A Photonic Interconnection Network for Hardware Accelerator Enabled Utility Computing

Cathy Chen[1], Howard Wang[1], Johnnie Chan[2], and Keren Bergman[1]
[1]*Department of Electrical Engineering, Columbia University, 1300 S. W. Mudd, 500 West 120th St., New York, New York 10027*
[2]*Department of Computer Science, Columbia University, 463 Computer Science Building, 1214 Amsterdam Ave., New York, New York 10027*
Email: cache@ee.columbia.edu

Abstract: High-bandwidth connectivity provided by WDM optical interconnects is an important enabler for delocalized hardware accelerators in utility computing. We validate a proposed architecture with an experiment that leverages optical interconnects to demonstrate error free (BER<10e-12) active switching and multicasting of FPGA generated and received packets.
OCIS codes: (200.4650) Optical interconnects; (060.0060) Optical Communications; (200.0200) Optics in Computing

1. Introduction

With the recent growth in cloud computing, the concept of utility computing (architecture model in which hardware resources are offered as on-demand services) has emerged as a new paradigm. 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 programing execution [1]. Known as hardware accelerators, these compute nodes 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. Graphics Processing Units (GPUs) and other forms of hardware acceleration are already being used commercially in the finance industry, and can achieve up to a 24x increase in performance, lower latency, and consumption of less power than systems without hardware acceleration [1].

Communication between the Central Processing Unit (CPU) and these hardware accelerators must be high bandwidth, low latency, and energy efficient [2]. 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 underutilization of these accelerators (can't access more accelerators than those *local* to it) [2]. *Delocalizing* the accelerators into an architecture with a central bank of accelerators would allow them to be dynamically allocated to different tasks. However, in order for this system to be successful, current electronic networks are inadequate. Optical interconnects offer the bandwidth, latency, and energy specifications necessary to support delocalized accelerators [3]. We propose a novel system architecture with an optically-connected bank of hardware accelerators. This specialized hardware would offer applications always-on, always-accessible acceleration for their specific computational tasks.

While this architecture would require additional hardware to manage the accelerator network, this hardware is minimal and does not introduce significant complexity to the design. Programming hardware accelerators is an additional hurdle, but with the development of languages like the Open Computing Language (OpenCL), which enables programs to be written in a way so that they will work across CPUs and GPUs, and IBM's Liquid Metal, a comprehensive compiler and run-time system that enables the use of a single language to program heterogeneous computing platforms, seamless co-execution of resultant programs on CPUs and accelerators is now possible [3, 4].

2. Multicasting in Hardware Accelerator Architectures

Multicasting data is an integral part of many hardware accelerator architectures. In the SPADE application for bargain discovery for example, Trade Quotes are multicasted 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 [2]. 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 [5]. In this data center, the original image could be multicasted to four separate jpeg hardware accelerators and encoded in parallel. An ideal architecture for hardware accelerators would be radically data parallel and enjoy direct access to main memory [2]. An optical network is capable of this data parallelism and has previously been shown to interface well with memory [3]. 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.

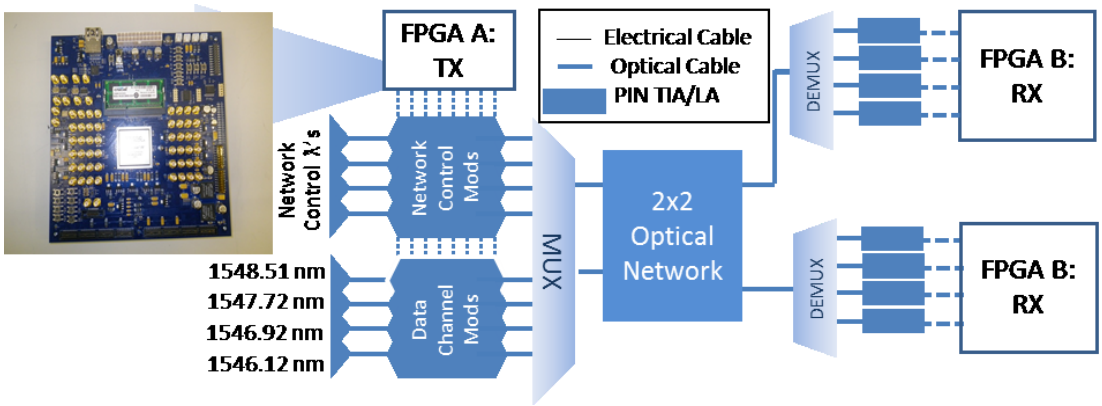

Fig. 1 - FPGA A modulates four payload channels and four network control wavelengths over a 2x2 actively switched network test-bed. FPGA B and the BERT receive these payloads from the optical network using four PIN-TIA receivers

3. Experimental Setup and Results

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 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 [Fig1.].

