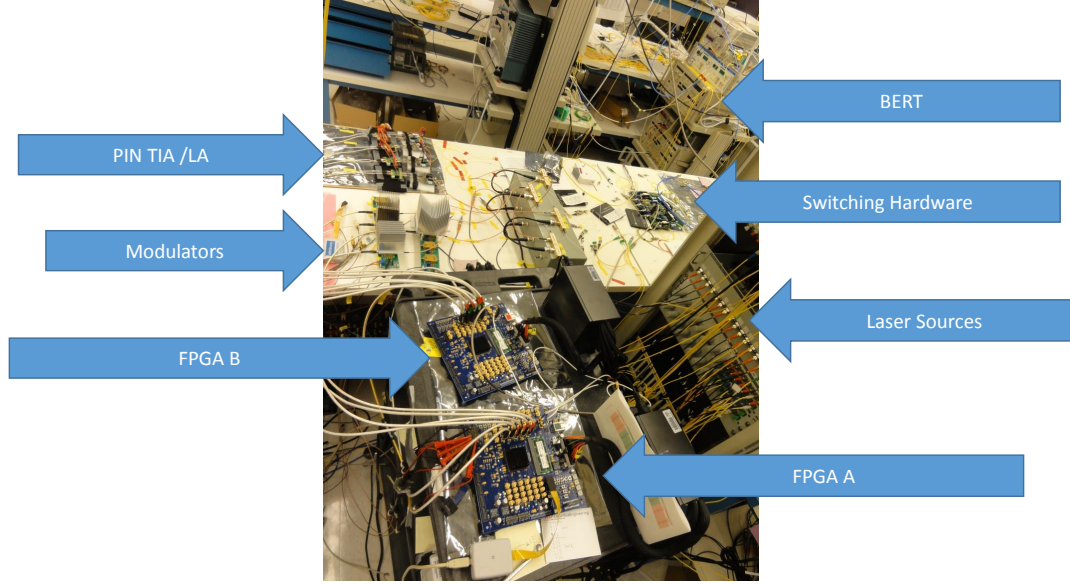


Figure 4.5: Photograph of Experimental Set-up - FPGA A modulates four payload channels and four network control wavelengths from the laser source trays over a 2x2 actively switched network test-bed. FPGA B and the BERT receive these payloads from the optical network using four PIN-TIA/LA receivers.

In the switching node, the electronic header controls enter a Complex Programmable Logic Device (CPLD)[78] where logic is in place to control a series of SOAs that either allow the data packet to pass, or suppress it. As seen in Table 4.1, the output of each port is individually controlled by a specific control signal. The output of each port is individually controlled using a bitwise XNOR of that ports control signal and a framing or reference signal, as seen in the logic table 4.1. This logic adds an additional wavelength to the network control, a minimal cost to the system. To avoid issues of clock skew, this output control is calculated on the positive edge of a phase-shifted, pulsed sample clock, as illustrated in Figure 5.4.

4.5 Results

Error free propagation was verified on one receive path with a Bit Error Rate Tester (BERT) with bit-error rates (BERs) less than 10^{-12} for all 4 payload wavelengths using 11.3 Gb/s data rates. The other receive node (FPGA B) received 100us packets that were actively multicast and switched through the network. Figure 4.6 shows the optical XOR Phase-encoded header scheme headers as detected on the CSA. A few of the 8 possible states are shown in this figure. Figure 4.7 shows the dynamic switching of 100 us packets.

We validate a proposed architecture with an experiment that leverages optical interconnects to demonstrate error free active switching and multicasting of FPGA generated and received packets.

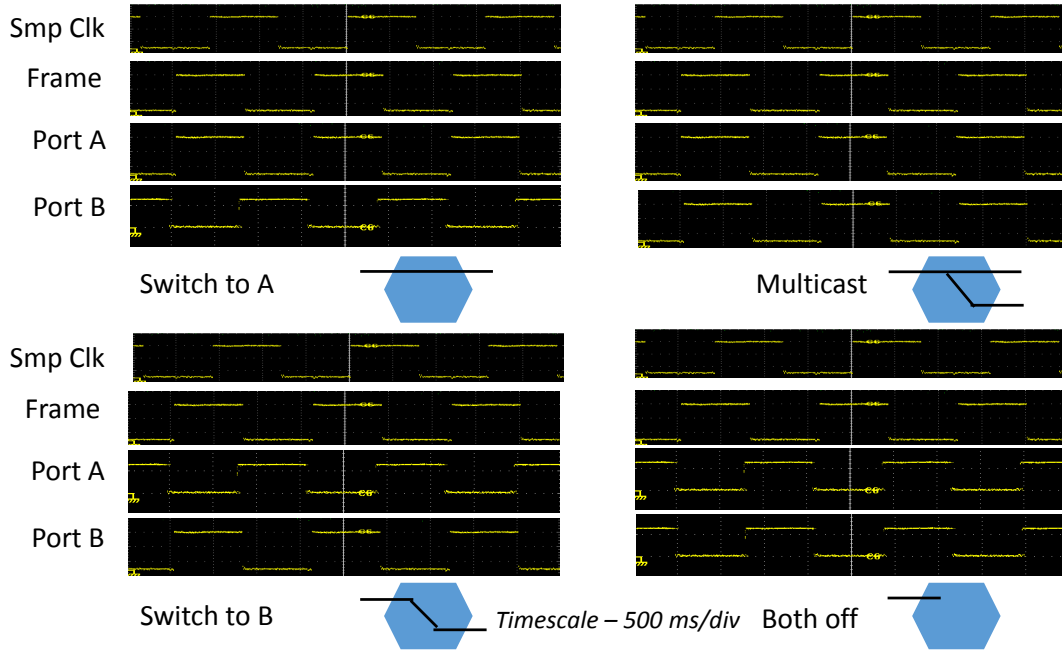


Figure 4.6: XOR Phase Encoded Header - XOR-Phase-encoded header network control- for (a) switch to output A (b) multicast (c) switch to output B and (d) both off

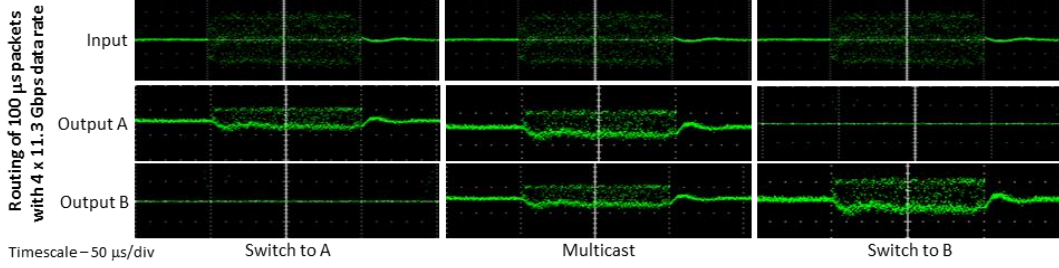


Figure 4.7: Packet Routing - input and output Packets when (a) switch to output A (b) multicast (c) switch to output

4.6 Discussion

This chapter demonstrates how OCHA enable novel system configurations, and therefore more efficient and resilient architectures. Heterogeneous systems, and accelerator use is continuing to grow in scale and complexity, and will likely incorporate multiple thousands of processors and accelerators, leading to many potential idle components and inefficiencies, as well as possible failure points.

