A CORAL SYSTEM AND IMPLICATIONS FOR FUTURE HPC HARDWARE AND DATA CENTERS

Michael K Patterson
Senior Principal Engineer; Power, Packaging & Cooling
Intel, Technical Computing Systems Architecture and Pathfinding
Acknowledgement

Contributors, borrowed content, and data sources

Mark Seager, Intel
Ram Nagappan, Intel
Susan Coghlan, Argonne National Lab
Helmut Satzger, LRZ, Munich
Jim Rogers, ORNL
Cray; Aurora system partners

Appreciation

NCAR Team for continuing to invite us.
Intel Core™ i7 with over 1 billion transistors and over 6 miles (10 km) of wires in the chip to connect them.
It would only take eight Core™ i7s to make a wire long enough to connect Glasgow to Edinburgh!
Intel Investments Paving the Way
Holistic Approach to Cluster Solutions

**CPU**
- Intel® Xeon® Processors
- Intel® Xeon Phi™ Product Family

**Software & Tools**
- Intel® Parallel Studio
- Intel® Enterprise Edition for Lustre* software

**Fabric**
- Intel® Omni-Path Architecture

**Storage**
- Intel® Solid-State Drives (NVMe)

*Other names and brands may be claimed as the property of others.*
Intel’s Scalable System Framework
A Configurable Design Path Customizable for a Wide Range of HPC & Big Data Workloads

Small Clusters Through Supercomputers
Compute and Data-Centric Computing
Standards-Based Programmability
On-Premise and Cloud-Based

Intel® Xeon® Processors
Intel® Xeon Phi™ Coprocessors
Intel® Xeon Phi™ Processors
Intel® True Scale Fabric
Intel® Omni-Path Architecture
Intel® Ethernet
Intel® SSDs
Intel® Lustre-based Solutions
Intel® Silicon Photonics Technology
Intel® Software Tools
HPC Scalable Software Stack
Intel® Cluster Ready Program
Acquire DOE 2018 – 2022 Leadership Computing Capability
Three leadership class systems – one each at ALCF, LLNL, OLCF
  - With arch diversity between ALCF and OLCF
ALCF: Intel (Prime) Cray (Integrator)
OLCF: IBM (Prime)
LLNL: IBM (Prime)
THE FUTURE
The Most Advanced Supercomputer Ever Built
An Intel-led collaboration with ANL and Cray to accelerate discovery & innovation

>180 PFLOPS
(option to increase up to 450 PF)
>50,000 nodes
13MW
2018 delivery

18X higher performance*
>6X more energy efficient*

Source: Argonne National Laboratory and Intel, *Versus ANL's current biggest system named MIRA (10PFs and 4.8MW)
Other names and brands may be claimed as the property of others.
Aurora | Built on a Powerful Foundation

Breakthrough technologies that deliver massive benefits

**Compute**

- **3rd Generation Intel® Xeon Phi™**
- **2nd Generation Intel® Omni-Path**

- **>17X performance†**
  - FLOPS per node

- **>12X memory bandwidth†**
  - >30PB/s aggregate in-package memory bandwidth

- **Integrated Intel® Omni-Path Fabric**
  - Processor code name: Knights Hill

**Interconnect**

- **Intel® Omnipath™**

- **>20X faster†**
  - >500 TB/s bi-section bandwidth

- **>2.5 PB/s aggregate node link bandwidth**

**File System**

- **Intel® **

- **>3X faster†**
  - >1 TB/s file system throughput

- **>5X capacity†**
  - >150TB file system capacity

Source: Argonne National Laboratory and Intel
*Other names and brands may be claimed as the property of others.

† Comparisons are versus Mira—Argonne National Laboratory’s current largest HPC system, Mira. See Aurora Fact Sheet for details

