OPTICAL INTERCONNECTS FOR EXASCALE SYSTEMS

Moray McLaren
HP Labs
13th October 2010
TALK OUTLINE

- Characteristics of optical interconnects
- CMOS nanophotonics and high radix routers
- Exascale Systems
PHOTONICS AS A DISRUPTIVE TECHNOLOGY
OPTICAL INTERCONNECTS
DISRUPTIVE TECHNOLOGY OR...

....GOLD PLATED PLUMBING?
POTENTIALLY DISRUPTIVE CHARACTERISTICS

- Freespace capability
- Broadcast
- Circuit switching
- Distance independence
- Power efficiency
- Bandwidth density
- EMI immunity
FREESPACE

- Board to board interconnect using 4x4 VCSEL arrays
- Telecentric lenses allow for misalignment
- No fibers..
- Disadvantages
  - Range limited by divergence
  - Very specific to packaging solution
  - Hard to scale with WDM
BROADCAST
Broadcast based optical fabric

Optical backplane demonstrator
  • Replace CMOS fabric ASICS with multiple broadcast buses
  • Very low cost plastic waveguides
  • Scalable to 16 line cards

Advantages
  • Low power
  • Simple passive backplane
  • Upgradeable through CWDM

Broadcast for Exascale?
  • Potential application to memory buses
  • (but bandwidth, not capacity is the problem)
OPTICAL SWITCHING

– “I’ve got a 10ps optical switch, why can’t you use it”
– “It’s only a bit of logic to turn a circuit switch into a packet switch isn’t it?”
– Many possible implementations
  • MEMS, ring resonators
– Advantages
  • Very low power operation
  • Very low through latency
– Disadvantages
  • It’s not a packet switch...
– Successful Telco use model – resilience and provisioning

Glimmerglass MEMS optical switch
# DISTANCE INDEPENDENCE

## ELECTRONIC HIERARCHY

- Short on chip  <3mm
- Global on hip  <20mm
- Local off chip  <200mm
- Chassis level  <1000mm
- Cabled        <6m
- Active cable   <12m
- Active optical <100m

## OPTICAL HIERARCHY

- Short on chip  <3mm
- Global on chip  <20mm
- Photonic       <100m
CHOKE POINTS TODAY

At the enclosure boundary
  • Screened cables

At the card edge
  • Connector density
  • Crosstalk limitations

At chip edge
  • Pin count limitation < 6000
  • Pin data rate limitation

OPTICAL BANDWIDTH SCALING

– Multiple wavelengths..

– Multiple fibers

– Multiple cores..
INTEGRATED PHOTONICS

Only way to fully exploit bandwidth density

- The 2000 telecom bubble based on discrete opt
  - Components are measured in mm
  - Hand alignment
  - Expensive and not scalable

- Integrated nanophotonics
  - Manufacture many thousands per die
  - Advances in lithography -> better devices

- Current generation CMOS photonics
  - Relatively large MZ modulators
  - No power advantage over VCSELs
  - Limited WDM capability

- Use resonant devices....

Source: Newport Corp., Assembly Magazine, September 2001
SI MICRORINGS

– Example: 5 cascaded microring resonators, slightly different radii ~ 1.5 mm.

– High Q of 9,000 (BW ~ 20 GHz) and high extinction ratio of 16 dB.

RING RESONATORS

One basic structure, 3 applications

- A modulator – move in and out of resonance to modulate light on adjacent waveguide
- A switch – transfers light between waveguides only when the resonator is tuned
- A wavelength specific detector - add a doped junction to perform the receive function
- 10 μm silicon ring resonators
  - Charge injection
  - 1310 nm (compatible with Ge detectors)

- Experimental Results
  - Q ~ 10,000
  - 0.18 nm shift
  - 18 dB extinction
  - 3 Gbps modulation
  - 54 fJ/bit