As these utilities are following a CAPEX model of business, these failures and inefficiencies cut into cost and computing power. As a result, next-generation heterogeneous utility computing systems will be required to surpass today's level of resilience and efficiency. The limitations of electronic interconnects prohibit modern accelerator technologies from implementing the architectures and acceleration need in future heterogeneous utility systems. To address these challenges, OCHA replace the electronic accelerator bus with an optical interconnection network, thereby enabling these novel accelerator architectures and distribution techniques.

The OCRM used here is a first-of-its-kind prototyping module that enable efficient, high performance all-optical communication between a processor and delocalized hardware accelerators. As an FPGA-based platform, it allows for reconfigurable test-bed implementations for explorations into the architectural

and network design of optically connected acceleration (both ASIC and FPGA based) in heterogeneous computing.

By drawing on the lessons learned using off-the-shelf FPGA-based development boards, the custom OCRM was created to address the challenges facing OCHA development. These hindrances would otherwise have remained unknown—including the importance of burst-mode receivers, and the AC coupling-bandwidth mismatch of the switching elements and the OCRM. This in turn allows novel accelerator architectures to be developed, prototyped and experimentally characterized in a lab environment.

In this work, we propose and validate a photonic network for delocalized hardware acceleration in utility computing. This technology leverages the bandwidth distance product gained by WDM optical transmission to create a system where delocalized hardware accelerators can be dynamically allotted to different applications at run-time. High-bandwidth connectivity provided by WDM optical interconnects is an important enabler for delocalized hardware accelerators in utility computing.

A high throughput, dynamically reconfigurable optical network is demonstrated on an FPGA-controlled platform, with a CPU emulator, Bit Error Tester and hardware accelerator emulators. This work represents a significant step toward integration of optical links into hardware accelerator enabled data center architectures. The sum of the work in Chapter 2 and this chapter illustrate the need and feasibility of optically connected resources, and their flexibility.

This work demonstrates the need for low-latency, high-performance optical interconnects within future large-scale heterogeneous utility systems. These OCHA systems must also be flexible, use optical switching, and optical multicasting, and be bidirectional to enable innovative systems architecture for future heterogeneous utility systems with improved bandwidth, latency, and efficiency.

Chapter 5

FPGA Implemented Bidirectional OCHA

In this chapter, a dynamic, high bandwidth, bidirectional OCHA is presented and experimentally characterized that can dynamically allocated hardware accelerators, can also address network incast bottle issues plaguing these systems today, and allows for the delocalization of hardware acceleration in heterogeneous utility computing. In this experiment we validate our proposed architecture with a FPGA-based bidirectional emulation test bed. The optical packets generated by the FPGA are sent through a semiconductor optical amplifier (SOA)-based, wave-length stripped, optical network, and utilizes a phase-encoded header for routing.

5.1 Background

In this chapter, the issue of dynamic provisioning of resources is explored. By combining heterogeneous systems with economies of scale, comparatively lower capital expenditures, and most importantly, dynamic provisioning, the benefits of heterogeneous clouds and utility systems can be further realized [23]. The challenges of network functionality and bidirectionality in Optically Connected Hardware Accelerators (OCHA) are addressed by expanding on existing protocols and developing hardware accelerator access protocols that leverage the optical network architecture described in Chapter 2.

In order for this architecture to succeed, the data patterns found in heterogeneous computing systems must be understood and properly addressed. The advantages of optical interconnects must also be leveraged. The FPGA was chosen due to its unique ability to offer fast prototyping at low cost. The photonic interconnection network employs a bidirectional switching design to maximize the dynamic re-configurability capabilities. By exploiting the bidirectional transparency of the SOA [79], the underlying optical medium, and by utilizing optical circulators, we are able to build a network that uniquely addresses many issues facing heterogeneous utility computer architectures today.

The two characteristics addressed in this work are the data movement protocols and the incast bottleneck described in Chapter 1 [2, 4]. Large amounts of data are often offloaded to the accelerator, and, after some processing, a smaller amount of data is sent back to the CPU. Current networks for heterogeneous systems do not leverage the data parallelism inherent in many hardware accelerated applications, nor can the electronics do much to mitigate the network incast bottleneck. This experiment emulates this scheme and explores the feasibility of the architecture to provide bidirectional, reconfigurable, dynamic hardware accelerator allocation.

5.2 Overview of OCHA

In contrast with current models in which accelerators are hardwired to a specific location and function, in the proposed scheme, hardware accelerators are dynamically allocated at runtime, and thus become an application-specific configuration. The remote hardware accelerator architecture presented here utilizes the bandwidth and latency gains of optical networks to allow for a bank of hardware accelerators to be located remotely, yet still appear to be physically near the CPU [5.1]. The bank of hardware accelerators can be dynamically allocated to CPUs, and not restricted to the single local CPU.

This design scheme makes it possible for multiple hardware accelerators to be dynamically utilized by different CPUs. By integrating this architecture into a utility computing system, powerful, application-specific hardware will be available

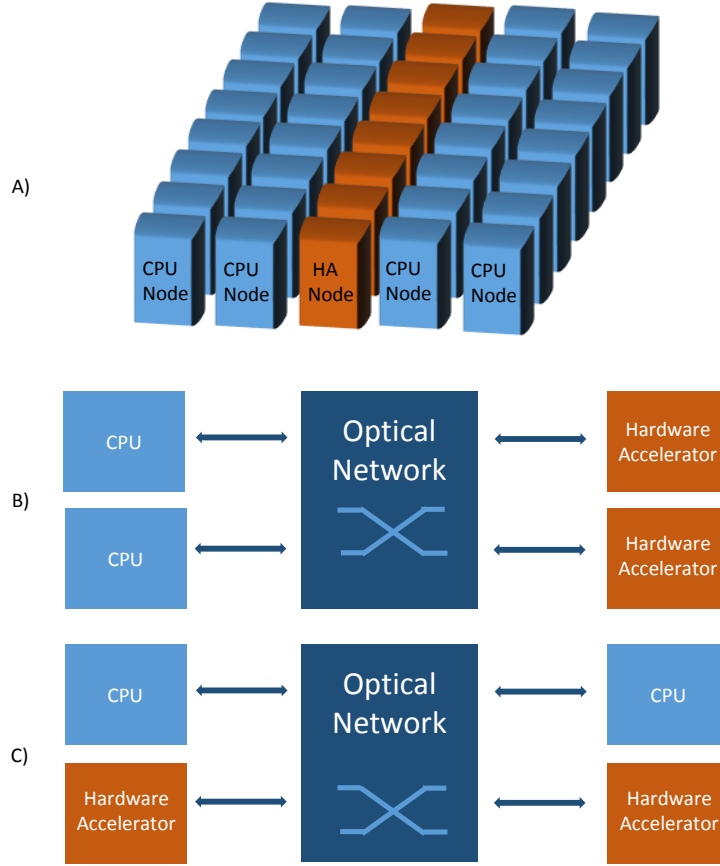


Figure 5.1: Architectural Design - (A) CPU nodes with separate optically-connected hardware accelerator nodes that can be dynamically configured. The hardware modules could either be configured as a scheme where (B) the hardware accelerators are organized as a central bank or (C) each CPU has a dedicated hardware accelerator that it rents out to the system when not in use.

for applications to use. Leveraging the speedups provided by hardware accelerators, as well as the latency gains of optics, this system would speed up compute times without sacrificing energy efficiency, data center footprint, or cost.