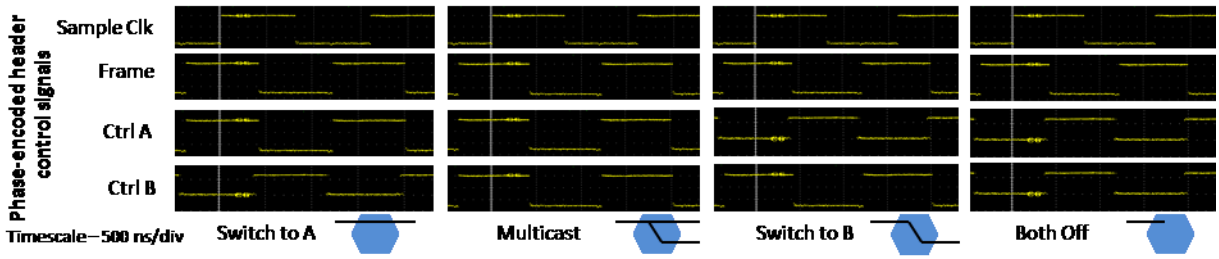

Fig. 2. Phase-encoded header network control- for a) switch to output A b) multicast c) switch to output B and d) both off

Due to various limitations of the PIN-TIA receivers in our test-bed, and the PLLs of the FPGA transceiver, we implement a phase-encoded header network control protocol. The output of each port is individually controlled using a bitwise XNOR of that port's control signal and a framing or reference signal [Fig. 2.]. To avoid issues of clock skew, this output control is calculated on the edge of a phase-shifted sample clock. This logic adds an additional wavelength to the network control, a minimal cost to the system. Each transceiver bank interfaces with optical components to generates and receive 4 x 11.3 Gbps Wavelength-Division Multiplexed (WDM) data transactions. Optical packets are sent through the network and shown to switch and multicast to the two receive nodes and shown to be error free (BER < 10e-12). Packet Routing is shown in Fig. 3.

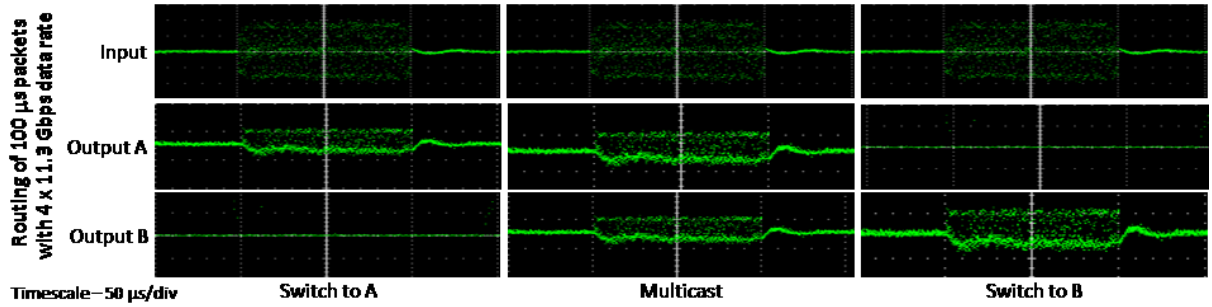

Fig. 3. Packet Routing- input and output Packets when a) switch to output A b) multicast c) switch to output B

4. References

[1] D.K. Yesalavich, "Switch to Videogame Chips Speeds Trading," *The Wall Street Journal,* April 30, 2010.
[2] S. Schneider *et al*, "Evaluation of Streaming Aggregation on Parallel Hardware Architectures" *DEBS '10,* pp. 248-257.
[3] D. Brunina *et al,* "Building Data Centers with Optically Connected Memory" JOCN **3** (8), A40-A48 (2011).
[4] J. Auerbach *et al*, "A Compiler and Runtime for Heterogeneous Computing," *DAC '12,* pp. 271-276.
[5] P. Vajgel *et al*, "Needle in a haystack: efficient storage of billions of photos" [Facebook Notes post]. https://www.facebook.com/note.php?note_id=76191543919 (April 2009).