Intel Confidential — Do Not Forward
Aurora Fact Sheet

<table>
<thead>
<tr>
<th>System Feature</th>
<th>The Aurora Details</th>
<th>Comparison to Mira</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak System Performance (fLOPs)</td>
<td>160 - 400 fLOPs/Do</td>
<td>10 fLOPs/Do</td>
</tr>
<tr>
<td>Processor</td>
<td>Xeon Phi™ Knights Hill</td>
<td></td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>&gt;50,000</td>
<td>49,192</td>
</tr>
<tr>
<td>Compute Platform</td>
<td>Intel system based on Grif, Blacha next generation supercomputing options</td>
<td>IBM Blue Gene/Q</td>
</tr>
<tr>
<td>Aggregated High Bandwidth On-Package Memory, Local Memory and Persistent Memory</td>
<td>&gt;7,000 Terabytes</td>
<td>191 Terabytes</td>
</tr>
<tr>
<td>Aggregated High Bandwidth On-Package Memory Bandwidth</td>
<td>&gt;0.5 Petabytes/s</td>
<td>2.3 Petabytes/s</td>
</tr>
<tr>
<td>System Interconnect</td>
<td>2nd Generation Intel Omni-Path Architecture with silicon interconnect</td>
<td>IBM SD Canvas interconnect with QDR, interconnect</td>
</tr>
<tr>
<td>Interconnect Aggregate Node Link Bandwidth</td>
<td>&gt;2.0 Petabytes/s</td>
<td>2 Petabytes/s</td>
</tr>
<tr>
<td>Interconnect Bandwidth</td>
<td>&gt;500 Terabytes/s</td>
<td>24 Terabytes/s</td>
</tr>
<tr>
<td>Interconnect Interface</td>
<td>Integrated</td>
<td>Integrated</td>
</tr>
<tr>
<td>Data Buffer Storage</td>
<td>Intel® SSDs, using both 1st and 2nd Generation Intel® Omni-Path Architecture</td>
<td>None</td>
</tr>
<tr>
<td>File System</td>
<td>Intel® Lustre File System</td>
<td>IBM GPFS File System</td>
</tr>
<tr>
<td>File System Capacity</td>
<td>&gt;100 Petabytes</td>
<td>26 Petabytes</td>
</tr>
<tr>
<td>File System Throughput</td>
<td>&gt;1 Terabytes</td>
<td>200 Gigabytes</td>
</tr>
<tr>
<td>Intel Architecture [Intel® K6] Compatibility</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Peak Power Consumption</td>
<td>13 Megawatts</td>
<td>4.0 Megawatts</td>
</tr>
<tr>
<td>fLOPs Per Watt</td>
<td>&gt;12 fLOPs/Do per watt</td>
<td>&lt;2 fLOPs/Do per watt</td>
</tr>
<tr>
<td>Delivery Timeline</td>
<td>2019</td>
<td>2012</td>
</tr>
<tr>
<td>Facility Area for Compute Clusters</td>
<td>~3,000 sq ft</td>
<td>~1,536 sq ft</td>
</tr>
</tbody>
</table>

For further information on Aurora, visit: intel.com/Aurora

All the details: Aurora Fact Sheet at intel.com
Intel SSF enables Higher Performance & Density

A formula for more performance....

advancements in CPU architecture

- advancements in process technology
  - integrated in-package memory
    - integrated fabrics with higher speeds
      - switch and CPU packaging under one roof
        - all tied together with silicon photonics

= much higher performance & density
So what have we learned over the last three years?

Today's focus is on Power, Packaging, and Cooling (PPC)

- **Power**
  - 480Vac
  - >100 kW / cabinet

- **Packaging**
  - High density computing – significant computing in a small package
  - Weight becomes a key design parameter

- **Cooling**
  - Liquid cooling; for a number of reasons
  - Cooler is better, to a point
Power

Trends....

- Power now 480 Vac 3ph (400 Vac in Europe)
- >100 kW / cabinet
- In-cabinet 380 Vdc for optimized delivery
- Power management and power monitoring allows optimized performance and efficiency
Power Delivery Challenges in the horizon

Variable Power Cap

- Several reasons
  - Peak Shedding
  - Reduction in renewable energy

Power rate of change

- Ex: Hourly or Fifteen minute average in platform power should not exceed by X MW.

Controlled Power Ramp up/down – economic or technical issues

- Challenge to do this at a reasonable cost and with energy efficient mechanisms
Packaging

Rack and cluster weight and density

- Packaging
  - High density computing – network topology optimization and high node count per rack make for dense cabinets

- Rack weight density
  - Design limit: Floor tiles at 500 lbs/sf ~ 2500 kg/m2

- White space vs utility space
  - Compute density increasing, infrastructure support equipment is not

- What's the trend for machine room area?
I must need a huge data center for PetaScale and ExaScale computing – Right?
Quick Survey

In what year, do you expect to see “rack level” performance exceed 1 PF?

