How co-designs and high level predictions may help matching new technology trends

Philippe Thierry, Principal Engineer, Codesign and Pathfinding,
Gabriele Paciucci, HPC Solutions Architect, Scalable Datacenter Solutions
Intel
June 22, 2017

Agenda

• Intel OPA current status
• Challenges looking forward
• Co design & Application point of view
Legal Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL’S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A “Mission Critical Application” is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL’S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS, COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS’ FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The cost reduction scenarios described in this document are intended to enable you to get a better understanding of how the purchase of a given Intel product, combined with a number of situation-specific variables, might affect your future cost and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to: Learn About Intel® Processor Numbers

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

The High-Performance Linpack (HPL) benchmark is used in the Intel® FastFabrics toolset included in the Intel® Fabric Suite. The HPL product includes software developed at the University of Tennessee, Knoxville, Innovative Computing Libraries.

Intel, Intel Xeon, Intel Xeon Phi™ are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States or other countries.

Copyright © 2015, Intel Corporation
CPU-Fabric Integration
with the Intel® Omni-Path Architecture

Future Generations
Additional integration, improvements, and features

Next generation Intel® Xeon® Phi™ coprocessor
Future Intel® Xeon® processor
Intel® Xeon Phi™ processor 7200
Intel® Xeon® processor E5-2600 v4
Intel® Xeon® processor E5-2600 v3

KEY VALUE VECTORS
✓ Performance
✓ Density
✓ Cost
✓ Power
✓ Reliability

INTEGRATION

TIME

Intel® OPA HFI Card

Tighter Integration
Multi-chip Package Integration
## Intel® Omni-Path Architecture

<table>
<thead>
<tr>
<th>HFI Adapters</th>
<th>Edge Switches</th>
<th>Director Switches</th>
<th>Silicon</th>
<th>Software</th>
<th>Cables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single port</td>
<td>1U Form Factor</td>
<td>QSFP-based</td>
<td>OEM custom designs</td>
<td>Open Source</td>
<td>Third Party Vendors</td>
</tr>
<tr>
<td>x8 and x16</td>
<td>24 and 48 port</td>
<td></td>
<td>HFI and Switch ASICs</td>
<td>Host Software and Fabric Manager</td>
<td>Passive Copper Active Optical</td>
</tr>
<tr>
<td>x16 Adapter (100 Gb/s)</td>
<td>48-port Edge Switch</td>
<td>768-port Director Switch (20U chassis)</td>
<td>HFI silicon Up to 2 ports (50 GB/s total b/w)</td>
<td>Cables</td>
<td></td>
</tr>
<tr>
<td>x8 Adapter (58 Gb/s)</td>
<td>24-port Edge Switch</td>
<td>192-port Director Switch (7U chassis)</td>
<td>Switch silicon up to 48 ports (1200 GB/s total b/w)</td>
<td>Third Party Vendors</td>
<td></td>
</tr>
</tbody>
</table>

### Description

| Traffic Flow Optimization | “Quality of Service“: Transmission of lower-priority packets can be paused so higher priority packets can be transmitted | Ensures high priority traffic is not delayed (Faster time to solution) |
| Packet Integrity Protection | * Allows for rapid and transparent recovery of transmission errors on an Intel® OPA link without additional latency | Much lower latency than Forward Error Correction (FEC) defined in the InfiniBand® specification<sup>1</sup> |
| Dynamic Lane Scaling | * Maintain link continuity in the event of a failure of one of more physical lanes (Operates with the remaining lanes) | Enables a workload to continue to completion. Note: InfiniBand will shut down the entire link in the event of a physical lane failure |

---

<sup>1</sup> Source: Intel internal information. Design win count based on OEM and HPC storage vendors who are planning to offer either Intel-branded or custom switch products, along with the total number of OEM platforms that are currently planned to support custom and/or standard Intel® OPA adapters. Design win count as of November 1, 2015 and subject to change without notice based on vendor product plans. Other names and brands may be claimed as property of others.
Interconnect at scale

High performance and efficient communications
- Low latency – low-overhead small messages
- High bandwidth – high-efficiency for large messages
- High message rate
- Handle small to big message for any IO or MPI