The photonic interconnection network design makes it possible for multiple hardware accelerators to be dynamically utilized by different CPUs. By integrating this architecture into a utility computing system, powerful, application-specific hardware will be available for applications to use. By leveraging the

speedups provided by hardware accelerators, as well as the latency gains of optics, this system can minimize latency without sacrificing energy efficiency, data center footprint, or cost.

The WDM parallelism further enables several accelerator chips or the entire bank of accelerators to be accessed in parallel over a single fiber, thus increasing the accelerator-CPU bandwidth [22]. These hardware modules could either be a central bank that is allocated by a controller [5.1B], or rented out from other CPUs [5.1C]. Overall, delocalizing the hardware accelerators and replacing electronic buses with optical interconnection networks will not only increase bandwidth, reduce system wiring complexity, and lower energy consumption; delocalized accelerators will also allow the current trends for increased hardware acceleration in utility computing to continue scaling.

5.3 Experimental Set Up and Results

The test bed uses four custom Altera Statix IV FPGA boards to emulate the CPU and hardware accelerators. Error-free propagation is verified on the receive boards themselves, with each board utilizing the bidirectionality to send back an error count on a unique wavelength. As seen in Figure 5.2, the send node (FPGA 0), the CPU emulator, generates the four lanes of PRBS data in 100 us packets and control wavelengths. In order to maximize efficiency and minimize switching time, we implement a wavelength-stripped, phase-encoded header network control protocol.

5.3.1 PRBS Generation

The FPGA 0 acts as our CPU emulator. In this experiment, it generates a 40 bit PRBS. The PRBS was chosen due to the fact that although it is deterministic, it seems to be random. The PRBS is generated using a series of linear feedback shift registers (LFSR), using an exclusive-nor (XNOR) of the bits 40, 38, 21, and 19 (as per [78]) to generate the PRBS. This logic is illustrated in Figure 5.3.

This PRBS data is sent on a 4x11.3 Gb/s network using WDM of ITU channels 36, 37, 38, and 39 (1548.51, 1547.72, 1546.92, and 1546.12 nm) [69], using

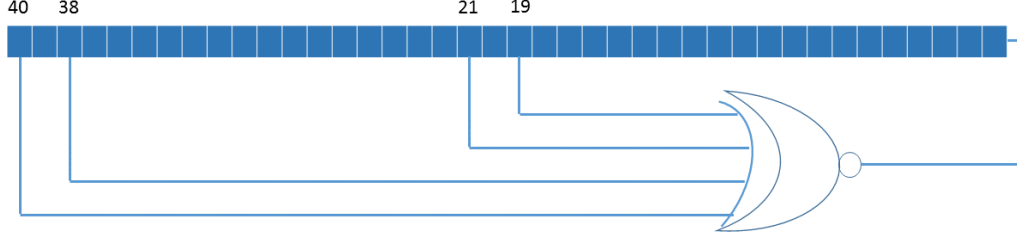


Figure 5.3: PRBS generated using Linear Feedback Shift Registers - A 40-bit PRBS generator. The XNOR gate provides feedback to the registers as it shifts from right to left. The maximal sequence consists of every possible state.

wavelengths could be used to switch to three end nodes instead of two. Each end node can be individually switched to, and thus this phase-encoded header would require a $N+1$ header wavelengths for N destination nodes.

At the input to the switch, the packet goes through a 70/30 splitter. The 70 side of the split, containing both the header wavelengths and payload wavelengths continues on as the data packet. Meanwhile the 30 side is decoded for the header wavelengths. The header wavelengths then go through an Optical-Electronic conversion at the photodiode Trans Impedance Amplifier / Limiting Amplifier (PIN-TIA/LA) input port of the wavelength-stripped switch, while the rest of the packet is kept on a FDL.

By eliminating the typical OEO conversions normally seen in a switching node, we save both time and energy, while speeding up performance (Please see 2.1) for an schematic and picture of this switching hardware). These elements are timed such that the delay through the fiber is the same as the electronic delay through the switching node (approx 10 ns).

In the switching node, the electronic header controls enter a Complex Programmable Logic Device (CPLD) where logic is in place to control a series of SOAs that either allow the data packet to pass, or suppress it. As seen in Table 5.1, the output of each port is individually controlled by a specific control signal. To avoid issues of clock skew, this output control is calculated on the positive edge of a phase-shifted, pulsed sample clock, as illustrated in Figure 5.4.

5.3 Experimental Set Up and Results

A2	A1	A0	Output
0	0	0	All off
0	0	1	Switch to Node 0
0	1	0	Switch to Node 1
0	1	1	Switch to Node 0 and 1
1	0	0	Switch to Node 2
1	0	1	Switch to Node 0 and 2
1	1	0	Switch to Node 1 and 2
1	1	1	Switch to Node 0, 1, and 2

Table 5.1: Phase-encoded Header logic table - On the positive edge of the sample clock, which is time delayed, the output SOAs are controlled by the state of their control bit. In the above table, these bits are labeled A2, A1, and A0. The logic explains what Nodes would be switched to in each scenario. This state is held until the next positive edge of the sample clock

5.3.3 Error Checker and Results

Each transceiver bank interfaces with optical components to generate (FPGA 0) and receive (FPGAs A, B, and C) 4x11.3 Gb/s WDM data transactions. At the hardware accelerator emulator nodes, FPGAs A, B, and C, the data is decoded and goes through an error checker. Once the transceiver phase lock loop locks onto the incoming data stream, verifying that valid data is on the transceiver, the PRBS data is deserialized and sent to be checked on an error checker.

The error checker takes the predictable PRBS pattern generator and compares the received PRBS sequence with the expected one. With each new packet, the error checker generates the next, predictable number in the pattern with its own LFSR XNOR of the bits 40, 38, 21, and 19. This number is cached and on the next packet, the cached prediction is compared with the new packet. If there is a discrepancy, the error count is incremented, if the numbers match, the error count remains the same.

In generating the PRBS on the emulator nodes, we emulate the behavior of a hardware accelerator. The error count is then sent back on the nodes return path,

5.3 Experimental Set Up and Results

through an optical circulator, to the original CPU-emulator node. Each node has a dedicated wavelength for its error count. This node then reads each error count, ensuring that the packets were sent error-free. In this work we verified error-free propagation of 4x11.3 Gbps 40-bit PRBS [80] data over a bidirectional network. The results of a few of these iterations can be seen in Figure 5.5.

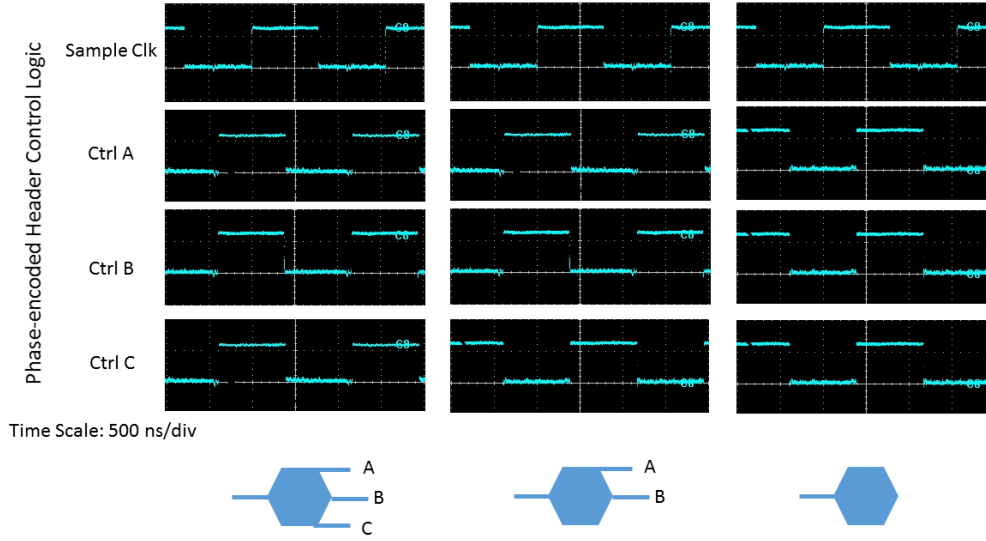


Figure 5.4: Phase-encoded header design scheme - On the positive edge of the sample clock, which is time delayed, the output SOAs are controlled by the state of their control bit. This allows the system to individually control each output port and allows for a logical multicast. In these scenarios, this system would multicast, switch to A and B, and turn all the ports off, respectively.

