Hornet: The new Cray **XC40** Supercomputer at HLRS

17\textsuperscript{th} Results and Review Workshop at HLRS
2014-09-30
Stefan Andersson
Agenda this talk

All basics about the hardware of the XC
● Haswell
● XC Packaging
● Aries and the Dragonfly network
● DataWarp I/O acceleration technology
● Hermit vs Hornet
Haswell
Most of the following slides were provided by Intel.
Haswell builds upon innovations in the 2\textsuperscript{nd} and 3\textsuperscript{rd} Generation Intel\textsuperscript{®} Core\textsuperscript{TM} i3/i5/i7 Processors (Sandy Bridge and Ivy Bridge).
Intel® Xeon® E5-2600v3 Processor Overview

- 22nm Process (Tock)
- PCI Express 3.0 EP: 40 Lanes
- Intel® Hyper-Threading Technology (2 threads/core)
- Intel® Turbo Boost Technology
- Up to 18 Cores
- Integrated Voltage Regulator

**Power Management**
- Per Core P-State (PCPS)
- Uncore Frequency Scaling (UFS)
- Energy Efficient Turbo (EET)

**Memory Technology:**
- Socket R3
- 4xDDR4 channels
- 1333, 1600, 1866, 2133 MTS

- ~2.5 MB Last Level Cache/Core
- Up to 45 MB total LLC

- Intel® AVX 2.0 / Haswell New Instruction (HNI)
- Intel® QuickPath Interface (x2)
- 9.6GT/s
# Haswell Buffer Sizes

Extract more parallelism in every generation

<table>
<thead>
<tr>
<th></th>
<th>Nehalem</th>
<th>Sandy Bridge</th>
<th>Haswell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out-of-order Window</td>
<td>128</td>
<td>168</td>
<td>192</td>
</tr>
<tr>
<td>In-flight Loads</td>
<td>48</td>
<td>64</td>
<td>72</td>
</tr>
<tr>
<td>In-flight Stores</td>
<td>32</td>
<td>36</td>
<td>42</td>
</tr>
<tr>
<td>Scheduler Entries</td>
<td>36</td>
<td>54</td>
<td>60</td>
</tr>
<tr>
<td>Integer Register File</td>
<td>N/A</td>
<td>160</td>
<td>168</td>
</tr>
<tr>
<td>FP Register File</td>
<td>N/A</td>
<td>144</td>
<td>168</td>
</tr>
<tr>
<td>Allocation Queue</td>
<td>28/thread</td>
<td>28/thread</td>
<td>56</td>
</tr>
</tbody>
</table>
Haswell Execution Unit Overview

Unified Reservation Station

Port 0
- Integer ALU & Shift
- FMA FP Multiply
- Vector Multiply
- Vector Logical
- Branch
- Divide
- Vector Shifts

Port 1
- Integer ALU & LEA
- FMA FP Multi
- FP Add
- Vector Int ALU
- Vector Logical

Port 2
- Load & Store Address
- Store Data
- Integer ALU & LEA
- Integer ALU & Shift
- Vector Int ALU

Port 3
- Vector Logical

Port 4

Port 5
- Integer ALU & LEA
- Vector Shuffle
- Vector Int ALU
- Vector Logical

Port 6

Port 7
- Store Address

2xFMA
- Doubles peak FLOPs
- Two FP multiplies benefits legacy

4th ALU
- Great for integer workloads
- Frees Port 0 & 1 for vector

New Branch Unit
- Reduces Port 0 Conflicts
- 2nd EU for high branch code

New AGU for Stores
- Leaves Port 2 & 3 open for loads
Haswell New Compute Instructions

• Intel® Advanced Vector Extensions 2 (Intel® AVX2)
  – Includes
    ▪ 256-bit Integer vectors
    ▪ FMA: Fused Multiply-Add
    ▪ Full-width element permutes
    ▪ Gather
  – Benefits
    ▪ High performance computing
    ▪ Audio & Video
    ▪ Games

• New Integer Instructions
  – Indexing and hashing
  – Cryptography
  – Endian conversion – MOVBE