Cross section

PECVD Oxide cladding

Si

BOX

Si Substrate <100>

Top view

Eye diagram RZ 3 Gbps

Normalized intensity

Wavelength (nm)
**DWDM POINT TO POINT LINK**

- **LASER SOURCE (shared)**
  - unmodulated light
  - 33 lambdas

- **splitter**
  - to other channels
  - Splitter loss 0.1 dB per binary stage
  - Array of 33 modulator rings (8.25μm x 215μm)

- **Single mode fiber**
  - 10μm mode diameter
  - Delay 5ns/m
  - Loss 0.4dB/km

- **Silicon ridge waveguide**
  - 0.5μm wide 4.5μm pitch
  - Delay 118ps/cm
  - Loss 0.3 - 1dB/cm

- **Fiber coupler if going off chip**
  - Loss 1dB per connection

- **Array of 33 detector rings**
TECHNICAL CHALLENGES - TUNING

– Tuning is necessary to:
  • Compensate for fabrication variations
  • Correct for temperature variations

– Thermal tuning
  • Simple to implement
  • Minimize tuned thermal mass

– Alternate approaches
  • Self compensated rings
  • Still need to tune for process variation

M. Watts et. al. Sandia National Lab
COMPUTERS WITH LARGE NUMBERS OF INTEGRATED HEATERS HAVE BEEN BUILT....

Colossus Mark 2 computer
POINT-TO-POINT POWER BUDGET

- 10Gbit/s per wavelength
- 177fJ/bit assuming 32nm process
- No clock recovery and latching - not directly comparable to electronic numbers
- Idle link still needs laser and tuning power
STATE OF THE ART ELECTRONIC ROUTER

- Total chip IO limited by package constraints
- Increasing port data rates only possible by reducing port count
- Overall chip power dominated by SerDes and IO power
- Pin data rates limited by power and signal integrity
- Limited cable length, (<6m today, decreasing with data rate)

Electronic switch core - buffering, routing, and switching
Electronic SerDes & IO
INTEGRATED PHOTONIC IO ROUTER

- High bandwidth density at chip edge due to WDM fiber interconnect
- Greatly reduced IO power
- Electronic switch core unchanged. Buffering, routing and switch functions still implemented electronically
- Unlimited cable length
- Scale port bandwidth by adding wavelengths

Off chip optical power source

Electronic switch core - buffering, routing, and switching

Optical IO. One input and output fiber pair per port
INTERGRATED PHOTONIC IO & CORE

- Optical switch core eliminates long electronic wires, enables full crossbar, reducing power
- Chip area dominated by electronic input buffering
- Area (&cost) scales ~linearly with port count
- Folding possible for large switches
- Switch arbitration may be optical or electronic
## ROUTER IO POWER SCALING

IO power in watts for 64, 100 and 144 port switches

<table>
<thead>
<tr>
<th>Generation</th>
<th>Port BW(Gbps)</th>
<th>IO type</th>
<th>fJ/bit</th>
<th>64</th>
<th>100</th>
<th>144</th>
</tr>
</thead>
<tbody>
<tr>
<td>45nm</td>
<td>80</td>
<td>Electronic</td>
<td>7000</td>
<td>35.8W</td>
<td>56.0W</td>
<td>80.6W</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>Optical</td>
<td>451</td>
<td>2.3W</td>
<td>3.6W</td>
<td>5.2W</td>
</tr>
<tr>
<td>35nm</td>
<td>160</td>
<td>Electronic</td>
<td>5048</td>
<td>51.7W</td>
<td>80.8W</td>
<td>116.3W</td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>Optical</td>
<td>284</td>
<td>2.9W</td>
<td>4.5W</td>
<td>6.5W</td>
</tr>
<tr>
<td>22nm</td>
<td>320</td>
<td>Electronic</td>
<td>4049</td>
<td>82.9W</td>
<td>129.6W</td>
<td>186.6W</td>
</tr>
<tr>
<td></td>
<td>320</td>
<td>Optical</td>
<td>191</td>
<td>3.9W</td>
<td>6.1W</td>
<td>8.8W</td>
</tr>
</tbody>
</table>