Fig. 5.4 illustrates a few examples of this phase-encoded header control logic for a multicast, a switch to Nodes A and B, and all outputs off, using screen captures from the Continuous Spectrum Analyzer (CSA).

The use of circulators for the bidirectional, optical return path also addresses the incast bottleneck facing many heterogeneous systems. Optical circulators are passive devices that do not contribute to power consumption in the system. The optical return paths are multiplexed together onto one return fiber. In an electronic system, this incast would require dedicated links from each node, where in the optical system, this is not the case.

5.3 Experimental Set Up and Results

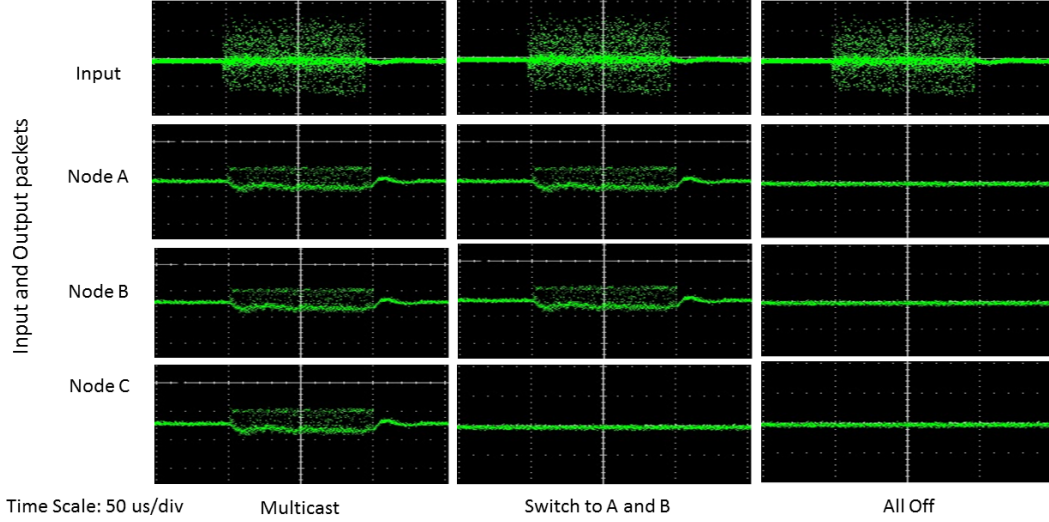


Figure 5.5: Data Packet Switching - Switching of a 100 us packet through the network in a multicast, a switch to A and B, and all off. The input packet can be seen in the top box, while the optical outputs can be seen in each scenario in the lower boxes.

Additionally, in some system designs, the accelerators could leverage a tunable laser (TL) instead of the fixed lasers used in this experiment. In this scheme, the accelerator could share and optical return path. By further leveraging modulation formats, this paradigm could be optimized further.

The use of wavelength-striped optical multicasting reduces overall access latency while also increasing bandwidth. Optical multicasting allows any set of OCHA node to be accessed in parallel, and thus combines the bandwidth of each OCHA node to provide the processor with greater aggregate accelerator bandwidth.

Latency is reduced, as detailed in Section 4.2.2, due to the the tens of nanoseconds of latency each independent accelerator access incurs. Constraints from energy dissipation, pin count, wiring complexity, and available space would require an electronically-connected memory system containing the accelerator devices proposed here to perform many accelerator accesses serially rather than simultaneously.

Here, the total accelerator access latency is reduced to that of a single accelerator access, regardless of the number of OCM nodes access by simultaneously access any set or subset of OCHA nodes with the use of the optical multicast approach. This experimental demonstration of accelerator data across one, two, or three nodes all incur the same access latency (time of flight + clock recovery time). Similarity, the latency incurred for accessing a entire OCHA node (be it 36 or 72 or more) would also be the time of flight + clock recovery time.

5.4 Discussion

This chapter demonstrates how OCHA enables novel architectures and accelerator access protocols and therefore more resilient and flexible heterogeneous systems. Large-scale computing systems continue to scale and incorporate acceleration and will soon likely include diverse many thousands of nodes systems with GPUs, FPGAs, FFP, and CPUs working in tandem on computational tasks. Consequently, next-generation heterogeneous utility systems will need to surpass today's level of reconfigurability and resilience.

The limitations of electronic interconnects limits and at times prohibits the ability of modern systems to in implementing the architectures and delocalization protocols necessary in future heterogeneous utility computing systems. To address these challenges and opportunities, the OCHA replaces the electrical connection bus to accelerators with and optical interconnection network, thereby enabling novel architectures that leverage the data parallelism in accelerator data, reduce latency, and offered reconfigurability.

The data behavior in this test bed is an emulation of the data patterns found in many heterogeneous utility computing systems, in that a large amount of data is multicasts or switched to a number of accelerator units, whom respond with a smaller data set after some computation. In this work, we further propose and validate a photonic network for delocalized hardware acceleration in utility computing.

In generating the PRBS on the emulator nodes, we emulate the behavior of a hardware accelerator. In calculating the PRBS and creating an error checker on the receive FPGAs, a hardware accelerator node is emulated. The error count

is then sent back on the nodes return path, through an optical circulator, to the original CPU-emulator node.

This technology leverages the bandwidth distance product gained by WDM optical transmission to create a system where delocalized hardware accelerators can be dynamically allotted to different applications at run-time. A bi-direction, dynamically reconfigurable optical network is demonstrated on an FPGA-controlled platform, with a CPU emulator and hardware accelerator emulators. This work represents a significant step toward integration of optical links into hardware accelerator enabled data center architectures.

Chapter 6

Summary and Conclusion

The work presented in this dissertation has focused on the design and implementation of optically-connected hardware accelerators for heterogeneous utility computing. This final chapter discusses the accomplishments present here and summarizes ongoing and future work that could lead to the commercialization of the work presented in this dissertation. In closing, a summary of this body of work is presented.

6.1 Overview

This work is primarily motivated by the growing trend in utility computing towards heterogeneity, the resulting issues with placement and networking to hardware accelerators and the need for a fundamental redesign of processor-accelerator communications. The use of specialized hardware either in the form of an ASIC like a GPU or FPP, or reconfigurable, like a FPGA.

However, the use of hardware acceleration creates a bottleneck and localization constraints due to its reliance on electrical interconnects that limit performance, efficiency, and scalability. The resulting restrictions placed the system become a limiting factor in the overall performance of heterogeneous utility computing systems.

