Scalable by Design

The Cray XT Series of Supercomputers
Cray XT5

(Jaguarpf/Kraken/Hopper)
Scalable Software Architecture: Cray Linux Environment (CLE)
“Primum non nocere”

• Microkernel on Compute PEs, full featured Linux on Service PEs.
• Service PEs specialize by function
• Software Architecture eliminates OS “Jitter”
• Software Architecture enables reproducible run times
• Large machines boot in under 30 minutes, including filesystem

Service Partition

Specialized Linux nodes
XT System Configuration Example

- Compute node
- Login node
- Network node
- Boot/Syslog/Database nodes
- I/O and Metadata nodes

GigE
10 GigE
GigE
SMW
Fibre Channels

RAID Subsystem
### Characteristics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Cores</td>
<td>8 or 12</td>
</tr>
<tr>
<td>Peak Performance</td>
<td></td>
</tr>
<tr>
<td>Shanghai (2.4)</td>
<td>76 Gflops/sec</td>
</tr>
<tr>
<td>Istanbul (2.6)</td>
<td>124 Gflops/sec</td>
</tr>
<tr>
<td>Memory Size</td>
<td>16 or 32 GB per node</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>25.6 GB/sec</td>
</tr>
</tbody>
</table>

- **6.4 GB/sec direct connect HyperTransport**
- **25.6 GB/sec direct connect memory**
- **Cray SeaStar2+ Interconnect**
Cray SeaStar2+ Interconnect

Cray XT5 systems ship with the SeaStar2+ interconnect

- Custom ASIC
- Integrated NIC / Router
- MPI offload engine
- Connectionless Protocol
- Link Level Reliability
- Proven scalability to 225,000 cores

Now Scaled to 225,000 cores
Cray XT5 Compute Blade
XT5 Axial Turbofan – 78% Efficient
Cray XT5 Compute Blade
Cray ECOphlex
Liquid Cooling
ECOphlex Cooling

Hot air stream passes through evaporator, rejects heat to R134a via liquid-vapor phase change (evaporation).

R134a absorbs energy only in the presence of heated air.
Phase change is 10x more efficient than pure water cooling.
ECOpHlex Technology in the Cray High Efficiency Cabinet

R134a piping
Exit Evaporators
Inlet Evaporator
Newer “Flat Top” ECOphlex Design
Other Changes

- New enhanced blower to handle the 130 Watt Magny-Cours Processor
- Enhanced sound kit to reduce noise
- More efficient design
- New VFD (Variable Frequency Diode) for blower
- An upgrade kit product code will be available for existing XT5 customers which will contain the required components
Enhanced Series 6 ECOphlex Cabinet

Air taken from top, no line of sight for sound

Foam lined duct for sound absorption

Extra foam added to front. Door now seals to front IO extension
Cray XT4

(Jaguar/Athena/Franklin)
Quad-Core Cray XT4 Node

- 4-way SMP
- >35 Gflops per node
- Up to 8 GB per node
- OpenMP Support within socket

12.8 GB/sec direct connect memory
(DDR 800)

6.4 GB/sec direct connect HyperTransport

Cray
SeaStar2+ Interconnect

2 – 8 GB
Cray XT4 Compute Blade

- 4 DIMM Slots with Chipkill
- Redundant VRMs
- Blade Control Processor
- Blade Backplane Connector (>100 GB/sec)
- Embedded HyperTransport Link
Software
Cray Software Ecosystem

- Site specific
- Public Domain
- ISV Applications

**Applications**
- Compilers
- Debuggers
- Schedulers
- Tools
- OS

**Public Domain Tools**
- CrayPat
- Cray Apprentice
- Libraries
- Public Domain Tools
- Cray Linux Environment

**The Portland Group**
- PathScale

**Altair PBS Works**

**Platform**
- Adaptive Computing
- Cluster Resources
Service nodes run a full-featured SLES10 Linux installation
  - We add our tools, libraries, and services