- Assumes each generation will require a doubling of port speed to match improvements in processor performance.
- Total device power must be less than 130W for normal forced air cooling with heatsink, 200W possible with special cooling. Assume max power budget of 50% (65W) for IO.
- Conservative modulation rate enables power efficient DDR clocking (data rates is 2x system clock rate).
**ROUTER CORE POWER SCALING**

**CORE POWER = CHIP POWER – IO POWER**

<table>
<thead>
<tr>
<th>Generation</th>
<th>Port BW</th>
<th>core type</th>
<th>64</th>
<th>100</th>
<th>144</th>
</tr>
</thead>
<tbody>
<tr>
<td>45nm</td>
<td>80Gbps</td>
<td>Electronic</td>
<td>41.8W</td>
<td>72.7W</td>
<td>120.7W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optical</td>
<td>13.2W</td>
<td>17.4W</td>
<td>31.9W</td>
</tr>
<tr>
<td>35nm</td>
<td>160Gbps</td>
<td>Electronic</td>
<td>38.0W</td>
<td>65.9W</td>
<td>109.0W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optical</td>
<td>22.9W</td>
<td>27.7W</td>
<td>50.9W</td>
</tr>
<tr>
<td>22nm</td>
<td>320Gbps</td>
<td>Electronic</td>
<td>52.4W</td>
<td>91.9W</td>
<td>153.8W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optical</td>
<td>34.2W</td>
<td>41.3W</td>
<td>76.3W</td>
</tr>
</tbody>
</table>

- Core speed doubles with each generation to match improvements in processor performance.
- Total device power must be less than 130W for normal forced air cooling with heatsink. IO power negligible with photonic IO.
- Photonic switch core allows low power operation at high port counts and high bandwidth.
- Router chip area (&cost) scales linearly with port count for optical switch core case.
PHOTONIC ROUTER

- Pair of fibers per port
- Buffering and routing logic in CMOS
- OEOEO structure with electronic buffering at inputs and outputs shown
- OEO switches possible with no buffering at outputs.
- Fully non-blocking crossbar
- Clustering to reduce number of rings in standby. Multiple ports share single modulator block in optical core (2-way shown)
- Optical or electronic arbitration
  
  **Optical** simple, fast, low-power
  
  **Electronic** allows more complex arbitration algorithms, age-based, QoS etc.

Need to evaluate trade-off at system level
EXASCALE INTERCONNECT CHALLENGES

Subtitle Placeholder Goes Here

- Topology choices
  - Cabling complexity
  - Network diameter

- Enclosure level deployment

- Cost trade-off in global bisection bandwidth
  - What size of domain is it desirable to maintain full bisection bandwidth?

- Central switches
  - Impractical at Exascale level

- Power
  - Proportionality
  - Ability to power down unused resources
ENCLOSURE LEVEL CONNECTIVITY

- Blade card
  - 16, ~10Tflops processors
  - 96 port packet switch
  - 2, 160Gbyte/s links per processor

- 10TB/s blade to backplane bandwidth
  - 128 fibers, 32mm at 250um pitch
  - Bandwidth constrained by power (&cost) not card-edge

- 32 blade backplane
  - 32 way all to all wiring between blades within enclosure
  - Consolidate links to multi-fiber connectors

- External connectivity
  - 160TB/s external bandwidth
  - 32, 64 fiber connections
OVERALL SYSTEM

- 256 enclosures
- No central switches
- Flattened butterfly/HyperX connectivity
- 5/10 Petabytes/s bisection bandwidth
- Replicated fiber wiring
- Wide range of link lengths

(S₂ = 16) way all to all L2 wiring x 2
(S₃ = 16), way all to all L3 wiring x 2

Maximum bisection 64 cables per all-to-all

(S₂ x S₃ = 16 x 16) array of (T x S₁ = 512) processor racks
MIXED PACKET & CIRCUIT SWITCHING?

- Use circuit switches to configure connectivity between enclosures
- Modify circuit switches for changes in usage patterns and response to hardware failures (Telco model)
- Reduces power by minimizing number of times packets are actively switched
- Increase application bandwidth by minimizing deroutes