<table>
<thead>
<tr>
<th>Group</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bit Field Pack/Extract</td>
<td>BZHI, SHLX, SHRX, SARX, BEXTR</td>
</tr>
<tr>
<td>Variable Bit Length Stream Decode</td>
<td>LZCNT, TZCNT, BLSR, BLSMSK, BLSI, ANDN</td>
</tr>
<tr>
<td>Bit Gather/Scatter</td>
<td>PDEP, PEXT</td>
</tr>
<tr>
<td>Arbitrary Precision Arithmetic &amp; Hashing</td>
<td>MULX, RORX</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instruction Set</th>
<th>SP FLOPs per cycle</th>
<th>DP FLOPs per cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nehalem SSE (128-bits)</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Sandy Bridge AVX (256-bits)</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Haswell AVX2 &amp; FMA</td>
<td>32</td>
<td>16</td>
</tr>
</tbody>
</table>

FMA & Peak FLOPS

- 2 new FMA units provide 2x peak FLOPs/cycle of previous generation

- 2X cache bandwidth to feed wide vector units
  - 32-byte load/store for L1
  - 2x L2 bandwidth

- 5-cycle FMA latency same as an FP multiply

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

Intel® Microarchitecture (Haswell); Intel® Microarchitecture (Sandy Bridge); Intel® Microarchitecture (Meron); Intel® Microarchitecture (Banias)
Which workloads will run at what frequency?

Frequency range of E5-2699 v3@2.3GHz

- **Max All Core Turbo Frequency (Non-AVX)**
- **Base Frequency (Non-AVX)**
- **AVX Base Frequency**
- **AVX Max All Core Turbo Frequency**

Expected frequencies for:
- non-AVX workloads
- workloads with heavy AVX usage

<table>
<thead>
<tr>
<th>Frequency Range</th>
<th>Expected for:</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8</td>
<td>AVX Base Frequency</td>
</tr>
<tr>
<td>2.7</td>
<td>AVX Max All Core Turbo Frequency</td>
</tr>
<tr>
<td>2.6</td>
<td>Expected frequencies for most AVX workloads</td>
</tr>
<tr>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>1.9</td>
<td></td>
</tr>
</tbody>
</table>
## Core Cache Size/Latency/Bandwidth

<table>
<thead>
<tr>
<th>Metric</th>
<th>Nehalem</th>
<th>Sandy Bridge</th>
<th>Haswell</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Instruction Cache</td>
<td>32K, 4-way</td>
<td>32K, 8-way</td>
<td>32K, 8-way</td>
</tr>
<tr>
<td>L1 Data Cache</td>
<td>32K, 8-way</td>
<td>32K, 8-way</td>
<td>32K, 8-way</td>
</tr>
<tr>
<td>Fastest Load-to-use</td>
<td>4 cycles</td>
<td>4 cycles</td>
<td>4 cycles</td>
</tr>
<tr>
<td>Load bandwidth</td>
<td>16 Bytes/cycle</td>
<td>32 Bytes/cycle (banked)</td>
<td>64 Bytes/cycle</td>
</tr>
<tr>
<td>Store bandwidth</td>
<td>16 Bytes/cycle</td>
<td>16 Bytes/cycle</td>
<td>32 Bytes/cycle</td>
</tr>
<tr>
<td>L2 Unified Cache</td>
<td>256K, 8-way</td>
<td>256K, 8-way</td>
<td>256K, 8-way</td>
</tr>
<tr>
<td>Fastest load-to-use</td>
<td>10 cycles</td>
<td>11 cycles</td>
<td>11 cycles</td>
</tr>
<tr>
<td>Bandwidth to L1</td>
<td>32 Bytes/cycle</td>
<td>32 Bytes/cycle</td>
<td>64 Bytes/cycle</td>
</tr>
<tr>
<td>L1 Instruction TLB</td>
<td>4K: 128, 4-way 2M/4M: 7/thread</td>
<td>4K: 128, 4-way 2M/4M: 8/thread</td>
<td>4K: 128, 4-way 2M/4M: 8/thread</td>
</tr>
<tr>
<td>L1 Data TLB</td>
<td>4K: 64, 4-way 2M/4M: 32, 4-way 1G: fractured</td>
<td>4K: 64, 4-way 2M/4M: 32, 4-way 1G: 4, 4-way</td>
<td>4K: 64, 4-way 2M/4M: 32, 4-way 1G: 4, 4-way</td>
</tr>
<tr>
<td>L2 Unified TLB</td>
<td>4K: 512, 4-way</td>
<td>4K: 512, 4-way</td>
<td>4K+2M shared: 1024, 8-way</td>
</tr>
</tbody>
</table>

All caches use 64-byte lines
On-Die Interconnect Enhancements

E5-2600 v2

E5-2600 v3
Haswell EP Die Configurations

Not representative of actual die-sizes, orientation and layouts – for informational use only.