Scalability, reliability and resiliency
- Fault tolerant, redundant, Minimal memory footprint
- Consistent performance as communicating pair count grows
- Adaptive routing, Rich topology with congestion management
- Accessed through standards-based APIs
- Power consumption

Rich, application-oriented Native transports
- RDMA send/receive, read/write
- PGAS, True network atomics, Flexible non-blocking collectives

Message Sizes
More Efficient to send without additional setup and response time. Memory copy time is smaller than standard setup

Performance Requirements
- High Message Rate
- Highly Latency Sensitive
- Lower Bandwidth
- Medium Message Rate
- Latency Sensitive
- Medium Bandwidth
- Lower Message Rate
- Some Latency Sensitivity
- High Bandwidth

OFA Annual Workshop, 2017

OPA fabric performance management and monitoring
Todd Rimmer, https://www.youtube.com/watch?v=I8TKC0Nqpc0

OPA fabric topologies and routing

MPI and storage traffic performance needs.
What to improve. Challenges.

<table>
<thead>
<tr>
<th>#node</th>
<th>48-port switch: #Chips</th>
<th>48-port switch: latency (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>5</td>
<td>330</td>
</tr>
<tr>
<td>256</td>
<td>17</td>
<td>330</td>
</tr>
<tr>
<td>512</td>
<td>34</td>
<td>330</td>
</tr>
<tr>
<td>1024</td>
<td>67</td>
<td>330</td>
</tr>
<tr>
<td>2048</td>
<td>266</td>
<td>550</td>
</tr>
<tr>
<td>4096</td>
<td>531</td>
<td>550</td>
</tr>
<tr>
<td>8192</td>
<td>1062</td>
<td>550</td>
</tr>
<tr>
<td>16384</td>
<td>2123</td>
<td>550</td>
</tr>
</tbody>
</table>

Need to increase the current number of port:
- => will decrease number of switches / cluster
- => will decrease number of hop / comms
- => will decrease latency & power

- High port counts have major implications
  - New router μArchitectures
  - New network topologies
- Must balance cost/power/distance between electrical and optical
  - Electrical: Lowest power, very short distance and lower bandwidth
  - Optical: Much longer distance and higher bandwidth

**OPA High Message Rate:** < 200M messages/s per switch port.
**Low Latency:** Port-to-port latency: <110ns. Only 3 hops between any two nodes (330ns latency) for 1024-node.
## What to improve

<table>
<thead>
<tr>
<th></th>
<th>OPA now</th>
<th>Mandatory for Exascale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transfer rate / lane</strong></td>
<td>25Gbps</td>
<td>25 Gbps min. More if possible (HDW)</td>
</tr>
<tr>
<td><strong>MPI L4 / PGAS L4</strong></td>
<td>Performance Scaled Messaging (PSM)</td>
<td>+ MPI Offload + HW Atomic , Collectives (Hdw + Sfw)</td>
</tr>
<tr>
<td><strong>Adaptive routing</strong></td>
<td>Coarse and Medium</td>
<td>+ Fine grained (Hdw + Sfw)</td>
</tr>
<tr>
<td><strong>Switch latency</strong></td>
<td>110 ns</td>
<td>Decrease as “possible” (HDW)</td>
</tr>
<tr>
<td><strong>MPI Message Rate (1 rank, N rank)</strong></td>
<td>(&lt; 4M , &lt; 200M)</td>
<td>Increase as “possible” (HDW)</td>
</tr>
<tr>
<td><strong>MPI Bandwidth (N rank, bidi, 1port)</strong></td>
<td>23.5-24.5 GB/s</td>
<td>Increase as “possible” (HDW)</td>
</tr>
</tbody>
</table>

“possible” definition: All of them should lead to a significant improvement!
Process Technology Scaling Trends will lead to on-chip specialization

- we will be able to build more transistors than we can power simultaneously
- Architecture will become more specialized
- different algorithms will use different transistors to operate most efficiently
- transistors not in use will be shut off
- Interconnect will benefit from it too

Normalized Trends offered by Moore’s Law

=> Area budget Increasing faster than power budget

* Assumes 2X area and 1.4X power scaling per process node

Source: Intel

Process Technology Trends drive the need for Specialized HW Architecture (dim/dark Silicon)
Integrated Processor Network Offers Unique Optimizations

**Conventional**

- CPU
- NIC
- Router

**Integrated**

- CPU
- iNIC
- Router

**PCIExpress**
- 10 pJ/bit
- 64 GB/s/direction = 5W

**Custom Link**
- 3 pJ/bit
- 64 GB/s/direction = 1.7W

**Custom Link**
- 3 pJ/bit
- 256 GB/s/direction = 5.1W
Will the Current Programming Models Scale to Exascale?

Current models based on communications between sequential processes depend on check-pointing for resilience. Not easy to just say: “Rewrite all your programs in some new language!”

