Cray XC System Node Diagnosability

Jeffrey J. Schutkoske
Platform Services Group (PSG)
jjs@cray.com
This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.
Overview

- Introduction to System Diagnosability
- Node Level Diagnosability
  - Initialization & Hardware Errors
  - Performance Errors
  - Out-Of-Band Diagnosis & Debug
- Q&A
What is the System Diagnosability?

- System Diagnosability is a suite of software tools
- Diagnostics are just one aspect of the toolset
- Features built into SMW and CLE commands
- System Diagnostics validate hardware and software
- System Diagnostics can be periodically scheduled
System Diagnostics

- Reporting: Report and log errors, warnings, and faults
- Workload: Simulate customer workload
- Performance: Measure component performance
- Stress: Maximize hardware stress
- Confidence: Validate individual functionality
- Boot: Performed prior to booting CLE
On-line Diagnostic Execution

- On-line diagnostics are installed with CLE
  - Node/Aries: /opt/cray.diag/default
  - GPU: /opt/cray/cray-nvidia/default
  - KNC: /opt/cray/cray-intel/default

- Submit jobs through the batch or interactive mode

  aprun -n 2 -N 1 -L 28,29 ./xtfma_ata -R 2
Node Initialization

- BIOS initializes the Intel processor & memory
- BIOS also discovers, initializes, and trains
  - QPI bus
  - Aries PCIe bus
  - Nvidia GPU or Intel Co-Processor PCIe bus
  - I/O card PCIe bus
- Reports link width, speed, and status
- BIOS logs are copied to the SMW on failure
BIOS Aries Initialization Example

- **Aries Detected**
  
  Aries NIC detected @ B1|D0|F0, RevId=0x10

- **Aries Initialized**
  
  Aries NIC [B1|D0|F0] initialized, Width: x16, CurSpd: 8.0 Gbps

- **Aries Trained Successfully**
  
  Aries (B0:D2:F0) completed all PCIe Gen3 Phases successfully, LNKSTS2=0x1f
BIOS Aries Initialization Failure Examples

- **Aries Not at PCIe Gen 3 Speed**
  
  *Aries (B3:D0:F0) not running at Gen3, PCIe Phase 1 did not complete, LNKSTS=0xx, LNKSTS2=0xx*

- **Aries PCIe Link training failed**
  
  *Aries (B3:D0:F0) link training failed, LNKSTS=0xx*

- **Aries Not at PCIe Gen 3 Speed**
  
  *Aries (B3:D0:F0) did not train to Gen3, LNKSTS=0xx*
BIOS PCIe Initialization Failure Example

- xtbounce indicates a PCIe Link Speed mismatch

```
***** node_up *****
ERROR: c0-0c0s15n0 - 370 - SXM (GPU) PCIe link speed mismatch
```

- Ensures device functionality at system boot time
NODE HARDWARE ERRORS

- The CLE kernel captures node hardware errors
- CLE Kernel console log
- CLE Kernel sends the errors to HSS via RCA
- Hardware Error Log Channel connected to the BC
- Logs and Errors saved on the SMW
## Xthwerrlog DIMM Output Example

<table>
<thead>
<tr>
<th>Node</th>
<th>Count</th>
<th>Bank</th>
<th>Type</th>
<th>DIMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1-0c0s1n0</td>
<td>1</td>
<td>8</td>
<td>CORRECTABLE</td>
<td>J10</td>
</tr>
<tr>
<td>c1-0c0s1n1</td>
<td>16</td>
<td>9</td>
<td>CORRECTABLE</td>
<td>J7</td>
</tr>
<tr>
<td>c1-0c0s7n2</td>
<td>50</td>
<td>9</td>
<td>CORRECTABLE</td>
<td>J11</td>
</tr>
<tr>
<td>c1-0c0s7n2</td>
<td>1</td>
<td>10</td>
<td>CORRECTABLE</td>
<td>J12</td>
</tr>
<tr>
<td>c1-0c1s1n1</td>
<td>24</td>
<td>9</td>
<td>CORRECTABLE</td>
<td>J11</td>
</tr>
</tbody>
</table>
Advanced Error Reporting (AER)

- Enabled in the CLE kernel by default
  - Aries
  - Nvidia GPU
  - Intel KNC

```
c0-0c0s7a0n1  CorrectableMemErr  0:0:0
  AER Correctable: Non-fatal error (mask bit: 1)

c0-0c1s6a0n2  CorrectableMemErr
  Link CRC error (cnt: 3)
```
Node Performance

- Node validated using DGEMM
- Validates performance and data miscompares

<table>
<thead>
<tr>
<th>Cname</th>
<th>c0-0c0s7n0</th>
</tr>
</thead>
<tbody>
<tr>
<td>NID</td>
<td>nid00028</td>
</tr>
<tr>
<td>Iteration</td>
<td>0</td>
</tr>
<tr>
<td>GFlops</td>
<td>24.1141</td>
</tr>
<tr>
<td>Power (W)</td>
<td>89</td>
</tr>
<tr>
<td>Processor actual</td>
<td>515.110009109461</td>
</tr>
<tr>
<td>Processor expected</td>
<td>514.110009109461</td>
</tr>
</tbody>
</table>

