Cray® XC40™ System Diagnosability
Jeffrey J. Schutkoske (jjs@cray.com)
Introduction

● General System Diagnosability Review

● Diagnosability Enhancements
  ● Specifically focused on the Intel® Xeon Phi™ CPU 7250 processor

▪ Q&A
System Diagnosability Is…

- More than just Diagnostics
- Suite of software tools
- Built into SMW and CLE commands
System Diagnosability Review

Initialize

Monitor

Diagnose

Notify

Analyze
Initialize
Node Initialization

1. Request Config
2. Set Config
3. Store Config
4. Stage Config
5. Request Reinit
6. Read Config
7. Boot Config
Node Initialization Analysis

- Failures during reboot initiated by CAPMC logged in:
  - `/var/opt/cray/log/xtremoted-YYYYMMDD`
- `xtremoted` logs the full `xtbounce` output when `xtbounce` returns non-zero
- `xtremoted` logs the full `xtcli` boot output regardless of return code
- BIOS detected errors are logged to the BIOS logs
  - BIOS logs are forwarded to the SMW using LLM
** BIOS Initialization Analysis **

** ED C-0 Memory Init: cmdcrc_err = 1  
** ED C-4 Memory Init: cmdcrc_err = 1  
EDC Meminit Time Elapsed: 99ms  
EDC-0: memory Init Status 0x0003

LPC_SCRATCH_FAULT_REPORT_ENTRY
FaultNum: 1  
Type: 9  
Flags: 0x00  
CodeMajor: 0xA1  
CodeMinor: 0x06  
ApicId: 0x00  
CpuNum: 0  
LogData: 0x0000FFFF  
FaultMsg: CRAY_MCDRAM_WARNING  
A warning has been logged! Warning Code = 0xA1, Minor Warning Code = 0x6, Data = 0xFFFF

** BIOS error detected **

Cray BIOS error reported to HSS
Monitor
Node Error Monitoring

- CLE kernel captures node hardware errors
  - AER enabled in CLE by default
  - Logged in the node console log
  - Written to the Hardware Error Channel
- HSS reads the errors from the Hardware Error Channel
- *xthwerrlog* displays the hardware errors on the SMW

```
HWERR[c1-0c2s14n1] 0xfd0b:
   Uncorrectable: MFG[0]: CPUID[50671] SOCKET[0] APIC[0]:
   BANK[11]: STATUS[0xf6000800040009e]: MISC[0x0]:
   ADDR[0x153fffc300]: CTL2[0x0]
```
### Node Error Monitoring Analysis Decode

**Bank 16: IMC1: Integrated Memory Controller 1**

- **MCA Status = 0x8400004000800c0:**
  - MCACOD = 0x00c0
  - MSCOD = 0x0008
  - Other Info = 0x00
  - Corrected Error Count = 1

**Common Status Info:**
- **VALID = 1** = Valid Error Detected
- **OVER = 0** = No overflow
- **UC = 0** = Error Corrected by HW

- **EN = 0** =
- **MISCV = 0** =
- **ADDRV = 1** = Error address in MCi_ADDR

**Model-specific error: Correctable Patrol Scrub**

- Channel: 0
- MCA: Undefined Error
Node Power and Temp Monitoring

- SEDC monitors system health
- `xtgetsecdcvalues` returns the available SEDC values
- Query SEDC data from the PMDB

```
SELECT value FROM pmdb.bc_sedc_data WHERE bc_sedc_data.id where (sensor_id >= 1300 and sensor_id <= 1306)
```

<table>
<thead>
<tr>
<th>Node</th>
<th>Sensor ID</th>
<th>Sensor Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1-0c0s10</td>
<td>1306</td>
<td>BC_T_NODE3_CPU0_TEMP</td>
<td>36</td>
</tr>
<tr>
<td>c1-0c0s10</td>
<td>1304</td>
<td>BC_T_NODE2_CPU0_TEMP</td>
<td>34</td>
</tr>
<tr>
<td>c1-0c0s10</td>
<td>1302</td>
<td>BC_T_NODE1_CPU0_TEMP</td>
<td>36</td>
</tr>
<tr>
<td>c1-0c0s10</td>
<td>1300</td>
<td>BC_T_NODE0_CPU0_TEMP</td>
<td>38</td>
</tr>
</tbody>
</table>
Diagnose
Node Diagnose