Resilience, Programmability, Scalability
Why co-design: The application point of view

- Best way to evaluate new core concept to address both multi-threaded / single threaded performance
- Unique approach to resource management on the die both for core and uncore (energy and performance)
- Addressing programmability including innovative approaches to scaling in MPI (+ something or not)
- All aspects of system design of intra and inter-die communication, optimized towards energy efficiency
- Extensive reliability/resilience design to minimize failure rates due to error rates from factors such as use of near threshold voltage operation combined with the error rates of O(100k) nodes in an Exascale system.
- Emulation/simulation to allow more extensive investigations from mini-apps to real applications

Co design maintains the link between Architects and Real End Users
Conclusions (1)

Today:
Impact of increasing number of cores

MPI only (before MPI3)
• Not enough memory to continue MPI only
• Communication becomes >> Compute

Omp only
• Scalability issues within the node
• And then at machine level (for several reasons)

Future (the day after tomorrow)

• MPI + OMP, or MPI +X
• or any Shared memory extension
• Many + and –
• Long discussion about “threads” or “processes”
• BUT the node get more cores, And the balance of (comm, io) versus compute has (still) to be addressed

Interconnection challenges (hdw & sfw), “cores”, programming models, HPC applications
⇒ Cannot be treated separately

⇒ We can do measurements today, but we have to model what’s going to happen to the apps
Scalability prediction: the key question

What is breaking scalability in my apps?
- Amdahl et al. is: serial + // + overhead parts
- Overhead is different for strong and weak scaling
- How to predict overhead increase? (statistical or analytical)
- Is the node level performance really impacting?
- Industry care about strong scaling for the coming 2y
  - (workloads constant, just wants to be faster)
  - After 3y, weak scaling matters.

Strong scaling.
Inflection point due to Communication or IO > Compute
And/or due to huge latency increases

Weak scaling.
Inflection point due to transfer total size > Total Interconnection BW
Full system prediction Overview

- Mpi traces
- Real workloads running on “#nodes” for a given cpu + a given interconnect

- Proto application needed, including MPI region & Compute region
- Both being simulated differently

- High level discriminant runs
- Need analytical model and or statistical to compute strong and weak scalability

Deal oriented

- Similar microU and nb of cores
- Projection using New interconnect (lat+bw)
- Real workloads
- Associated with compute extrapolation

Interconnect Simulator

- Need compute projection or interface with other simulator as Sniper
- Use future network Spec
- Any nb of cores
- Not the real workloads

“UQ”

- Could be any arch and any interconnect
- Uncertainties over all parameters of the model including Uncertainties quantification + sensitivity analysis
- Real app, real workload
Execution-Driven Performance Projections

- **Simulator Requirement**
  - Enable execution-driven of unmodified WL Binary over Next-Gen Cluster Arch Simulator
  - Enable **Functional + Timing** Simulation for Performance
  - Enable “What-If” Architectural Exploration

![Diagram showing scale-out host cluster with parallel and distributed workload, host switches, host network switch, and performance visualization.]
Execution-Driven Performance Projections

- **Simulator Requirement**
  - Enable execution-driven of unmodified WL Binary over Next-Gen Cluster Arch Simulator
  - Enable **Functional + Timing** Simulation for Performance
  - Enable “What-If” Architectural Exploration
Conclusions (2)

Hardware: We know what to improve and can model the gains
Many physical challenges to address to stay in the “Exa” power budget (20 pJ / FpOps)

Programming models. MPI, OMP, SHMEM, ...
Hardware – software “agreement”: following and contributing to Standards for portability
As for HDW, we need to go there all together (almost)

Then Co design projects remain the key to integrate industrial application needs into new designs
but we need Functional + Timing Simulator at system level for Architectural Exploration
Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel.

Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804