- Supported on Node, GPU, and KNC
# Node Performance Example

<table>
<thead>
<tr>
<th>Node</th>
<th>GFlops Min</th>
<th>GFlops Max</th>
<th>GFlops Avg</th>
<th>Bin</th>
<th>Eff Bin</th>
</tr>
</thead>
<tbody>
<tr>
<td>c0-0c2s8n0</td>
<td>503.813</td>
<td>506.223</td>
<td>505.066</td>
<td>-332.98</td>
<td>-77.2488</td>
</tr>
<tr>
<td>c0-0c2s9n0</td>
<td>501.591</td>
<td>504.382</td>
<td>503.217</td>
<td>-341.659</td>
<td>-87.0555</td>
</tr>
<tr>
<td>c0-0c2s9n1</td>
<td>495.502</td>
<td>498.846</td>
<td>497.501</td>
<td>-365.447</td>
<td>-113.934</td>
</tr>
<tr>
<td>c0-0c2s8n1</td>
<td>489.865</td>
<td>493.652</td>
<td>492.368</td>
<td>-387.466</td>
<td>-138.814</td>
</tr>
<tr>
<td>c0-0c2s8n2</td>
<td>477.858</td>
<td>481.309</td>
<td>480.039</td>
<td>-434.366</td>
<td>-191.809</td>
</tr>
<tr>
<td>c0-0c2s8n3</td>
<td>477.336</td>
<td>479.782</td>
<td>478.613</td>
<td>-436.406</td>
<td>-194.114</td>
</tr>
<tr>
<td>c0-0c2s9n3</td>
<td>473.895</td>
<td>477.102</td>
<td>475.889</td>
<td>-449.848</td>
<td>-209.302</td>
</tr>
<tr>
<td>c0-0c2s9n2</td>
<td>472.54</td>
<td>474.801</td>
<td>473.787</td>
<td>-455.14</td>
<td>-215.283</td>
</tr>
</tbody>
</table>
Aries HSN Performance

- Aries All-To-All performance test, *xta2a*
- Measures performance on all-to-all communication

<table>
<thead>
<tr>
<th>Bytes</th>
<th>Min (GB/s)</th>
<th>Mean (GB/s)</th>
<th>Max (GB/s)</th>
<th>Dev</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>4096</td>
<td>5457</td>
<td>5603</td>
<td>5668</td>
<td>56</td>
<td>1.0%</td>
</tr>
<tr>
<td>4096</td>
<td>3626</td>
<td>4887</td>
<td>4891</td>
<td>74</td>
<td>1.5%</td>
</tr>
<tr>
<td>4096</td>
<td>3944</td>
<td>4916</td>
<td>4918</td>
<td>74</td>
<td>1.5%</td>
</tr>
<tr>
<td>4096</td>
<td>4068</td>
<td>4916</td>
<td>4918</td>
<td>74</td>
<td>1.5%</td>
</tr>
<tr>
<td>4096</td>
<td>3617</td>
<td>4915</td>
<td>4919</td>
<td>84</td>
<td>1.7%</td>
</tr>
</tbody>
</table>
Aries HSN Performance Results Analysis

Bandwidth low for set 994 nodes 4056 4059: 3617 GB/s

Bandwidth low for set 1051 nodes 4288 4291: 4012 GB/s
Out-Of-Band Diagnosis

- Validate the HSS hardware and software
- HSS diagnostic utility, *xtcheckhss*
  - Cabinet
  - Blade
    - Aries Network Card (ANC)
    - Processor Daughter Card (PDC)
  - Node
  - GPU
  - KNC
### xtcheckhss Example

**xtcheckhss** --volts --blade=c0-0c0s7

<table>
<thead>
<tr>
<th>Component:</th>
<th>c0-0c0s7n2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module:</td>
<td>qpdc0_n0_s0_mem_vrm</td>
</tr>
<tr>
<td>Sensor:</td>
<td>vdd_vdr01_s0_c_i</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HMIN</th>
<th>SMIN</th>
<th>Data</th>
<th>Unit</th>
<th>SMAX</th>
<th>HMAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200</td>
<td>1350</td>
<td>1339</td>
<td>V*1000</td>
<td>1560</td>
<td>1800</td>
</tr>
</tbody>
</table>
Intel In-Target Probe (ITP) Debug

- Intel In-Target Probe (ITP) is a JTAG bus
- Scripts reside on the SMW ➔ xtitep
- Provide useful hardware and software debug information
  - PCIe configuration and status
  - QPI configuration and status
  - Processor information, MCA errors, and MSR data
  - Package Power Limit (turbo) registers
- Executing this command on the SMW temporarily pauses the processor
ITP Debug Example

```
xtitp -t c0-0c0s7 qpi-status 1
```

**Socket 0**

**QPI0:**

- Link Speed: 8.0 GT/s
- Configured Tx Width: Full
- Configured Rx Width: Full
- Tx Lane Status: 0xffffffff
- Rx Lane Status: 0xffffffff
- Error Counter 0: 0
- Error Counter 1: 0
Summary

✓ System Diagnosability Overview
✓ Node Level Diagnosability
  ✓ Node Initialization & Hardware Errors
  ✓ Performance Errors – Processor & Aries
  ✓ HSS At Scale Out-Of-Band Diagnosis & Debug
Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

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

Cray hardware and software products 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.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, URIKA, and YARCDATA. The following are trademarks of Cray Inc.: ACE, APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.