<table>
<thead>
<tr>
<th>Chop</th>
<th>Columns</th>
<th>Home Agents</th>
<th>Cores</th>
<th>Power (W)</th>
<th>Transistor s (B)</th>
<th>Die Area (mm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCC</td>
<td>4</td>
<td>2</td>
<td>14-18</td>
<td>110-145</td>
<td>5.69</td>
<td>662</td>
</tr>
<tr>
<td>MCC</td>
<td>3</td>
<td>2</td>
<td>6-12</td>
<td>65-160</td>
<td>3.84</td>
<td>492</td>
</tr>
<tr>
<td>LCC</td>
<td>2</td>
<td>1</td>
<td>4-8</td>
<td>55-140</td>
<td>2.60</td>
<td>354</td>
</tr>
</tbody>
</table>
Haswell EP Die Configurations

10-12 Core (MCC)

Not representative of actual die-sizes, orientation and layouts – for informational use only.
# Haswell Processor Improvements

<table>
<thead>
<tr>
<th>Area</th>
<th>Change</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>On-die interconnect</strong></td>
<td>• Two Fully Buffered Rings</td>
<td>• Enables higher core counts and provides higher bandwidth per core.</td>
</tr>
<tr>
<td><strong>Home Agent / Memory Controller</strong></td>
<td>• DDR4 • Two Home Agents in more SKUs • Directory Cache</td>
<td>• Increased memory bandwidth and power efficiency • Greater socket BW with more outstanding requests • Lower average memory latency</td>
</tr>
<tr>
<td><strong>LLC</strong></td>
<td>• Cluster On Die (COD) mode • Improved LLC allocation policy • Cache Allocation Monitoring</td>
<td>• Increased performance, reduced latency • Enables improved performance by better application placement in a virtualized environment</td>
</tr>
<tr>
<td><strong>Power Management</strong></td>
<td>• Separate clock and voltage domains for each core and uncore (enables PCPS, UFS)</td>
<td>• Better performance per watt • Lower socket idle (package C6) power.</td>
</tr>
<tr>
<td><strong>QPI 1.1</strong></td>
<td>• Increase to 9.6GT/s</td>
<td>• Multi-socket coherence performance</td>
</tr>
<tr>
<td><strong>Integrated IO-Hub (IIO)</strong></td>
<td>• LLC cache tracks IIO cache line ownership • Increased PCIe buffers and credits</td>
<td>• Improves PCIe bandwidth under conflicts (concurrent accesses to the same cache line). • Increase PCIe bandwidth and latency tolerance</td>
</tr>
<tr>
<td><strong>PCI Express 3.0</strong></td>
<td>• DualCast - Allows a single write transaction to multiple targets. • Relaxed ordering</td>
<td>• Utilized to minimize memory channel bandwidth – data can be sent to memory and on the NTB port. Storage applications are typically memory bandwidth limited.</td>
</tr>
</tbody>
</table>
DDR4 Benefits

Lower Power
- Lower voltage (1.5v -> 1.2v) DIMMs
- Smaller page size (1024 -> 512) for x4 devices
- Initial results show savings of ~2W per DIMM at the wall.

Improved RAS
- Command/Address Parity error recovery

Higher bandwidth
- 14% higher STREAM results (DDR4-2133 vs. DDR3-1866)
- Increased DIMM frequency when multiple DIMMs per channel are installed

<table>
<thead>
<tr>
<th>Dimms / Ch</th>
<th>DDR3 1.5v</th>
<th>DDR3 1.35v</th>
<th>DDR4 RDIMM</th>
<th>DDR4 LRDIMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1866</td>
<td>1600</td>
<td>2133</td>
<td>2133</td>
</tr>
<tr>
<td>2</td>
<td>1600</td>
<td>1333</td>
<td>1866</td>
<td>2133</td>
</tr>
<tr>
<td>3</td>
<td>1066</td>
<td>800</td>
<td>1600</td>
<td>1600</td>
</tr>
</tbody>
</table>
Summary

