ONIC: Optical Network Interface for Multi-Wavelength Interconnects in Ubiquitous CMP Computing Systems

Keren Bergman and Gary Carter
Advanced Computing Systems

HPC Systems

- Top-of-the-line
  - Cell, Opteron, Xeon
  - InfiniBand
- Highly specialized
  - Scientific computing
  - Financial analysis
  - Task parallelization

Data Centers

- Commodity
  - Celeron, Athlon
  - Ethernet
- General-purpose
  - Web/database apps
  - Virtualization
  - Cloud computing

Components

Applications

Users

Research/Government

Enterprise

ACS Interconnects Workshop - April 16, 2010
• The common bottleneck: the *interconnect*
  - Network bandwidth: a scarce resource
    - results in low server utilization
  - Power consumption: an even scarcer resource
  - Cost/Complexity (e.g. cabling), Location/Space
• Electronic technologies *will not scale*
Alleviating the Bottleneck

- Photonic technologies can enable scaling
  - Wavelength Division Multiplexing
    - Fiber bandwidth >> Cable bandwidth
      - Terabits vs Gigabits
    - Excess bandwidth traded off to simplify control
  - Bit-rate transparency
    - Switches operate at message rate, not bit rate
  - Low loss
    - Large bandwidth-distance product
    - No regeneration, pre-emphasis, equalization
    - Lower power density
      - Google: ~500 W/ft²
  - Photonic integration
    - Reduce power and cost
Network Requirements

HPC Systems

- Highly orchestrated
  - Complex, long-running algorithms
  - Extremely parallel and distributed
- Stringent latency requirements
  - Overhead of memory accesses limiting (1000s of clock cycles)

Data Centers

- Unpredictable
  - Bursty at the edge
  - Short, random messages
  - Long, extended flows
- Throughput-limited
  - Long flows account for majority of bandwidth
  - Localized hot-spots
  - Relaxed latencies

No “one-size fits all” network
We propose a unique end-to-end WDM optical network test bed platform featuring:

- A configurable hybrid photonic network building-block
  - Optical/electronic switching
  - Packet/circuit traffic support
- A network-agnostic optical network interface
  - Support for standard protocols (e.g. InfiniBand, etc.)
- Support for optically-connected memories
Photonic Network Test Bed

The test bed will feature flexibility and scalability to evaluate various topologies/configurations

- Reprogrammability to support various architectures and protocols
- Configurable levels of hybrid electronic/optical switching
- All-optical packet switching
- Optical circuit switching
- Wavelength-division multiplexed (WDM) data streams
- Centralized, distributed, and/or hierarchical control
All-Optical Hybrid Switch

- Simultaneous Circuits and Packets
  - Traffic adaptability
  - Resource reallocation
- 4 \times 4 wideband support
- \lambda\text{-striped messages}
  - Simplicity
  - Bandwidth: Tb/s
- Programmable/Reconfigurable
  - FPGA-based control scheme
- Building block for data center networks
  - Provide max BW at the top of the DC hierarchy
Architecture and Packet Structure

HEADERS

PAYLOAD

Payload

Control Logic

Power splitter

wavelength filter

receiver

optical fiber

electronic trace/cable

in

out

Time

λ

UMBC
Architecture and Packet Structure

- **Power Splitter**
- **Wavelength Filter**
- **SOA**
- **Receiver**
- **Optical Fiber**
- **Electronic Trace/Cable**

**Headers**

**Payload**

**Circuit or Packet**

**Control Logic**

ACS Interconnects Workshop - April 16, 2010
Architecture and Packet Structure

ACS Interconnects Workshop  April 16, 2010
Implementation

- 4 sub-modules
  - 1 per input port
  - 4 SOAs
  - Xilinx CPLD
  - PIN-TIA-LA
Implementation

- Each sub-module centrally managed
  - Xilinx Virtex 5-based board
  - Manages output contentions
    - Prioritizes circuits over packets
  - Port configuration
    - Packet or Circuit
    - Configured via Ethernet
    - Managed via Windows GUI
Demonstration

- All-optical packet routing
  - $4 \times 10$ Gb/s $\Lambda$-striped packets
  - $\text{in}_0$ to $\text{out}_0/\text{out}_1$
- Concurrent circuit generation
  - $10$ Gb/s stream
  - $\text{in}_3$ to $\text{out}_3$
- Error-free $10^{-12}$
Photonic Network Test Bed

• Modular 4 × 4 switching node design
  ▪ Functionality
    o Network protocol support (Ethernet, etc.)
    o Circuit & Packet Switching
  ▪ Flexibility
    ▪ 1 × 2, 2 × 2, or 4 × 4
  ▪ Robustness
    o status monitoring
    o debugging functionality
  ▪ Currently Under Development
The Interface Challenge

• Interface Optical Packet-Switched Networks to Real-World Systems
  ▪ Non-Trivial Hurdle for Optical Networks & I/O Protocols

• Need a **seamless** interface between nodes and network
  ▪ Support standard protocols
  ▪ Network agnostic
  ▪ WDM
  ▪ Resilient
  ▪ Low Latency
**O-NIC Design Vision**

- **Virtual Switch Model**
  - Local Node Interacts with High-Radix ‘Virtual Switch’
  - O-NIC Performs Address & Flow Control Translation
Comparison with Standard I/O