- Out-of-band supported by hierarchy of controllers
- In-band used sparingly based on specific requirements
Node OOB Diagnose - xtcheckhss

- **xtcheckhss** reports the component, sensor, data, and unit for all detailed telemetry data
  - HLMIN – Hardware Limit Minimum
  - SLMIN – Software Limit Minimum
  - SLMAX – Software Limit Maximum
  - HLMAX – Hardware Limit Maximum

<table>
<thead>
<tr>
<th>HLMIN: c0-0c0s7n2_qpdc0_n0_s0_mem_vrm</th>
<th>SLMIN: vdd_vdr01_s0_c_i_v*1000</th>
<th>Actual Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200</td>
<td>1350</td>
<td>1339</td>
</tr>
<tr>
<td>1650</td>
<td>1800</td>
<td></td>
</tr>
</tbody>
</table>
Node OOB Diagnose - xtitp

- Embedded ITP used as a processor debug tool
- Scripts provide useful hardware debug information
  - PCIe config and status (Aries, SSD, etc.)
  - Processor MCA errors and MSR data
- **WARNING:** Temporarily pauses the node

```
xtitp -t c0-0c0s13 mca-error-check-all 2
MCA found in bank 14, socket 0, core 0
  IA32_MC14_STATUS = 0xf4000400040009e
  IA32_MC14_ADDR = 0x18be8bcb0
```
Node In-Band Diagnose

- New set of online diagnostics
- *xtphiperf* – Computationally intensive processor test
- *xtphimemory* – Targets DDR and MCDRAM memory
- *xtphinuma* – Validates the NUMA capabilities of node
- *xtphinls* – Stress test for the nodes
- *xtphicheck* – Gathers basic information about nodes
Node In-Band Diagnose - xtphiperf

- Targets DDR4, MCDRAM, or both
- Outputs the performance, power, and temperature
- Outputs actual and expected values on failure

<table>
<thead>
<tr>
<th>CNAME</th>
<th>Iteration</th>
<th>GFLOPS</th>
<th>Power</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:52:20, c0-0c2s12n2, nid00178</td>
<td>3</td>
<td>2039.3</td>
<td>197.806</td>
<td>44</td>
</tr>
<tr>
<td>13:52:21, c0-0c2s12n2, nid00178, Failed:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPU actual: 502.630097504761,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPU expected: 502.63009941210</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In-Band Diagnose – Workload Test Suite (WTS)

- **xtsystest** - Control script for WTS
  - Executes pre-compiled, pre-configured benchmarks
  - Executes diagnostics

- **Recommendation: Larger systems with multiple rows**
  - Execute an instance per row
  - Launch each instance from a different login node

- **Used in Cray Manufacturing**

- **Supports ALPS and SLURM**
  - SLURM Patch set for CLE 6.0 UP02, CLE 6.0 UP03, and CLE 6.0 UP04
Notify
Notify

- Utilize Open Source Simple Event Correlator (SEC) package
- Interfaces to email and System Snapshot Analyzer (SSA)
- Alerts and alarms trigger appropriate rules
  - Detect excessive cabinet power draw
  - Cabinet EPO and environmental alerts
  - Node memory errors
  - Aries PCIe link change
  - RDMA timeout
  - Gets ALPS Process ID (APID) on job failures
  - DataWarp SSD reaches 90% of its life
Analyze
Analyze

- `xtcheckhss` reports the PCIe attached SSD cards.
- `xtcheckhss` reports the targeted and trained PCIe speed and width

<table>
<thead>
<tr>
<th>Slot</th>
<th>CNAME</th>
<th>Device Name</th>
<th>Target Speed</th>
<th>Target Width</th>
<th>Trained Speed</th>
<th>Trained Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>c0-0c2s0n0</td>
<td>0</td>
<td>Samsung_SM951_M.2_SSD</td>
<td>Gen2</td>
<td>x4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CNAME 0 Samsung_SM951_M.2_SSD Gen2 x4 Gen2 x4
Summary
Summary

● Support for the KNL
  ● Initialize
  ● Monitor
  ● Diagnose
  ● Notify
  ● Analyze

● Enhanced existing tools to support the KNL
● Created new tools for the KNL
● Capture failure data the first time
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, and URIKA. The following are trademarks of Cray Inc.: APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, REVEAL, 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.