Cray XC30 System
Cray Aries Custom Interconnect
<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>3X increase in sustained injection bandwidth to 10 GB/sec</td>
<td>Improves the efficiency of a wide range of communication intensive applications</td>
</tr>
<tr>
<td>Large increase in global bandwidth, from 3X at the low end to 20X or more at the high end</td>
<td>Benefits applications with complex communication patterns (unstructured meshes, adaptive mesh refinement, search and sort, data mining etc.)</td>
</tr>
<tr>
<td>Hardware support for collectives</td>
<td>Especially helpful on large jobs (1000s of cores) which utilize collective operations. HW support is about a 2X improvement over our best SW algorithms</td>
</tr>
<tr>
<td>3X increase in the rate of small puts and gets to 120M/sec</td>
<td>Key for new programming environments (PGAS), important to our DARPA mission partners and some of our more advances customers</td>
</tr>
<tr>
<td>PCI-Express Gen3 interface</td>
<td>Benefits Cray, reduces our dependency on a particular CPU vendor</td>
</tr>
</tbody>
</table>
Cray XC30 Network

- The Cray XC30 system is built around the idea of optimizing interconnect bandwidth and associated cost at every level.
Cray XC30 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 XC30 Modular Blades
Cray XC30 Compute Blade Architecture

Global links, optical (5 links x 2 ports) “Rank-3 Network” 12.5 Gbps signaling

Intra-group, backplane (15 links x 1 port) “Rank-1 Network” 14 Gbps signaling

Dual QPI SMP Links

4 Channels DDR3

Intra-group, copper cables (5 links x 3 ports) “Rank-2 Network” 14 Gbps signaling

PCle-3 16 bits at 8.0 GT/s per direction

Aries 48-port Router 8 ports to 4 NICs, 40 external network ports
Cray XC30 Fully Populated Compute Blade
Cray XC I/O Module

PCle Card Slots

Intel 2600 Series Processor

Riser Assembly

Aries
Cray XC30 Dragonfly Topology
Cray XC30 Rank1 Network

- Chassis with 16 compute blades
- 128 Sockets
- Inter-Aries communication over backplane
- Per-Packet adaptive Routing
Cascade – Local Electrical Network

- 16 Aries connected by backplane “Green Network”
- 6 backplanes connected with copper cables in a 2-cabinet group: “Rank-2 Network”
- Active optical cables interconnect groups “Rank-3 Network”
- 4 nodes connect to a single Aries
- 2 Cabinet Group 768 Sockets
Cray XC30 Rank-2 Cabling

- Cray XC30 two-cabinet group
  - 768 Sockets
  - 96 Aries Chips
With adaptive routing we select between minimal and non-minimal paths based on load. The Cray XC30 Class-2 Group has sufficient bandwidth to support full injection rate for all 384 nodes with non-minimal routing.

Minimal route between any two nodes in a group is just two hops.

Non-minimal route requires up to four hops.
Cray XC30 – Rank-3 Network

- An all-to-all pattern is wired between the groups using optical cables (blue network)

- The global bandwidth can be tuned by varying the number of optical cables in the group-to-group connections

Example: A 7-group system is interconnected with 21 optical “bundles”. The “bundles” can be configured between 2 or more cables wide, subject to the group limit.
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
Copper & Optical Cabling
Cray Software
# Cray Programming Environment Distribution

Focus on Performance and Productivity

## Programming Languages
- Fortran
- C
- C++
- Python

## Programming models
- Distributed Memory (Cray MPT)
  - MPI
  - SHMEM
- Shared Memory
  - OpenMP 3.0
  - OpenACC
- PGAS & Global View
  - UPC (CCE)
  - CAF (CCE)
  - Chapel

## Compilers
- Cray Compiling Environment (CCE)
- GNU
- 3rd Party Compilers

## Tools
- Environment setup
- Modules
- Debuggers
- DDT
- Igdb
- Debugging Support Tools
  - Abnormal Termination Processing
  - STAT

## Performance Analysis
- CrayPat
- Cray Apprentice

## Optimized Scientific Libraries
- LAPACK
- ScaLAPACK
- BLAS (libgoto)
- Iterative Refinement Toolkit
- Cray Adaptive FFTs (CRAFFT)
- FFTW
- Cray PETSc (with CASK)
- Cray Trilinos (with CASK)