a) 2016  
b) 2018  
c) 2020  
d) 2022
Performance density continues to increase

<table>
<thead>
<tr>
<th>System Feature</th>
<th>LRZ Phase 1</th>
<th>LRZ Phase 2</th>
<th>Mira</th>
<th>Titan</th>
<th>Summit</th>
<th>Aurora</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
<td>2012</td>
<td>2015</td>
<td>2012</td>
<td>2012</td>
<td>2017</td>
<td>2018</td>
</tr>
<tr>
<td><strong>Perf P flop/s</strong></td>
<td>3.2</td>
<td>3.2</td>
<td>10</td>
<td>27</td>
<td>150</td>
<td>180</td>
</tr>
<tr>
<td><strong># of Nodes</strong></td>
<td>9216</td>
<td>3096 Haswell</td>
<td>49,152</td>
<td>18,688</td>
<td>3500</td>
<td>&gt;50,000 KNH</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>2.3 MW</td>
<td>1.1 MW</td>
<td>4.8 MW</td>
<td>9 MW</td>
<td>9.6 MW</td>
<td>13 MW</td>
</tr>
<tr>
<td><strong>Cluster Area (m²) est.</strong></td>
<td>546</td>
<td>182</td>
<td>143</td>
<td>400</td>
<td>418</td>
<td>279</td>
</tr>
<tr>
<td><strong>Cluster Area (ft²) est.</strong></td>
<td>5875</td>
<td>1960</td>
<td>1540</td>
<td>4300</td>
<td>4500</td>
<td>3000</td>
</tr>
<tr>
<td><strong>TF/m² Est.</strong></td>
<td>6</td>
<td>18</td>
<td>70</td>
<td>67.5</td>
<td>359</td>
<td>645</td>
</tr>
<tr>
<td><strong>TF/ft² est.</strong></td>
<td>0.5</td>
<td>1.6</td>
<td>6.5</td>
<td>6.3</td>
<td>33.4</td>
<td>60</td>
</tr>
</tbody>
</table>

Source: lrz.de
Do I need a huge data center?

- Facility area for Compute Cluster does not have to be huge. Significant compute density in small packages
  - At Aurora density, the 3.2 LRZ PF step could fit in 5 m²

- Don’t forget:
  - If Storage is going to be large then you will need additional floor space.
  - If you are going to be using Xeon instead of Xeon Phi then you may need additional floor space
  - Utility and infrastructure space continues to grow
Cooling

Why liquid?

- Power per node continues to rise
- Rack density limits airflow path
- Increased thermal performance of liquid (vs air) allows more free-cooling
  - Thermal resistance from chip to liquid in a cold plate is smaller than chip to air over a heat sink
- Warmer or cooler? “Warm-water cooling” has a good ring to it!
What does Warm-Water Cooling really mean?

- W1: 17°C (63°F)
- W2: 27°C (81°F)
- W3: 32°C (90°F)
- W4: 45°C (113°F)

Not to scale
Just say no....

Warm Water Cooling

Instead, define either specific temperatures or functional....

Define the temperature at the facilities and IT water loop interface

W2, W3, W4
ASHRAE values help system vendor design system, guarantee performance

Define how the water temperature is made

Chiller
Cooling Tower
Dry Cooler
A proposal....

- As a starting point, use the coolest water you can make without a chiller
- Always be above the dewpoint (to prevent condensation in the machine)
- Cooler temperatures promote:
  - Lower leakage
  - More turbo frequencies
  - Higher stability
  - More time to recover in an upset condition
  - Better reliability
  - Reduced flow rates

Note - May consume more water, not applicable if after heat recovery
Why use “warm” water, when “cool” water costs the same?

- \( W_2 = 27^\circ C \)
- \( W_3 = 32^\circ C \)
- \( W_4 = 45^\circ C \)
Summary

Planning for Exascale needs to happen now; 180 PF in 2018

Designing for the future:
It's going to be Large!
- kg/rack, kW/rack, perf/rack, power ramps and peak, pipe sizes, m2/m2
It may get Smaller!
- Cluster footprint

High packaging density, high power, liquid cooling all enable best performance, efficiency, and TCO
Aurora

It’s one more landmark.

It’s the next one we have to reach.

But the journey does not stop there.
Thanks for your attention

Questions?

michael.k.patterson@intel.com
Legal Disclaimer

Today's presentations contain forward-looking statements. All statements made that are not historical facts are subject to a number of risks and uncertainties, and actual results may differ materially.

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. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.

Intel does not control or audit the design or implementation of third party benchmarks or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmarks are reported and confirm whether the referenced benchmarks are accurate and reflect performance of systems available for purchase.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. See www.intel.com/products/processor_number for details.

Intel, processors, chipsets, and desktop boards 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.

Intel, Intel Xeon, Intel Core microarchitecture, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others

Copyright © 2011, Intel Corporation. All rights reserved.