The OptoHPC simulator: Bringing OptoBoards to HPC-scale environments

Pavlos Maniotis, Nikos Terzenidis, Nikos Pleros
Aristotle University of Thessaloniki (AUTH), Greece

OMNeT++ Community Summit 2016
15 September 2016, Brno, Czech Republic
Outline

- Introduction
- The OptoHPC simulator architecture
- An OptoHPC use case: comparison performance analysis using the OptoHPC
- Conclusion
Motivation

Data Movement is the Bottleneck to Performance, Not Flops

Source: Al Geist in “Paving the Roadmap to Exascale”, SciDAC Review 2010

TH2
Located in China

Ranked as the world’s fastest supercomputer (Nov. 2015)
- 33.9 PFLOPS
- 17.6 MW

* has only reached 4% of the exascale target (set for ~2020-2025)
* has already reached 89% of the 20 MW power limit target *


The OptoHPC simulator
Motivation

Data Movement is the Bottleneck to Performance, Not Flops

*Source: Al Geist in “Paving the Roadmap to Exascale”, SciDAC Review 2010*

**Challenges and the role of Optical interconnects**

- As computation density increases (more cores/chip) leads to higher capacity requirements…
- …but Copper wires have significant limitations as:
  - they can offer High capacity only for very short distances
  - they present increased power consumption as speed and distance increases

- Optical interconnects emerge as a promising solution for replacing copper at short distances in future DC and HPC systems
  - they can offer High capacity for both short and higher distances combined with low power consumption

Optical Interconnects Evolution & RoadMap

~2010
Electrical system, optical fibers at card edge only

~2011
Optical fibers across the boards

Today
Optical waveguides in/on boards

~2020
Optical interconnects integrated with the processor

Optical Interconnects Evolution & RoadMap

~2010
- Electrical system, optical fibers at cards

~2011
- Optical fibers across the boards

Today
- Optical waveguides in/on boards

~2020
- Optical interconnects integrated with the processor

Active Optical Cables
On-board subassemblies
Optical PCBs
Optical Network-on-chip

PhoxTroT deals with optical:
(1) On-board, (2) Board to board and (3) Rack to Rack interconnects
The PhoxTroT Research Project & its Vision

PhoxTroT deals with optical:
(1) On-board,
(2) Board-to-board, and
(3) Rack-to-Rack interconnects

How do all these technology improvements will affect the system-scale performance of an HPC?

Opto-HPC is an OMNeT++ based simulator that targets in simulating complete HPC network systems that make use of PhoxTroT technologies (and generally optical technologies)
titanStyleNetwork network module:
- Defines the connections among the HPC racks and declares the use of the (a) statisticsManager, (b) networkAddressesManager and (c) trafficPatternsManager simple modules
- Can be configured to any 3D Torus and Mesh network desired size
The Opto-HPC simulator

statisticsManager simple module:
- Responsible for collecting the global statistics
The Opto-HPC simulator

networkAddressesManager simple module:
- Responsible for addresses allocation to network’s nodes and routers (for both decimal and XYZ addresses)
- Responsible for defining the dateline routers that are necessary for resolving Deadlocks in Torus networks
The Opto-HPC simulator

trafficPatternsManager simple module: Responsible for defining and managing the applications running on the HPC

10 available options:
1) Random Uniform
2) Bit Complement
3) Bit Reverse
4) Bit Rotation
5) Shuffle
6) Transpose
7) Tornado
8) Neighbor
9) User defined statistical distributions
10) Packet traces
cabinet compound module:
- Defines the connections among the chassis placed in the cabinet and the outer world
The Opto-HPC simulator

chassis compound module:
- Defines the connections among the PCBs placed in the cabinet and the outer world
The Opto-HPC simulator

PCB compound module:
- Defines the connections among the nodes and routers inside the PCB and the outer world
The Opto-HPC simulator

Node compound module:
- Represents the CPU chips used in the HPC
- Embodies all the key simple modules for having “cpu operation”

Router compound module:
- Represents the router chips used in the HPC
- Embodies all the key simple modules for having “router operation”
- Supports DOR and minimal Valiant routing algorithms
- Utilizes 3 auxiliary classes:
  1) shortestPathsManager
  2) routingTableManager
  3) routingManager
Buffer simple module:
- Implements FIFO queue buffering for the incoming data
- Separated in Virtual Buffers in order to avoid warp-around link deadlocks
The OptoHPC simulator has a simple module named `resourcesManager` which is responsible for:
- the router resources allocation (output ports)
- sending credit packets to the previous nodes/routers

It utilizes 3 auxiliary classes:
1. `pendingDataManager`
2. `gateAllocationManager`
3. `creditManager`
The OptoHPC simulator

switchFabric simple module:
Forwards the data transmitted by the buffers/resourcesManager to the proper output port
The Opto-HPC simulator

trafficGenerator simple module:
Responsible for:
- Creating the node’s data according to the running application
- Sinking the incoming data from network
- Forwarding credit packets to the buffer