• Haswell is the next Intel® “tock” microarchitecture, builds upon Sandy Bridge to deliver:
  – Scalability across broad range of domains and workloads
  – Per core performance for the vast majority of workloads
  – Lower power for better performance and smaller envelopes

• Developer-friendly features
  – Fundamental performance and power improvements for legacy workloads, including AVX
  – New instructions addressing key developer requests
    ▪ Intel® AVX2 with FMA and 256-bit integer vectors
    ▪ Intel® Bit Manipulation Instructions
    ▪ Intel® TSX for thread parallelism through lock elision

• Focus on power
  – Microarchitecture improvements: deeper idle states, lower active power
  – Finer grain control: more voltage and frequency domains, improved power sharing
XC30 + XC40 Packaging
Cray XC System Building Blocks

- **Compute Blade**: 4 Compute Nodes
- **Chassis**: Rank 1 Network, 16 Compute Blades, No Cables, 64 Compute Nodes
- **Group**: Rank 2 Network, Passive Electrical Network, 2 Cabinets, 6 Chassis, 384 Compute Nodes
- **System**: Rank 3 Network, Active Optical Network, Hundreds of Cabinets, Up to 10s of thousands of nodes
Cray XC Fully Populated Compute Blade

**SPECIFICATIONS**

- **Module power:** 2014 Watts
- **PDC max. power:** 900 Watt
- **Air flow req.:** 275 cfm (7.8 m³/min)
- **Size:** 2.125 in x 12.95 in x 33.5 in
- **Weight:** <40 lbm (18 kg)
PDC’s (Processor Daughter Card) are Upgradeable to New Technology
Cray XC Quad Processor Daughter Card

- Intel Processor (4)
- Voltage Reg (2)
- Southbridge (2)
- DDR Memory (16)
Cray XC I/O Module

SPECIFICATIONS
Module power: 1650 Watts
PDC max. power: 225 Watt
Air flow req.: 275 cfm
Size: 2.125 in x 12.95 in x 33.5 in
Weight: 35 lbs
XC Cooling
XC Cooling Overview

- Single air inlet and exhaust
- Preconditioning option relaxes inlet air requirements
- Outlet air conditioned (room neutral)
Cray XC Transverse Cooling Advantages

● **Performance**
  ● Transverse cooling and graduated heat sink pitch ensure that all processors operate in the same thermal envelope
  ● “Turbo mode” works like it should in a parallel job

● **Simplicity**
  ● No airflow issues to manage or adjust
  ● System is 100% water-cooled
  ● No pumps, refrigerant, treated water, or plumbing on the blades

● **Cost of Ownership**
  ● Excellent PUE characteristics
  ● 25% better density than other ‘direct’ water cooled solutions
  ● All cooling infrastructure is retained across multiple generations of computing technology

● **Maintainability**
  ● Blades can be warm-swapped without disturbing any plumbing
  ● Blowers can be hot-swapped if required and can provide N+1 redundancy
Aries
Aries

- Aries is the Cray custom interconnect ASIC used in the Cray XC product family
  - 40nm process
  - Die size: 16.6 x 18.9mm
  - Gate count: 217M
  - 184 lanes of high speed SerDes
    - SerDes = Serializer/Deserializer
      (SerDes pronounced sir-deez)
  - 30 optical network lanes
  - 90 electrical network lanes
  - 64 PCI Express lanes
- **4 NICs**
  - Each Aries connects 4 nodes to the interconnect (Gemini connects 2)
Cray XC Compute Blade Architecture

- Dual QPI SMP Links
- 4 Channels DDR3 or DDR4
- PCIe-3 16 bits at 8.0 GT/s per direction
XC Compute Blade

- Intra Cabinet/Blue (Optic Cable x 10 links) 12.5 Gbps
- Chassis/Green (Backplane x 15 links) 14 Gbps
- PCIe-3 16 bits at 8.0 GT/s per direction
- Dual QPI SMP Links
- 4 Channels DDR3 or DDR4
- Inter Chassis/Black (Copper Cable x 15 links) 14 Gbps
- Aries 48-port Router 4 NICs, 2 router tiles each 40 router tiles for interconnect
Cray XC Dragonfly Topology
Cray XC Network

- The Cray XC system is built around the idea of optimizing interconnect bandwidth and associated cost at every level.

