Cray XT and Cray XE
System Overview
Overview Topics

• System Overview
  – Cabinets, Chassis, and Blades
  – Compute and Service Nodes
  – Components of a Node
    ▪ Opteron Processor
    ▪ SeaStar ASIC
      • Portals API Design
    ▪ Gemini ASIC

• System Networks

• Interconnection Topologies
Cray XT System
The cabinet contains three chassis, a blower for cooling, a power distribution unit (PDU), a control system (CRMS), and the compute and service blades (modules).

All components of the system are air cooled.

- A blower in the bottom of the cabinet cools the blades within the cabinet.
  - Other rack-mounted devices within the cabinet have their own internal fans for cooling.

- The PDU is located behind the blower in the back of the cabinet.
The system contains two types of blades:

- Compute blades
  - 4 SeaStar or 2 Gemini ASICs
  - 4 nodes

- Service blades
  - SIO (XT systems)
    - 4 SeaStar ASICs
    - 2 nodes
    - One dual-slot PCI-X or PCIe riser assemblies per node
  - XIO (XE systems)
    - 2 Gemini ASICs
    - 4 nodes
    - One single slot PCIe riser per node (except on the boot node)
PCI Cards

• Cray supports the following types of PCIe cards:
  – Gigabit Ethernet
  – 10Gigabit Ethernet
  – Fibre Channel (FC2, FC4, and FC8)
  – InfiniBand
Node Components

- Opteron processor
  - Single-, dual-, quad-, six-, eight-, twelve-core versions
- Node memory
  - 2- to 8-GB of DDR1 memory in Cray XT3 compute nodes
  - 2- to 8-GB of DDR1 memory in Cray XT service (SIO) nodes
  - 4- to 16-GB of DDR2 memory in Cray XT4 compute nodes
  - 4- to 32-GB of DDR2 memory in Cray XT5 compute nodes
  - 8- to 64-GB of DDR3 memory in Cray XT6 compute nodes
  - 8- to 64-GB of DDR3 memory in Cray XE6 compute nodes
  - 8- to 16-GB of DDR2 memory in Cray XE6 service (XIO) nodes
- SeaStar or Gemini ASIC
SIO Service Blade
SeaStar ASIC

- Direct HyperTransport connection to Opteron
- The DMA engine transfers data between the host memory and the network
  - Separate read and write sides in the DMA engine
  - The Opteron does not directly load/store to network
  - On the receive side a set of content addressable memory (CAM) locations are used to route incoming messages
    - There are 256 CAM entries
- Designed for Sandia Portals API
SeaStar Diagram

- Core
- Northbridge
- DDR-SDRAM interface
- Memory
- HyperTransport Interface
- HyperTransport Interface
- SeaStar DMA engine
- Hyper Transport
- Memory
- SSI
- PPU
- L0 controller
- To PCI-X on Service blades

10/18/2010 Cray Private
Node Connection Bandwidths

- **Opteron core(s)**
- **HyperTransport**
- **Memory**
- **Cray XT3**
  - 6.4 GB/s bidirectional
  - 2.17 GB/s sustained
- **Cray XT4**
  - 8.0 GB/s bidirectional
  - 9.6 GB/s
  - 8.5 GB/s sustained
  - 50 ns latency - measured

**SeaStar**
- High Speed Network
  - 9.6 GB/s bidirectional
  - 6.5 GB/s sustained

**Cray XT3 DDR1**
- 6.4 GB/s (400 MHz DIMMs)
- 5.7 GB/s sustained
- 51.6 ns latency - measured

**Cray XT4 DDR2**
- 10.4 GB/s (667 MHz DIMMs)
- 8.5 GB/s sustained
- 50 ns latency - measured

**Cray XT5 DDR2**
- 12.8 GB/s (800 MHz DIMMs)
- 10 GB/s (667 MHz DIMMs)
- 12.8 GB/s (800 MHz DIMMs)
Portals API Design

- Designed to message among thousands of nodes
  - Performance is critical only in terms of scalability
  - Success is measured by how large an application is allowed to scale, not by a 2-node ping-pong time
- Designed to avoid scalability limitations
  - Is network independent
  - Bypasses OS – no memory copies into or out of the kernel; few interrupts
  - Bypasses application – does not require activity by the application to ensure progress
  - Reliable, ordered delivery of messages between two nodes
Portals API Design