Utilizes 2 auxiliary classes:
1) nodeMessagesManager
2) nodeStatisticsManager
# Stats for Nerds

## 6 Compound Modules
1. titanStyleNetwork.ned
2. cabinet.ned
3. chassis.ned
4. pcb.ned
5. node.ned
6. router.ned
(5 & 6 implement also C++ classes)

## 7 Simple Modules
1. networkAddressesManager.ned
2. trafficPatternsManager.ned
3. statisticsManager.ned
4. trafficGenerator.ned
5. buffer.ned
6. resourcesManager.ned
7. switchFabric.ned

## 5 msg definitions
1. bufferTimer.msg
2. resourcesManagerTimer.msg
3. data.msg
4. flit.msg
5. credit.msg

## C++ code
1. 23 new C++ class definitions
2. a total of ~8000 lines of C++ code
3. $O(n^2)$ complexity for the Dijkstra algorithm
4. $O(1)$ complexity for all the major functions (routing decisions, traffic generation etc…)
An OptoHPC use case: Titan CRAY XK7 blade vs OPCB

World’s #2 HPC

The OptoHPC simulator
An **OptoHPC** use case: Titan CRAY XK7 blade vs OPCB

The **OptoHPC** simulator

*Siokis A. et al. “Laying out Interconnects on Optical Printed Circuit Boards”*
### An OptoHPC use case: Titan CRAY XK7 blade vs OPCB

#### Comparison Table

<table>
<thead>
<tr>
<th>Router Port Type</th>
<th>Conventional Router</th>
<th>OE-Router-88ch *</th>
<th>OE-Router-168ch *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node-Router (Gbps)</td>
<td>83.2</td>
<td>64</td>
<td>120</td>
</tr>
<tr>
<td>X dimension (Gbps)</td>
<td>75</td>
<td>64</td>
<td>120</td>
</tr>
<tr>
<td>Y dimension (Gbps)</td>
<td>75 (Mezzanine)</td>
<td>96</td>
<td>192</td>
</tr>
<tr>
<td></td>
<td>37.5 (Cable)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z dimension (Gbps)</td>
<td>120 (Backplane)</td>
<td>128</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>75 (Cable)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max Capacity (Tbps)</td>
<td>0.706</td>
<td>0.704</td>
<td>1.344</td>
</tr>
</tbody>
</table>

*Siokis A. et al. “Laying out Interconnects on Optical Printed Circuit Boards”*
Performance Analysis Results – CRAY XK7 for both DOR & MOVR

- DOR ~20% better
- DOR ~15% better
Performance Analysis Results

The OptoHPC simulator
## Performance Analysis Results

### Mean node Throughput Results

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Conventional Router (Gbps)</th>
<th>OE-Router-88ch (Gbps)</th>
<th>OE-Router-168ch (Gbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform Random</td>
<td>14.28</td>
<td>48 (3.36x)</td>
<td>92 (6.44x)</td>
</tr>
<tr>
<td>Bit Rotation</td>
<td>20.2</td>
<td>27.2 (1.34x)</td>
<td>51.46 (2.54x)</td>
</tr>
<tr>
<td>Bit Complement</td>
<td>11.7</td>
<td>23.67 (2.02x)</td>
<td>48 (4.10x)</td>
</tr>
<tr>
<td>Bit Reverse</td>
<td>12</td>
<td>17 (1.41x)</td>
<td>32.8 (2.73x)</td>
</tr>
<tr>
<td>Shuffle</td>
<td>17.4</td>
<td>19.25 (1.10x)</td>
<td>36.43 (2.09x)</td>
</tr>
<tr>
<td>Tornado</td>
<td>5.23</td>
<td>11.51 (2.20x)</td>
<td>24 (4.58x)</td>
</tr>
<tr>
<td>Transpose</td>
<td>15.45</td>
<td>21.63 (1.40x)</td>
<td>41.76 (2.70x)</td>
</tr>
<tr>
<td>Nearest Neighbour</td>
<td>36</td>
<td>30.7 (0.85x)</td>
<td>57.6 (1.60x)</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>~16.5</td>
<td>~24.9 (1.5x)</td>
<td>~48 (2.90x)</td>
</tr>
</tbody>
</table>
**Conclusions**

Successfully developed a queue-based simulator for complete HPC systems

- Offers support for both electrical and optical components
- Currently supports 3D Torus and Mesh Topologies
- Supports 8 synthetic traffic patterns as well as user-defined statistical distributions and trace files
- Features both SF and VCT operation like most state-of-the-art routers in the market
- Implements DOR and Minimal Oblivious Valiant Algorithms (with VC support) allowing for deadlock free operation
- Comparison between Conventional & O/E technologies using OptoHPC has shown 1.5x mean higher throughput for 88ch. case, 2.9x mean higher throughput for 168ch. case
Thank you for your attention!