- Optical Switching Occurs in Physical Layer
  - Requires Links to be Made & Broken Rapidly
  - Continuous Link-Based Protocols Assume Transport Layer Switching
  - Clock & Data Recovery Must be Restarted with Every Switch in Destination
  - Infiniband Link Training
    - 100ms + at 2.5 Gbps, 200ms + at 5/10 Gbps per lane
  - Current O-NIC Design < 2ms at 2.5 Gbps per lane
Packet Traversal

- This configuration of our optical interface is of a 2-node prototype currently under construction
Packet Traversal

- Typical Layers of the Protocol Stack Running on the Nodes
- Similar Design Across Standards
Packet Traversal

- Command & Data in App Layer Flows Down NIC Protocol Stack
- Transport & Link Layers Add Protocol Header & Tail
- Physical Layer Slices Packet for Multiple Transmitters
- Physical Layer Inserts Packet into Transmitter Stream
Packet Traversal

- Continuous Link
- In Idle, PRBS is Sent
  - Maintains Clock Recovery, DC Balance
- Packet is Inserted into PRBS Stream
Packet Traversal

- Packet Flows Up Protocol Stack to Virtual Switch
Packet Traversal

- Virtual Switch Sends Packet to Optical Network Side of O-NIC
- Optical Network Arbitration Layer
  - Packet Aggregation
  - Optical Injection Arbitration
  - Switch Emulator Interface
Packet Traversal

- Optical Network Transport Layer
  - Address Translation Table
  - Network Ack / Nack
  - Network Timeouts
  - Network Flow Control
Packet Traversal

- Optical Network Packet Encoder
  - Creation of Optical Packet
  - Consumption of Optical Packet
Packet Traversal

- Optical Physical Layer
  - Optical Transceivers
  - Electronic Transceivers
O-NIC Packet Format

- HEADERS
  - Header Sequence
  - Payload
  - Tail

- PAYLOAD
  - Data
  - Skip

GUARD TIME

For CDR Unit

\[ \ldots, \text{COMMA, D10.2 \times15, } \ldots \]
O-NIC Packet Format

HEADERS

PAYLOAD

DATA

SKIP

For CDR Unit

..., COMMA, D10.2 ×15, ...

GUARD TIME

Header Sequence

Payload

Idle

Idle

Time

λ

OCS Interconnects Workshop ■ April 16, 2010
O-NIC Logic Design

INFINIBAND LINK LAYER

FPGA

OPTICAL PACKET ENCODER

LOOKUP TABLE

HEADER GENERATOR

OPTICAL PACKET DECODER

INFINIBAND INTERCONNECTS WORKSHOP

ACS Interconnects Workshop  April 16, 2010

32
O-NIC Logic Design

![Diagram of O-NIC Logic Design]

- **O-NIC Logic Design**
- **Optical Packet Encoder**
- **Optical Packet Decoder**
- **Lookup Table**
- **Header Generator**
- **PCI Express Link**
- **FPGA**

ACS Interconnects Workshop ▪ April 16, 2010
O-NIC Logic Design Test Mode

IB Pkt Gen

Optical Packet Encoder

Lookup Table

Header Generator

Optical Packet Decoder

IB Pkt Checker

Error Counter

FPGA

ACS Interconnects Workshop • April 16, 2010

UMBC
O-NIC : Electronic Hardware

- 4 Stratix II GX FPGAs
  - 14 Transceivers at 5.0 Gbps
- Infiniband & PCIe Adapters
UMBC Setup

Modulators

FPGA

Receivers

ACS Interconnects Workshop  April 16, 2010
Electronic Memory

- Devices share command and address lines for parallel access
- Many wires/pins
- Path-length matched for skew
- Bus clocked at higher rate than memory devices

- Extremely complex wiring
- High energy dissipation
- Large pin-out
- Does not scale with performance, capacity, or physical distance
Optically-Connected Memory

- Scales with performance, capacity, and physical distance
  - Bit-rate transparency
  - Distance immunity at computer-scale
  - Large bandwidth-distance product
- Simplified wiring
  - Board space near processor now available for other uses
Accomplishments

- Demonstration of circuit-switched microprocessor/SDRAM communication over 4×4 Optical Test-Bed
  - CPU is emulated on a high-speed FPGA
- All communication between CPU and Memory Controller is electronic
- All communication between Memory Controller and SDRAM is optical
Advantages of Optical Interface

• Decoupled energy-distance relationship
  ▪ SDRAM can be arbitrarily distant
  ▪ Entirely new design space for computer systems
• No long traces to drive
  ▪ Less power
  ▪ Higher bandwidth, lower latency
• Less pins on DIMM module and going into chip
  ▪ Waveguides can achieve dramatically higher density due to WDM
• Ideal for scaling to large number of SDRAM devices
  ▪ Memory architectures with higher capacity and greater bandwidth than electronics can provide
Next Steps

• Optical Network Interface Card
  ▪ Reduce CDR Time
    o Optimize CDR Settings
    o DC-Coupling of All Transceiver Elements
    o Explore Alternate Clock Recovery Options
  ▪ Network-Level Operations
    o E.g. Port Discovery, Flow Control

• Memory Interface
  ▪ Optimize Link for Circuit Switching
    o Currently Using Tweaked O-NIC Burst-Mode Design
  ▪ Switch to 5 Gbps per lane with 8 lanes
  ▪ Build Packet-Switched Variation
Questions?

Visit Us At

http://lightwave.ee.columbia.edu