- Receiver (target) managed – no reserved memory for thousands of potential senders
  - Is connectionless; no explicit connection is established with other processes in the application
    - When connected, maintains a minimum amount of state
  - The target process determines how to respond to the incoming message
    - The target node can choose to accept or ignore a message from any node
    - Destination of any message is not an address
      - Instead “Match bits” enable the receiver to place it
  - Unexpected messages are handled by the unexpected message queue
  - All buffers are in user space
X6 Compute Blade with Gemini

Node 0

Node 1

Node 2

Node 3

Gemini 0

Gemini 1
XIO Service Blades

• Four nodes; each node contains:
  – One AMD Opteron processor with up to 16 GB of DDR2 memory
    ▪ Processor is a six-core Opteron
  – A connection to a Gemini ASIC
  – Voltage regulating modules (VRMs)

• L0 controller

• Gemini mezzanine card
  – Contains two Gemini ASICs

• Four PCIe risers
  – One riser per node
XIO Blade

node 3 connects to the lower mezzanine

node 1 connects to the lower mezzanine
XIO Riser Configuration

Upper bay

- Inner mezzanine
  - 75 Watt slot
  - 225 Watt slot
- Outer mezzanine
  - 225 Watt slot

Lower bay

- Inner mezzanine
  - 75 Watt slot
  - 225 Watt slot
- Outer mezzanine
  - 225 Watt slot

PCIe G2 bridge
- x16

Kernel DDR2 memory
- x16

HyperTransport

Opteron node 0
- 75 Watt slot
- 225 Watt slot

Opteron node 1

Opteron node 2

Opteron node 3

HyperTransport

Gemini

10/18/2010
Cray Private
XIO Boot Node Configuration

Upper bay

Inner mezzanine → PCIe G2 bridge → HyperTransport → DDR2 memory → HyperTransport → Opteron node 3

Opteron node 3 → HyperTransport → Gemini

Opteron node 2 → HyperTransport → Gemini

Opteron node 1 → HyperTransport → Gemini

Opteron node 0 → HyperTransport → Gemini

Lower bay

Inner mezzanine → PCIe G2 bridge → HyperTransport → DDR2 memory → HyperTransport → Opteron node 1

Opteron node 1 → HyperTransport → Gemini

Opteron node 2 → HyperTransport → Gemini

Opteron node 3 → HyperTransport → Gemini

Outer boot mezzanine → PCIe G2 bridge → HyperTransport → DDR2 memory → HyperTransport → Opteron node 2

Opteron node 2 → HyperTransport → Gemini

Opteron node 3 → HyperTransport → Gemini

Opteron node 0 → HyperTransport → Gemini

PCIe G2 bridge

Inner mezzanine

Outer mezzanine

DDR2 memory

x16

x16

x16

x8

HyperTransport

HyperTransport

HyperTransport

HyperTransport

HyperTransport

HyperTransport

HyperTransport

HyperTransport

10/18/2010

Cray Private
Opteron Processors

AMD Opteron single-core processor
Cray XT3 System
Socket 940: DDR1 buffered DIMMs and 3 HyperTransports

AMD Opteron dual-core processor
Cray XT3 System
Socket 940: DDR1 buffered DIMMs and 3 HyperTransports
Cray XT4 System
Socket AM2: DDR2 unbuffered DIMMs and one HyperTransport

AMD Opteron quad-core processor
Cray XT4 System
Socket AM2: DDR2 unbuffered DIMMs and one HyperTransport
Cray XT5 System
Socket F: DDR2 registered (buffered) DIMMs and 3 HyperTransports

AMD Opteron hex-core processor
Opteron Processors

AMD Opteron six-core processor
Cray XT5 System
Socket F: DDR2 unbuffered DIMMs and 3 HyperTransports
L3 Cache: 6 MB

AMD Opteron twelve-core processor
Dual-die socket
(for an eight-core socket remove two CPUs from each side)
Cray XT6 and Cray XE6 Systems
Socket G34: DDR3 unbuffered DIMMs and 4 HyperTransports
L3 cache: each die has 6 MBs, 12 MB per socket
Gemini

- Gemini is designed to pass data to and from the network with less control
  - Gemini is suited for SMP
  - Gemini allows for adaptive routing
  - Hardware support for PGAS languages
    - Remote memory address translations, as with Cray X2 systems
  - Atomic memory operations
- The Gemini chip uses HyperTransport version 3
Terminology