Optically-connect hardware accelerators have been presented here as a solution to the bottleneck, localization limitations, and network latency found in current heterogeneous utility computing systems, owing to its ability to eliminate

the electronic bus to hardware accelerators. The need for a switch to optics in utility systems and the increasing heterogeneity of these systems has become apparent to the optical research community and industry, and the work presented in this dissertation take the first steps in achieving this endeavour.

By replacing the wide electronic bus (eg, PCIe) to hardware accelerators with an optical interconnection network, future utility computing systems will be about to dynamically access vast quantities of specialized and reprogrammable acceleration, and maintain continued scaling of these systems. Throughout this dissertation, three key metrics for next-generation systems are addressed:

- **Bandwidth** - The low bandwidth-density of electrical interconnects limits hardware accelerators to localization. This work overcomes these challenges through the high bandwidth-density of optics, which allows terabits-per-second of data to traverse a single fiber using WDM. Furthermore, optical interconnection networks enable greatly increased bandwidth through simultaneous access to multiple OCHA devices with optical multicasting.
- **Latency** - The network latency in heterogeneous systems is a large part of the overall system latency. Future heterogeneous utility computing systems will need to implement networks to hardware accelerators that provide ultra-low latencies, in addition to high bandwidths. The optical network architectures presented here can address this issue through transparent optical routing, in which high-bandwidth WDM messages traverse an optical network with time-of-flight latency. The use of the wavelength-striped phase encoded header optical routing enables a custom network-aware arbiter to execute accelerator allocation with optimal latency.
- **Efficiency** - Two major metrics of performance in utility systems are energy efficiency (green computing) and computing efficiency (computations per unit time) With the growing number of hardware accelerators in cloud and high performance computing systems, the electronic bus linking accelerators to processors is becoming a significant source of power dissipation. Each hardware accelerator added to the system will increase computing efficiency,

if properly utilized, but will also increase the pin count, adding more power-hungry data buffers.

Larger and more accelerator devices speed up computing times but increase the total physical wiring distancing, wiring complexity and those power dissipation. The speedups provided for optical interconnected heterogeneous systems improve on the efficiency of electronic systems. This work demonstrated how integration of optical networks with dynamic allocation will leverage the benefits of heterogeneous utility computing while simultaneously leveraging the energy efficiencies provided by optical networking.

6.2 Future Work

With the accomplishments from these first, critical steps in creating optically-connect accelerator systems, it is important to learn from previous work and address the next steps and remaining challenges. Our next steps in the work will be to continue exploring the feasibility and advantages of this system by exploring a number of characteristics and applications as they apply to this network architecture.

6.2.1 Architectural Design

Due to the varying applications, and accelerators in heterogeneous systems, OCHAs will not be a one-size-fits-all solution. Architectures will have to be designed to determine the optimal bandwidth, placement, and allocations to allow for minimized latency, and maximized bandwidth, energy, and compute efficiencies.

In order to enable next generation systems to perform to their full potential, more work will need to be done to build on the work in this dissertation. Of interest are to explore the architectural configurations for reconfigurable hardware like FPGAs, as opposed to ASICs. Other work ongoing work on GPU integration and popular hardware accelerators like Floating Point Processors (FPP) will also need to be investigated.

Advanced schedule and non-blocking arbitration would also need to be added to the work presented in this dissertation to further investigate the advantages of

this architecture. In order to leverage the advantages of optics without disruption to the ongoing engineering and innovations in utility computing systems, it would be advantageous to explore possibilities in which there is a certain level of abstraction and transparency in the optical network. By leveraging the work presented in this dissertation, a path towards innovation is possible.

Differing communication and parallelism in applications that run on heterogeneous systems will also need to be investigated. In many financial service applications, where market data is streamed uncompressed, more bandwidth would need to be allocated. While, in applications like Gaming as a service, designs with varying GPU configurations for different games (ie graphics intensive, vs physics intensive) would need to be investigated.

6.2.2 Optically Connected Memory

Demands on the memory systems in utility systems have driven the development of processing-memory interconnects that require greater performance and flexibility. As computation systems increase in complexity and heterogeneity, the latency caused by this "memory wall" will be exacerbated, and in order for OCHA to succeed, the memory interface will also need to be addressed.

A great deal of work into optically connected memory shows promise, leveraging many of the same advantages outlined in this dissertation, but uniquely exploiting the optics for problems facing memory systems.

The OCRM presented in this dissertation allows for the easy expansion into incorporating optically connected memory systems under investigation into the heterogeneous systems investigated here. Integrating the use of optically connected memory alongside optically connect hardware acceleration will enable future-generation heterogeneous utility systems to continue maximizing efficiency, minimizing latency, and exploiting optical bandwidth.

6.2.3 Silicon Photonic Integration

Nanophotonic devices like microring resonators and ring-based WDM modulators must be integrated as closely as possible with the electronic driver circuitry

in order to eliminate power-hungry, bandwidth-limited electrical wires. CMOS-compatible silicon photonic devices are especially promising to achieve this goal. The microring resonator shows promise in being a building block for WDM modulators, switches, filters, and photo detectors, all of which are elements needed for the architecture outlined in this dissertation. The work presented here did not progress to include the use of silicon photonics, but future systems will need this integration.

Microring-based optical switches, with a large free spectral range, are particularly attractive for accelerator applications due to their ability to switch many wavelengths. This is a perfect application to the WDM packets demonstrated in this work. In contrast to electronic switches, where higher bandwidth equates to higher power consumed by the switch, in ring resonator based switches, the more data in a packet, the more energy efficient it becomes. However, there are still work to be done in thermal stabilization of microring resonators. Ongoing research in this arena show promise, and would need to be integrated into this system as an architecture entity.

Fabrication technological challenges facing integration of optics into processors must also be addressed. And, until optical devices are integrated in a 3D stack or monolithically, the bandwidth of going off-chip will still impose electronic wiring bottlenecks on utility computing systems. Silicon photonics will also be an integral part of future-generation 3D stacked heterogeneous computing chips, functioning as an additional transport layer.

6.2.4 Burst-Mode Receivers

The need for burst-mode receivers is made ever more clearer throughout the work in this thesis. When processors and accelerators (and even memory) communicate over an optical network, each and every message could become subject to a costly clock and data recovery overhead that would reduce the throughput of the already latency-clogged heterogeneous utility system. Moreover, there is the possibility that optical network architectures may cause messages to arrive at each receiver with different phases and/or power levels, further motivating the need for robust burst-mode receiver circuitry.

The lack of fast burst-mode receivers availability limited the abilities of the work in this dissertation. The OCRM presented here on which a majority of this work has been implemented holds the potential and capacity to support suitable burst-mode receivers through the high-bandwidth expansion port, and further and future work must develop such a receiver and characterize the impact it will have on OCHA systems.

6.2.5 Runtime Allocation Integration

The first commercial OCHA systems will likely be deployed as server or rack sized cluster-scale systems. This system will require a solution that incorporates runtime allocation and software integration into the systems. In this setup, current research into compilation and runtime scheduling of accelerator nodes by a network aware arbiter will be imperative to the success.

By integrating a compiler like Liquid Metal or LegUP the OCHA system would abstract away the complex programming and allocation difficulties holding back current heterogeneous utility computing systems. In this paradigm, software programmers would be unaware of the physical placements and networking of accelerators and thusly the server systems would face less barriers to adoption, as existing software applications would not have to be rewritten for the new OCHA systems.