<table>
<thead>
<tr>
<th>Rank-1</th>
<th>Rank-2</th>
<th>Rank-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC Board: ₳₪₪</td>
<td>Passive CU: $</td>
<td>Active Optics: ₳₪₪₪</td>
</tr>
</tbody>
</table>

Cray XC Rank1 Network

- Chassis with 16 compute blades
- 128 Sockets
- Inter-Aries communication over backplane
- Per-Packet adaptive Routing
Cray XC Rank-2 Copper Network

- 6 backplanes connected with copper cables in a 2-cabinet group: “Black Network”
- 2 Cabinet Group 768 Sockets
- 16 Aries connected by backplane “Green Network”
- 4 nodes connect to a single Aries
- Active optical cables interconnect groups “Blue Network”
Cray XC Rank-2 Cabling

- Cray XC40 two-cabinet group
  - 768 Sockets
  - 96 Aries Chips
- All copper and backplanes signals running at 14 Gbps
Cray XC Routing

With adaptive routing we select between minimal and non-minimal paths based on load.

Minimal routes between any two nodes in a group are just two hops.

Non-minimal route requires four hops.

The Cray XC40 Class-2 Group has sufficient bandwidth to support full injection rate for all 384 nodes with non-minimal routing.
Aries Adaptive Routing Algorithm

Route tables (per port)

<table>
<thead>
<tr>
<th>Global minimal</th>
<th>Global non-minimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local minimal</td>
<td>Local non-minimal</td>
</tr>
</tbody>
</table>

Choice of 4 possible routes, 2 minimal and 2 non-minimal. A new selection is made for every packet.

Compare current loads for each of the 4 possible outputs, select the best.

Output packet on selected port.
Cray XC Network Overview – Rank-3 Network

- An all-to-all pattern is wired between the groups using optical cables (blue network)
- Up to 240 ports are available per 2-cabinet group
- The global bandwidth can be tuned by varying the number of optical cables in the group-to-group connections

Example: An 4-group system is interconnected with 6 optical “bundles”. The “bundles” can be configured between 20 and 80 cables wide
Adaptive Routing over the Blue Network

- An all-to-all pattern is wired between the groups.

Assume Minimal path from Group 0 to 3 becomes congested.

Traffic can “bounce off” any other intermediate group.

Doubles load on network but more effectively utilizes full system bandwidth.
Optical network saturation using the OSU MPI BM
This runs were done on Hornet with the 25% opt. cables config

- Saturating the optical network using communications between more and more nodes within 2 groups:
  - 2x300 nodes: 875 GB/s
Copper & Optical Cabling
Does Aries handle MPI Traffic with I/O Traffic?

I/O Messages

MPI Messages
Mix of application and streaming I/O traffic

- Analysis of the impact of big I/O traffic on performance of other codes
- Compared two runs
  1. Four miniGhost jobs spread out across the whole machine vs.
  2. Three miniGhost plus one performing big many-to-few I/O

I/O Job sustaining 400GB/sec (95% clients to 5% servers)
Impact to compute jobs is tiny (64.5 sec to 65 sec)
Why is the Dragonfly topology a good idea?

- **Scalability**
  - Topology scales to very large systems
- **Performance**
  - More than just a case of clever wiring, this topology leverages state-of-the-art adaptive routing that Cray developed with Stanford University
  - Smoothly mixes small and large messages
  - *Cray invested in bringing it to market – IBM and Mellanox have not*
- **Simplicity**
  - Implemented *without* external switches
  - No HBAs or separate NICs and Routers
- **Cost**
  - Dragonfly maximizes the use of backplanes and passive copper components
  - Dragonfly minimizes the use of active optical components
DataWarp I/O acceleration technology
DataWarp

- **DataWarp** is designed to improve application performance and drive down TCO for I/O intensive applications by:
  - bringing the fastest I/O applications as close as possible to the compute resources;
  - providing a flexible and balanced infrastructure mapping to different application use cases;
  - ensuring a cost-effective balance of I/O across all storage tiers;
  - creating a central pool of fast I/O resources that can be allocated in diverse ways;
  - and meeting the worst case data I/O surge needs.