• BTE – Block Transfer Engine
• DMAPP – Distributed Memory Application
• FMA – Fast Memory Access
• GHAL – Generic Hardware Abstraction Layer
• GNI – Generic Network Interface
• PGAS – Partitioned Global Address Space
  – Programming language extensions such as Unified Parallel C (UPC) and Co-Array Fortran (CAF)
Performance and Resiliency Features

• Automatic link-level and HT3 retries
  – HT3 supports an improved CRC

• Congestion feedback to enable routing around network bottlenecks
  – Packets are reordered in receive buffer
    ▪ Separate completion notification when all are stored

• Automatic link failure handling with software help
  – To route around a failed link, system traffic must be quiesced to prevent deadlock and preserve ordering
    ▪ This feature is also used to support hot blade swap

• Router errors are detected and reported at the point of error
Gemini ASIC Diagram

HT3

NIC 0

NIC 1

Netlink block

48-port YARC router

Tile 0, coordinate 0,0

Tile 7, coordinate 0,7

Tile 47, coordinate 5,7
Node Types

- **Compute nodes**
  - No user login
  - Only for execution of parallel applications
  - Diskless - no permanent storage
  - CNL operating system
    - Linux symmetrical multi-processing (SMP)
    - Supports OpenMP programming model
    - `apinit` part of ALPS (Application Level Placement Scheduler)

- **Service Nodes**
  - Provide services based on hardware/software configuration
  - Node names: boot node, SDB, syslog, login, I/O, and network nodes
  - Run Linux operating system - SLES
Service Nodes

- **Boot node**
  - First node booted
  - Has its own root file system
  - Exports the “shared root” file system to the other service nodes

- **System database (SDB) node**
  - Stores system configuration information and system state
  - Implemented with MySQL Pro

- **syslog node**
  - Collects syslog data from other service nodes

- **Login nodes**
  - Provide users with a familiar Linux environment
  - Edit, compile, submit, and monitor interactive or batch jobs
    - PBS Pro is the standard batch system
Service Nodes

- I/O nodes
  - Attach to RAID storage devices
  - Perform remote DMA I/O through the HSN
  - Function as a metadata server (MDS) and object storage servers (OSSs) for Lustre file systems

- Network nodes
  - 10 Gigabit Ethernet (10GbE) adapter
  - Connect to other file servers
Service Node Software Components

- ALPS - Application Level Placement Scheduler
  - Application placement, launch and management for CNL
  - Includes several daemons and client processes

- RCA – Resiliency Communication Agent
  - Part of the kernel, on all nodes
  - Sends a heartbeat message to the SMW (CRMS/HSS/CMS)
  - Detects and responds to failing application launchers
The High-speed Network (HSN) or interconnect network
- Connects the service and compute nodes
  - The preferred topology is a folded torus
    - A torus is a circle
    - Folding enables the use of the shortest cables possible to connect all nodes within a dimension
    - Depending on the class, the topology may be a mesh
  - The class of the system governs the configuration
- Applications move data across the HSN

The Hardware Supervisory System (HSS) network
- Monitors and controls the system
- Ethernet network with the System Management Workstation (SMW) is at the top of the network
Interconnect Network Topology Classes

• Configurations are based on topology classes:
  – Class 0 contains 1 to 9 chassis (1 to 3 cabinets)
    ▪ 1 row
  – Class 1 contains 4 to 16 cabinets
    ▪ 1 row
  – Class 2 contains 16 to 48 cabinets
    ▪ 2 equal length rows
  – Class 3 contains 48 to 320 cabinets
    ▪ Three rows of equal length
      • A mesh in one or more dimensions
    ▪ Even number of 4 or more equal length rows
      • Torus in all dimensions
A Single-chassis Configuration

• The basic building block is a single chassis
  – A chassis is 1 x 4 x 8
    ▪ Dimensions are: X x Y x Z
  – Each node on a blade is connected in the Y dimension (mezzanine)
  – Each node in a chassis is connected in the Z dimension (backplane)
  – All X-dimension connections are cables
Class 0 Topology
Class 0 Cable Drawing, 3 Cabinets

Three Cabinets
(9 x 4 x 8)
Class 1 Topology
Class 1 Cable Drawing, 4 Cabinets

Four Cabinets
(4 x 12 x 8)
Class 2 Topology
Class 2 Cable Drawing, 16 Cabinets

16 Cabinets
2 Rows of 8
(8 x 12 x 16)
Class 3 Topology
Cray XT5m

- Cray XT5m topology
  - 1 to 18 chassis, 1 to 6 cabinets
  - All cabinets are in a single row
  - The system scales in units of chassis
  - 2D torus topology (y and z)
    - \((Y \times 4) \times 8\) topology
    - \(Y\) is the number of populated chassis in the system
Identifying Components