6.2.6 Commercialization

By leveraging the work presented in this dissertation, as well as ongoing work in photonic integration, heterogeneous architectural design and compilers and runtimes for heterogeneous systems, commercially-viable optically-connected hardware accelerators in heterogeneous systems could be realized on a 5-to-10 year timescale. High-performance optical transceivers already exist that could be packaged with processors and accelerators to improve network latency and of large-scale utility computing systems.

The integration of compilers and runtimes with heterogeneous systems would alleviate allocation and programming challenges in heterogeneous systems and abstract away many of the programming challenges, thus lowering the barriers to

entry and adoption. The integration of photonic transceivers would enable the use of optical switching and optical networks detailed in this dissertation and provide architectural benefits and overall system performance and efficiency. These benefits would then in turn provide the motivation to further investigate and integrate optical components into heterogeneous computing systems to achieve even greater computational and energy performance improvements when compared to existing electrically-connected, localized hardware accelerator systems.

6.3 Summary

The latency caused by the network is quickly becoming a major cause of bottlenecks in heterogeneous utility computing systems. In addition, the location and function of hardware accelerators is increasingly becoming an application-specific computation problem. In general, the response to this issue has been to either deploy as many hard accelerators as electronic wiring permits and let accelerators sit idle when not need by their master CPU, or to optimize processing nodes to cope with a large network latency time.

The inability to break free of this paradigm stems for the reliance on electronic connection networks at the board and rack level of the computing system, and the resultant penalties in energy dissipation and overall scalability. The reluctance to adopt new physical layer technology and the architectures enabled by them is primarily due to the relatively mature state of electronic interconnect technology, its established protocols and therefore pervasiveness, and the resulting high cost and unknowns of migrating to a new technology.

Nonetheless, at present, the growing demands of HPCs and Cloud Computing technologies are exceeding the limits of electrical interconnects. However, computer-directed photonic technologies are reaching the point of maturity, and thus rack-to-rack interconnects are increasingly becoming optical due to the low barrier of entry. Concurrently, revolutions in both software programming and runtime scheduling as well as optical interconnect networks

This dissertation takes the initial steps in integrating optical interconnects into hardware accelerated heterogeneous utility computing systems. By developing architectures, protocols, and physical-layer systems , this work demonstrates the

many advantages of optically connected hardware acceleration, while identifying previously unknown integration challenges.

Some of these challenges have been address, such as a control method for wavelength-striped phase-encoded headers, while other challenges, such as optimal architecture designs, burst-mode receiver integration, and runtime allocation integration are to be addressed in future work.

References

- [1] NVIDIA CORP. **Nvidia GRID Cloud Gaming**. <http://www.nvidia.com/object/cloud-gaming.html>. Accessed: 2015-03-27. iv, 1, 6, 10, 11
- [2] A. HOUSTON AND J. SPORN. **Introducing Nvidia Tesla GPUs For computational Finance**. http://www.nvidia.com/content/tesla/pdf/Finance_brochure_2014_fin2.pdf. Accessed:2015-03-27. iv, 9, 10, 12, 14, 57
- [3] SCOTT SCHNEIDERT, HENRIQUE ANDRADE, BURA GEDIK, KUN-LUNG WU, AND DIMITRIOS S NIKOLOPOULOS. **Evaluation of streaming aggregation on parallel hardware architectures**. In *Proceedings of the Fourth ACM International Conference on Distributed Event-Based Systems*, pages 248–257. ACM, 2010. v, 13, 14
- [4] R. WILSON. **Heterogeneous Computing Meets the Data Center**. *Altera Archives*, May 2014. 1, 8, 12, 13, 20, 57
- [5] MICHAEL A RAPPA. **The utility business model and the future of computing services**. *IBM Systems Journal*, **43**(1):32–42, 2004. 1, 2
- [6] D.K. YESALAVICH. **Switch to Videogame Chips Speeds Trading**. *The Wall Street Journal*, April 2010. 1, 8, 10
- [7] IEEE. **IEEE P802.3ba 40Gb/s and 100Gb/s Ethernet Task Force**. <http://grouper.ieee.org/groups/802/3/ba/index.html>. Accessed:2015-03-27. 1
- [8] M. ARMBRUST, A. FOX, R. GRIFFITH, A.D. JOSEPH, R. KATZ, A. KONWINSKI, G. LEE, D. PATTERSON, A. RABKIN, I. STOICA, AND M. ZAHARIA. **A view of cloud computing**. *Communications of the ACM*, **53**(4):50–58, 2010. 2, 6, 13
- [9] C. VECCHIOLA, S. PANDEY, AND R. BUYYA. **High-performance cloud computing: A view of scientific applications**. In *Pervasive Systems, Algorithms, and Networks (ISPAN), 2009 10th International Symposium on*, pages 4–16. IEEE, 2009. 3
- [10] TOP500 ORG. **The Top 500 List - November 2014**. <http://www.top500.org/>. Accessed: 2014-04-7. 3
- [11] C. WILLARD, A. SNELL, L. SEGERVALL, AND M. FELDMAN. **Top Six Predictions for HPC in 2015**. *Intersect360 Reports*, February 2015. 4, 8
- [12] M.A. TAUBENBLATT. **Optical interconnects for high-performance computing**. *Journal of Lightwave Technology*, **30**(4):448–457, 2012. 4
- [13] K. KEAHEY. **Cloud Computing for Science**. In *SS-DBM*, page 478, 2009. 4
- [14] C. KACHRIS AND I. TOMKOS. **A survey on optical interconnects for data centers**. *Communications Surveys & Tutorials, IEEE*, **14**(4):1021–1036, 2012. 6, 7
- [15] L. WANG, J. TAO, M. KUNZE, A.C. CASTELLANOS, D. KRAMER, AND W. KARL. **Scientific Cloud Computing: Early Definition and Experience**. In *HPCC*, **8**, pages 825–830, 2008. 6
- [16] K. MCINTYRE. **GRID The Future of Gaming**. http://karen-mcintyre.com/wp-content/uploads/2015/02/karenMcIntyre_GRID_v2-1.pdf, February 2015. Accessed:2015-04-1. 6
- [17] WILLIAM VOORSLUYS, JAMES BROBERG, AND RAJKUMAR BUYYA. **Introduction to cloud computing**. *Cloud computing: Principles and paradigms*, pages 3–37, 2011. 6
- [18] IBM. **IBM 2014 Proxy Statement to Stockholders**. http://www.sec.gov/Archives/edgar/data/51143/000110465914025583/a14-2281_2defa14a.htm, April 2014. Accessed:2015-04-1. 6, 33
- [19] P. BAHL, R.Y. HAN, L.E. LI, AND M. SATYANARAYANAN. **Advancing the state of mobile cloud computing**. In *Proceedings of the third ACM workshop on Mobile cloud computing and services*, pages 21–28. ACM, 2012. 6, 33
- [20] N. FARRINGTON, G. PORTER, S. RADHAKRISHNAN, H.H. BAZAZ, V. SUBRAMANYA, Y. FAJMAN, G. PAPER, AND A. VAHDAT. **Helios: a hybrid electrical/optical switch architecture for modular data centers**. *ACM SIGCOMM Computer Communication Review*, **41**(4):339–350, 2011. 7, 45
- [21] G. WANG, D.G. ANDERSEN, M. KAMINSKY, K. PAPAGIANNAKI, T. NG, M. KOZUCH, AND M. RYAN. **c-Through: Part-time optics in data centers**. *ACM SIGCOMM Computer Communication Review*, **41**(4):327–338, 2011. 7
- [22] C. CHEN, H. WANG, J. CHAN, AND K. BERGMAN. **A photonic interconnection network for hardware accelerator enabled utility computing**. pages 98–99, May 2013. 7, 59
- [23] K. CRAIGO, S. AND DUNN, P. EADS, L. HOCHSTEIN, D. KANG, M. KANG, K. MODIUM, D. AND SINGH, J. SUH, AND J.P. WALTERS. **Heterogeneous cloud computing**. In *Cluster Computing (CLUSTER), 2011 IEEE International Conference on*, pages 378–385. IEEE, 2011. 7, 56
- [24] JESSE BENSON, RYAN COFELL, CHRIS FRERICKS, CHEN-HAN HO, VENKATRAMAN GOVINDARAJU, TONY NOWATZKI, AND KARTHIKEYAN SANKARALINGAM. **Design, integration and implementation of the DySER hardware accelerator into OpenSPARC**. In *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*, pages 1–12. IEEE, 2012. 8
- [25] C. WILLARD, A. SNELL, L. SEGERVALL, AND M. FELDMAN. **HPC User Site Census: Systems**. *Intersect360 Reports*, March 2015. 9
- [26] V.W. ROSS. **Heterogeneous high performance computer**. In *Users Group Conference, 2005*, pages 304–307. IEEE, 2005. 9