Compute nodes run a slim-line Linux kernel with only necessary services
  - Only run what’s needed so the application can rule the roost

Libraries
  - MPT – Message Passing Toolkit
  - LibSci – Cray Scientific Libraries (BLAS, LAPACK, SCALAPACK, FFTW, etc)
  - I/O Libraries – HDF5 & NetCDF

Tools
  - Compilers – PGI, Cray, GNU, Pathscale, Intel
  - CrayPAT – Performance Analysis Tools

ALPS
  - Application placement, job launching, application clean-up
  - Users interface with ALPS primarily via aprun

PBS/TORQUE & MOAB
  - All jobs on the local XTs are batch jobs
  - MOAB is an advanced job scheduler that is used on Jaguar and Kraken
New CLE Features for 2009 / 2010

- Parallel Data Virtualization Service support
- Scalable Dynamic Libraries
- Virtual Cluster Environment
- Core Specialization for codes with high synchronization requirements
- NodeKARE (Node Knowledge and Reconfiguration) resiliency features
- Checkpoint / Restart
Mounting Other Filesystems with DVS

- Compute Node DVS Client
- Compute Node DVS Client
- Compute Node DVS Client
- Compute Node DVS Client
- Compute Node DVS Client

SeaStar Interconnect

DVS Server SIO Node 4
Panfs Client

DVS Server SIO Node 4
Stornext Client

DVS Server SIO Node 4
GPFS Client

DVS Server SIO Node 4
NFS Client

Servers & Storage

Stornext Servers & Storage

GPFS Servers & Storage

NFS Servers & Storage

Cray XT System

IB or 10GigE
Dynamic Shared Libraries

- Benefit: root file system environment available to applications
- Shared root from SIO nodes will be available on compute nodes
- Standard libraries / tools will be in the standard places
- Able to deliver customer-provided root file system to compute nodes
- Programming environment will support static and dynamic linking
- Performance impact negligible, due to scalable implementation
Scaling Shared Libraries with DVS

- Requests for shared libraries (.so files) are routed through DVS Servers
- Provides similar functionality as NFS, but scales to 1000s of compute nodes
- Central point of administration for shared libraries
- DVS Servers can be “re-purposed” compute nodes
Cray Linux Environment – Adaptive Vision

- Less Compatibility
  - Capability (Ultra-light Linux Image)
  - Capacity/Production (Mid-weight Linux Image)
  - Shrink-wrap 3rd Party Application (Full Linux Image and all services)

- Full Compatibility
  - High Scale
  - Low Scale
A Very Skinny Penguin – Core Specialization

- **Benefit:** Eliminate noise with overhead (interrupts, daemon execution) directed to a single core

- **Rearranges existing work**
  - Without core specialization: overhead affects every core
  - With core specialization: overhead is confined, giving app exclusive access to remaining cores

- **Helps some applications, hurts others**
  - POP 2.0.1 on 8K cores on XT5: 23% improvement
  - Larger jobs should see larger benefit
  - Future nodes with larger core counts will see even more benefit

- This feature is adaptable and available on a job-by-job basis
Feature Also Known As “Node Health Checker”
- Benefit: verify that nodes are healthy so that jobs are not started on unhealthy nodes, that is, improved application completion rates
- Checks more possible sources of error: file system checks, memory usage, application termination, site-specific check
- Configurable: when to run, what to do on errors, callout to site-specific script
- Suspect Mode minimizes burden on administrator
- Future release will dump and restart downed nodes

Checkpoint / restart
- Released in CLE 2.2 (Jul 09)
- Supported by PBS Pro (10.1 or later) and Moab/Torque
- MPI and SHMEM
Cray continues the partnership with PGI to provide compilers on XT

Cray Compilation Environment
- UPC implementation
- Co-Array implementation
- Smooth transition to Cascade
- Laying support for integrating accelerators

Intel compiler also available for XT systems
Cray acquired Pathscale Technology
Support for dynamic libraries and ISV codes
Thank You!