- **Cray’s DataWarp technology** is available as an added feature to the Cray XC40 supercomputer.
Indicative Flash IO node configuration

Hornet has 2 DataWarp nodes, but they are not yet available for users.
Storage Hierarchy Software

Three phase development approach
1. Static partitioning of burst buffer storage to compute nodes
2. Flexible Flash storage allocation, explicit migration API
3. Transparent caching

Static Partitioning
- Node to burst buffer association fixed at boot time
- Global filesystem to burst buffer association fixed at boot time
- Provides swap and scratch space per node

BB Storage Reservation Manager
- Persistent and flexible allocations
- BB node/Compute node affinity

Application Interface
- Simple, explicit migration interface for files between BB and PFS, invoked from compute node

Transparent Caching
- No application changes needed
- Compute/burst/compute/ burst functionality
- Provides access to intermediate data
Hornet vs Hermit
# Hermit vs Hornet comparison

<table>
<thead>
<tr>
<th></th>
<th>Hornet</th>
<th>Hermit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System</strong></td>
<td>Cray XC40</td>
<td>Cray XE6</td>
</tr>
<tr>
<td><strong>Nr of compute nodes</strong></td>
<td>3944</td>
<td>3552</td>
</tr>
<tr>
<td><strong>Processor (2 per nodes)</strong></td>
<td>Intel Haswell 2.5 Ghz Intel E5-2680v3,12 cores</td>
<td>AMD Interlagos 2.3GHz 16 cores</td>
</tr>
<tr>
<td><strong>Total #cores</strong></td>
<td>94656 (HT not counted)</td>
<td>113664</td>
</tr>
<tr>
<td><strong>Interconnect</strong></td>
<td>Cray Aries</td>
<td>Cray Gemini</td>
</tr>
<tr>
<td><strong>Interconnect Topology</strong></td>
<td>Dragonfly</td>
<td>3d Torus</td>
</tr>
<tr>
<td><strong>Memory per node</strong></td>
<td>128 GB DDR4</td>
<td>32 or 64 GB DDR3</td>
</tr>
<tr>
<td><strong>Memory BW/Node</strong></td>
<td>114 MB/sec</td>
<td>70 MB/sec</td>
</tr>
<tr>
<td><strong>Peak PFlops</strong></td>
<td>3.79</td>
<td>1.05</td>
</tr>
<tr>
<td><strong>HPL (Linpack)</strong></td>
<td>2.76</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>KW while running HPL</strong></td>
<td>~1550</td>
<td>~1900</td>
</tr>
</tbody>
</table>
## File system setup: 6+2 file systems

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>OSSes</th>
<th>OTSs</th>
<th>TBytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry 1</td>
<td>4</td>
<td>24</td>
<td>166</td>
</tr>
<tr>
<td>Industry 2</td>
<td>4</td>
<td>24</td>
<td>166</td>
</tr>
<tr>
<td>University Capability</td>
<td>16</td>
<td>96</td>
<td>664</td>
</tr>
<tr>
<td>University Capacity</td>
<td>12</td>
<td>72</td>
<td>498</td>
</tr>
<tr>
<td>Research Capability</td>
<td>16</td>
<td>96</td>
<td>664</td>
</tr>
<tr>
<td>Research Capacity</td>
<td>12</td>
<td>72</td>
<td>498</td>
</tr>
<tr>
<td>ws7 – NEW</td>
<td>24</td>
<td>168</td>
<td>2400</td>
</tr>
<tr>
<td>ws8 – NEW</td>
<td>24</td>
<td>168</td>
<td>2400</td>
</tr>
</tbody>
</table>
Some Hints

- The module “craype-haswell” is loaded when compiling
  - If not, you will lose performance
  - This is default on Hornet
- An executable build for Haswell will in general NOT run on any other processor
  - This is not a problem for a parallel application you build on Hornet for Hornet, but if you build a ‘package’ including some support applications, you have to make sure to compile each application with the correct options
- Turbo Boost is nice, but will differ between nodes.
- You have 128 GB per node. That is a lot
- Using AVX2 will in most cases drop the frequency but you’ll gain efficiency
  - 2xpeak performance, vectorization of gather/scatter loops
Early Haswell experiences:

Very good Power distribution

- The Haswell chip does dynamic power routing, that is, it moves power between cores and across units of the chip to where is needed. As such the chip always runs at a higher TDP (thermal design power). This is expected behavior, and actually ‘LOWERS’ the power used.
- You might get higher performance by NOT using all cores.
- Using some AVX instructions, including FMAs, will cause a frequency drop up to 400 MHz.
Thank you!

Questions?
Comments?