REFERENCES

- [27] XILINX. **7 Series FPGAs Overview (DS180 v1.15)**. http://www.xilinx.com/support/documentation/data_sheets/ds180_7Series_Overview.pdf. Accessed:2015-04-4. 11
- [28] ALTERA. **Stratix V Device Overview (SV51001)**. http://www.altera.com/literature/hb/stratix-v/stx5_51001.pdf. Accessed:2015-04-4. 11
- [29] PCI-SIG. **PCI Express Base Specification**. <http://www.pcisig.com/specifications/pciexpress/base3/>, 2010. Accessed:2015-04-4. 13
- [30] A BACH. **The financial industrys race to zero latency and terabit networking**. *OFC/NFOEC Keynote presentation*, 2011. 13
- [31] J STUECHELI, B BLANER, CR JOHNS, AND MS SIEGEL. **CAPi: A Coherent Accelerator Processor Interface**. *IBM Journal of Research and Development*, **59**(1):7–1, 2015. 14, 19
- [32] D. BRUNINA, C.P. LAI, A.S. GARG, AND K. BERGMAN. **Building Data Centers with Optically Connected Memory**. *Journal of Optical Communications and Networking*, **3**(8):A40–A48, July 2011. 14
- [33] P. VAJGEL. **Needle in a haystack: efficient storage of billions of photos**. https://www.facebook.com/note.php?note_id=76191543919. Accessed:2015-03-31. 14
- [34] J.E. STONE, D. GOHARA, AND G. SHI. **OpenCL: A parallel programming standard for heterogeneous computing systems**. *Computing in science & engineering*, **12**(1-3):66–73, 2010. 15
- [35] CUDA NVIDIA. **Programming guide**, 2008. 15
- [36] LE. DAGUM AND R. MENON. **OpenMP: an industry standard API for shared-memory programming**. *Computational Science & Engineering, IEEE*, **5**(1):46–55, 1998. 15
- [37] J. AUERBACH, D.F. BACON, I. BURCEA, P. CHENG, S.J. FINK, R. RABBAH, AND S. SHUKLA. **A compiler and runtime for heterogeneous computing**. pages 271–276, June 2012. 15, 16, 17
- [38] J. BOUTELLIER, P. JAASKELAINEN, AND O. SILVEN. **Run-Time Scheduled Hardware Acceleration of MPEG-4 Video Decoding**. pages 1–4, November 2007. 15, 17
- [39] UNIVERSITY OF TORONTO. **LegUp**. <http://legup.eecg.utoronto.ca/>. Accessed:2015-04-3. 17
- [40] ANDREW CANIS, JONGSOK CHOI, MARK ALDHAM, VICTOR ZHANG, AHMED KAMMOONA, JASON H ANDERSON, STEPHEN BROWN, AND TOMASZ CZAJKOWSKI. **LegUp: high-level synthesis for FPGA-based processor/accelerator systems**. In *Proceedings of the 19th ACM/SIGDA international symposium on Field programmable gate arrays*, pages 33–36. ACM, 2011. 17
- [41] DAVID MILLER. **Device Requirements for Optical Interconnects to CMOS Silicon Chips**. In *Photonics in Switching*, page PMB3. Optical Society of America, 2010. 17
- [42] A.H. GNAUCK, RW TKACH, AR CHRAPLYVY, AND T LI. **High-capacity optical transmission systems**. *Journal of Lightwave Technology*, **26**(9):1032–1045, 2008. 17
- [43] O. LIBOIRON-LADOUCEUR, A. SHACHAM, B.A. SMALL, B.G LEE, H. WANG, C.P. LAI, A. BIBERMAN, AND K. BERGMAN. **The data vortex optical packet switched interconnection network**. *Journal of Lightwave Technology*, **26**(13):1777–1789, 2008. 18
- [44] R. LUIJTEN, W.E. DENZEL, R.R. GRZYBOWSKI, AND R. HEMENWAY. **Optical interconnection networks: The OSMOSIS project**. In *The 17th Annual Meeting of the IEEE Lasers and Electro-Optics Society*, 2004. 18
- [45] A. SHACHAM, H. WANG, AND K. BERGMAN. **Experimental demonstration of a complete SPINet optical packet switched interconnection network**. In *Optical Fiber Communication Conference*, page OThF7. Optical Society of America, 2007. 18
- [46] R.R. GRZYBOWSKI, R. HEMENWAY, M. SAUER, C. MINKENBERG, F. ABEL, P. MÜLLER, AND R. LUIJTEN. **The OSMOSIS optical packet switch for supercomputers: Enabling technologies and measured performance**. In *Photonics in Switching, 2007*, pages 21–22. IEEE, 2007. 18
- [47] A. BENNER, D.M. KUCHTA, P.K. PEPELJUGOSKI, R.A. BUDD, G. HOUGHAM, B. V. FASANO, K. MARSTON, H. BAGHERI, E.J. SEMINARO, H. XU, ET AL. **Optics for high-performance servers and supercomputers**. In *Optical Fiber Communication Conference*, page OTuH1. Optical Society of America, 2010. 18
- [48] B.J. OFFREIN AND P. PEPELJUGOSKI. **Optics in supercomputers**. In *Optical Communication, 2009. ECOC'09. 35th European Conference on*, pages 1–2. IEEE, 2009. 18
- [49] CAROLINE P LAI AND KEREN BERGMAN. **Broadband multicasting for wavelength-striped optical packets**. *Journal of Lightwave Technology*, **30**(11):1706–1718, 2012. 18, 24
- [50] BENJAMIN G LEE, BENJAMIN A SMALL, JUSTIN D FOSTER, KEREN BERGMAN, QIANFAN XU, AND MICHAEL LIPSON. **Demonstrated 4×4 Gbps silicon photonic integrated parallel electronic to WDM interface**. In *Optical Fiber Communication Conference*, page OTuM5. Optical Society of America, 2007. 18
- [51] CORNING. **Corning SMF-28e Optical Fiber Product Information**. <http://www.corning.com/docs/opticalfiber/pi1344.pdf>. Accessed:2015-04-2. 19, 29
- [52] NORIYUKI MIURA, YUSUKE KOIZUMI, EIICHI SASAKI, YASUHIRO TAKE, HIROKI MATSUTANI, TADAHIRO KURODA, HIDEHARU AMANO, RYUICHI SAKAMOTO, MITARO NAMIKI, KIMIYOSHI USAMI, ET AL. **A scalable 3D heterogeneous multi-core processor with inductive-coupling thruchip interface**. In *Cool Chips XVI (COOL Chips), 2013 IEEE*, pages 1–3. IEEE, 2013. 20
- [53] R. HO, K.W. MAI, AND M.A. HOROWITZ. **The future of wires**. *Proceedings of the IEEE*, **89**(4):490–504, 2001. 20
- [54] A. SHACHAM AND K. BERGMAN. **An Experimental Validation of a Wavelength-Striped, Packet Switched, Optical Interconnection Network**. *Journal of Lightwave Technology*, **27**(7):841–850, April 2009. 23