- System components are labeled according to physical ID (HSS Identification), node ID, IP address, or class

<table>
<thead>
<tr>
<th>Component</th>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>s0, all</td>
<td>All components attached to the SMW.</td>
</tr>
<tr>
<td>Cabinet</td>
<td>cX-Y</td>
<td>Cabinet number and row; this is the L1 host name.</td>
</tr>
<tr>
<td>Cage</td>
<td>cX-Yc#</td>
<td>Physical cage in cabinet: 0, 1, 2. Cages are numbered from bottom to top.</td>
</tr>
<tr>
<td>Blade or slot</td>
<td>cX-Yc#s#</td>
<td>Physical blade slot in cage: 0 – 7, numbered from left to right; this is the L0 hosts name.</td>
</tr>
<tr>
<td>Node</td>
<td>cX-Yc#s#n#</td>
<td>Opteron Chip on a blade: 0 - 3 for compute blades, 0 and 3 for SIO blades</td>
</tr>
<tr>
<td>SeaStar ASIC</td>
<td>cX-Yc#s#s#</td>
<td>Cray SeaStar ASIC on a blade: 0 – 3</td>
</tr>
<tr>
<td>Gemini ASIC</td>
<td>cX-Yc#s#g#</td>
<td>Cray Gemini ASIC on a blade: 0 – 1</td>
</tr>
<tr>
<td>Link</td>
<td>cX-Yc#s#s#1#</td>
<td>Link port of a SeaStar ASIC: 0 – 5</td>
</tr>
<tr>
<td></td>
<td>cX-Yc#s#g#1#</td>
<td>Link port of a Gemini ASIC: 0 - 57</td>
</tr>
</tbody>
</table>
Node ID (NID)

- The Node ID is a unique hexadecimal or decimal number that identifies each CLE node
  - The NID reflects the node location in the network
    - In a SeaStar based system, the NIDs are assigned on 128-number boundaries for each cabinet
    - In a Gemini based system, NIDs are sequential
  - NID format is `nidnnnnn` (decimal) when it refers to a hostname of a node, for example: `nid00003`
### Cabinet 0 NID Numbering Example

<table>
<thead>
<tr>
<th>Cray XT (SeaStar based) system</th>
<th>Cray XE (Gemini based) system</th>
</tr>
</thead>
<tbody>
<tr>
<td>6, 18, 20, 22, 24, 26, 28, 30</td>
<td>6, 8, 10, 12, 14, 16, 18, 20</td>
</tr>
<tr>
<td>12, 14, 16, 18, 20, 22, 24, 26</td>
<td>2, 4, 6, 8, 10, 12, 14, 16</td>
</tr>
<tr>
<td>28, 30, 32, 34, 36, 38, 40, 42</td>
<td>4, 6, 8, 10, 12, 14, 16, 18</td>
</tr>
<tr>
<td>42, 44, 46, 48, 50, 52, 54, 56</td>
<td>6, 8, 10, 12, 14, 16, 18, 20</td>
</tr>
<tr>
<td>58, 60, 62, 64, 66, 68, 70, 72</td>
<td>8, 10, 12, 14, 16, 18, 20, 22</td>
</tr>
<tr>
<td>74, 76, 78, 80, 82, 84, 86, 88</td>
<td>20, 22, 24, 26, 28, 30, 32, 34</td>
</tr>
<tr>
<td>92, 94, 96, 98, 100, 102, 104, 106</td>
<td>36, 38, 40, 42, 44, 46, 48, 50</td>
</tr>
</tbody>
</table>

10/18/2010 Cray Private
External Services for Cray Systems

- Implementation of services external to the Cray XT and Cray XE systems to address customer requirements
  - More flexible user access
  - More options for data management, data protection
  - Leverage commodity components in customer-specific implementations
  - Provides faster access to new devices and technologies
  - Repeatable solutions that remain open to custom configuration
  - Each solution can be used, scaled, and configured independently
Specifically

- esFS
  - Provides globally shared data between multiple systems
    - Cray systems and others
  - Provides access to other file systems
  - Data Virtualization Service (DVS) is used to project Panasas or StorNext to the compute nodes

- esLogin
  - Less dependence on a single Cray system for data access and compiles
  - An enhanced user environment
    - larger memory, swap space, and more horsepower
    - Increased availability of data and system services to users

- esDM
  - More options for data management and data protection