## I/O Libraries
- NetCDF
- HDF5

---

Cray developed
Licensed ISV SW
3rd party packaging
Cray added value to 3rd party
Cray Software Ecosystem

Cray Linux Environment

- CrayPAT
- Cray Apprentice2
- Reveal
- Cray Iterative Refinement Toolkit
- Cray PETSc, CASK
- DVS
- GNU
- Intel
- LSTC
- accelrys
- CD-adapco
- The MathWorks
- Simulia
- Metcomp Technologies
- Altair
- slurm
- TotalView Technologies
- allinea
- TotalView Technologies
- ParaTools
- Lustre
- nag
- IMSL
- Adaptive Computing
- JOB MANAGEMENT
- DEBUGGING
- PERFORMANCE TOOLS
- IO & LIBRARIES
- COMPILERS
- APPLICATIONS
An Adaptive Linux OS optimized specifically for HPC

**ESM – Extreme Scalability Mode**
- No compromise *scalability*
- Low-Noise Kernel for scalability
- Native Comm. & Optimized MPI
- Application-specific performance tuning and scaling

**CCM – Cluster Compatibility Mode**
- No compromise *compatibility*
  - Fully standard x86/Linux
  - Standardized Communication Layer
  - Out-of-the-box ISV Installation
  - ISV applications simply install and run

CLE run mode is set by the user on a job-by-job basis to provide full flexibility
Cray Operating Systems Focus

Performance

• Maximize compute cycles delivered to applications while also providing necessary services
  • Lightweight operating system on compute node
  • Standard Linux environment on service nodes
  • Optimize network performance through close interaction with hardware
  • GPU infrastructure to support high performance

Stability and Resiliency

• Correct defects which impact stability
  • Implement features to increase system and application robustness

Scalability

• Scale to large system sizes without sacrificing stability
  • Provide system management tools to manage complicated systems
Benchmarks
Performance comparison: XE-IL vs XC30

- Typical XE-IL vs. XC-Sandybridge for 256-512 compute nodes
- Actually results will depend on system size and configurations and problem size and definition

<table>
<thead>
<tr>
<th>Tests (Units)</th>
<th>XE-Interlagos</th>
<th>XC</th>
<th>XC/XE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPL (Tflops)</td>
<td>~81%</td>
<td>~86%</td>
<td>106%</td>
</tr>
<tr>
<td>Star DGEMM (Gflops)</td>
<td>~87%</td>
<td>~102%</td>
<td>117%</td>
</tr>
<tr>
<td>STREAMs (Gbytes/s/node)</td>
<td>72</td>
<td>78</td>
<td>108%</td>
</tr>
<tr>
<td>RandomRing (Gbytes/s/rank)</td>
<td>~0.055</td>
<td>~0.141</td>
<td>256%</td>
</tr>
<tr>
<td>Point-to-Point BW (Gbytes/s)</td>
<td>2.8-5.6</td>
<td>&gt;8.5</td>
<td>157% - 314%</td>
</tr>
<tr>
<td>Nearest Node Point-to-Point Latency (usec)</td>
<td>1.6-2.0</td>
<td>&lt;1.4</td>
<td>116% - 145%</td>
</tr>
<tr>
<td>GUPs</td>
<td>2.66</td>
<td>15.6</td>
<td>525%</td>
</tr>
<tr>
<td>GFFT (Gflops)</td>
<td>628</td>
<td>2221</td>
<td>354%</td>
</tr>
<tr>
<td>HAMR Sort (GiElements/sec)</td>
<td>9.4</td>
<td>36.6</td>
<td>390%</td>
</tr>
</tbody>
</table>
WRF NewConus 2.5KM Benchmark

- Computation and halo exchange costs only

Pete Johnsen, Cray, Inc.
AcuSolve* (CFD) Performance
70M ASMO model

At 16 nodes:

- Cray XE6: 512 Interlagos cores = 331 sec
- Cray XC30: 256 Sandy Bridge cores = 187 sec

* Pre-release version of AcuSolve from Altair
LS-DYNA benchmark
Two car crash simulation, 2.4M elements, Hybrid parallel

XC30 provides significantly better per core performance
The End