REFERENCES

- [55] XILINX. **XC2C32A CoolRunner-II CPLD**. http://www.xilinx.com/support/documentation/data_sheets/ds310.pdf. Accessed: 2014-04-9. 24
- [56] D. BRUNINA, C.P. LAI, A.S. GARG, AND K. BERGMAN. **Building data centers with optically connected memory**. *Journal of Optical Communications and Networking*, 3(8):A40–A48, 2011. 24, 44
- [57] H. WANG, A.S. GARG, K. BERGMAN, AND M. GLICK. **Design and demonstration of an all-optical hybrid packet and circuit switched network platform for next generation data centers**. In *Optical Fiber Communication Conference*, page OTuP3. Optical Society of America, 2010. 24
- [58] C.P. LAI, D. BRUNINA, AND K. BERGMAN. **Demonstration of 8×40 -Gb/s wavelength-striped packet switching in a multi-terabit capacity optical network test-bed**. In *2010 23rd Annual Meeting of the IEEE Photonics Society*, pages 688–689. Citeseer, 2010. 24, 45
- [59] G.N. ROUSKAS. **Optical layer multicast: rationale, building blocks, and challenges**. *Network, IEEE*, 17(1):60–65, 2003. 26
- [60] D. BRUNINA, C.P. LAI, AND K. BERGMAN. **A data rate-and modulation format-independent packet-switched optical network test-bed**. *Photonics Technology Letters, IEEE*, 24(5):377–379, 2012. 29
- [61] IEEE. **IEEE P802.3bm 40 Gb/s and 100 Gb/s Fiber Optic Task Force**. <http://www.ieee802.org/3/bm/index.html>. Accessed: 2014-04-7. 29
- [62] IEEE. **IEEE P802.3bs 400 Gb/s Ethernet Task Force**. <http://www.ieee802.org/3/bs/index.html>. Accessed: 2014-04-7. 29
- [63] W. ZHANG, A.S. GARG, H. WANG, C.P. LAI, J. WU, J. LIN, AND K. BERGMAN. **Experimental demonstration of 10 G-gigabit Ethernet-based optical interconnection network interface for large-scale computing systems**. In *Proc. IPC*, pages 443–444. Citeseer, 2011. 34
- [64] IEEE. **IEEE 802.3 'STANDARD FOR ETHERNET' MARKS 30 YEARS OF INNOVATION AND GLOBAL MARKET GROWTH**. http://standards.ieee.org/news/2013/802.3_30anniv.html. Accessed: 2015-04-6. 34
- [65] ALTERA. **Stratix II Device Handbook, Volume 1**. https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/hb/stx2/stratix2_handbook.pdf. Accessed: 2014-04-7. 35
- [66] VIDEOLAN. **ViduoLan Client**. www.videolan.org/vlc. Accessed: 2014-04-8. 36
- [67] J.W. LEE, R. FRANCESCANGELI, J. JANAK, S. SRINIVASAN, S.A. BASET, H. SCHULZTRINNE, Z. DESPOTOVIC, AND W. KELLERER. **NetSerV: active networking 2.0**. In *Communications Workshops (ICC), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011. 36
- [68] IEEE. **IEEE 802.16e Task Group (Mobile Wireless-MAN)**. <http://www.ieee802.org/16/tge/>. Accessed: 2015-04-5. 36
- [69] ITU. **ITU grid Channels (100 GHz Spacing)**. <http://www.telecomengineering.com/downloads/DWDM-100GHz.pdf>. Accessed: 2015-03-31. 39, 40, 50, 59
- [70] J.P. COSTAS. **Synchronous communications**. *Proceedings of the IRE*, 44(12):1713–1718, 1956. 45
- [71] J.A. KASH, A. BENNER, F.E. DOANY, D. KUCHTA, B.G. LEE, P. PEPELJUGOSKI, L. SCHARES, C. SCHOW, AND M. TAUBENBLATT. **Optical interconnects in future servers**. In *Optical Fiber Communication Conference*, page OWQ1. Optical Society of America, 2011. 45
- [72] P.K. PEPELJUGOSKI, J.A. KASH, F. DOANY, D.M. KUCHTA, L. SCHARES, C. SCHOW, M. TAUBENBLATT, B.J. OFFREIN, AND A. BENNER. **Low power and high density optical interconnects for future supercomputers**. In *Optical Fiber Communication Conference*, page OThX2. Optical Society of America, 2010. 45
- [73] M. NADA, M. NAKAMURA, AND H. MATSUZAKI. **25-Gbit/s burst-mode optical receiver using high-speed avalanche photodiode for 100-Gbit/s optical packet switching**. *Optics express*, 22(1):443–449, 2014. 46
- [74] J. RYLYAKOV, A. AND PROESEL, S. RYLOV, B. LEE, J. BULZACHELLI, A. ARDEY, B. PARKER, M. BEAKES, C. BAKS, C. SCHOW, ET AL. **22.1 A 25Gb/s burst-mode receiver for rapidly reconfigurable optical networks**. In *Solid-State Circuits Conference-(ISSCC), 2015 IEEE International*, pages 1–3. IEEE, 2015. 46
- [75] J. MANUEL D. MENDINUETA, J.E. MITCHELL, P. BAYVEL, AND B.C. THOMSEN. **Digital dual-rate burst-mode receiver for 10G and 1G coexistence in optical access networks**. *Optics express*, 19(15):14060–14066, 2011. 46
- [76] ALTERA. **Volume 4: Device Datasheet and Addendum Stratix IV Device Handbook**. https://www.altera.com/en_US/pdfs/literature/hb/stratix-iv/stx4_5v4.pdf. Accessed: 2014-04-9. 47
- [77] VERILOG. **Verilog Resources**. <http://www.verilog.com>. Accessed: 2014-04-8. 47
- [78] XILINX. **Efficient Shift Registers, LFSR Counters, and Long Pseudo-Random Sequence Generators**. http://www.xilinx.com/support/documentation/application_notes/xapp052.pdf, July 1996. Accessed: 2015-03-27. 52, 59
- [79] H. WANG AND K. BERGMAN. **A bidirectional 2×2 photonic network building-block for high-performance data centers**. *Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference*, pages 1–3, March 2011. 57
- [80] P. BARDELL, W. MCANNEY, AND J. SAVIR. *Built-In Test for VLSI: Pseudorandom Techniques*. John Wiley and Sons, New York, 1987. 63