Detecting and Managing GPU Failures

Cray User Group 2015
Nicholas P. Cardo, CSCS
April 29, 2015
Theme

“Strive for perfection in everything. Take the best that exists and make it better. If it doesn’t exist, create it. Accept nothing nearly right or good enough.”

Sir Henry Royce
The Swiss National Supercomputing Centre, develops and provides the key supercomputing capabilities required to solve important problems to science and/or society. The Centre enables world-class research with a scientific user lab that is available to domestic and international researchers through a transparent, peer-reviewed allocation process.
GPUs

Power

Flops

GPUs
Piz Daint – XC30

- 5,272 Compute Nodes
  - 8-core E5-2670 2.6 Ghz
  - 32 GB DDR3
  - NVIDIA K20x GPU
    - 6 GB GDDR5
- 52 Service Nodes
- 1.5 TFlops/Node
- 7.8 PFlops/System

Top500: #6  Green500: #9
Failure

- Node Level Diagnostics – *good*
- GPU diagnostics – *lacking*
- Need to detect GPU related issues – *before users*
The following table lists the Xid errors along with the potential causes for each.

<table>
<thead>
<tr>
<th>XID Failure</th>
<th>Causes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HW Error</td>
</tr>
<tr>
<td>1  Invalid or corrupted push buffer stream</td>
<td>X</td>
</tr>
<tr>
<td>2  Invalid or corrupted push buffer stream</td>
<td>X</td>
</tr>
<tr>
<td>3  Invalid or corrupted push buffer stream</td>
<td>X</td>
</tr>
<tr>
<td>4  Invalid or corrupted push buffer stream</td>
<td>X</td>
</tr>
<tr>
<td>5  GPU semaphore timeout</td>
<td>X</td>
</tr>
<tr>
<td>6  Unused</td>
<td>X</td>
</tr>
<tr>
<td>7  Invalid or corrupted push buffer address</td>
<td>X</td>
</tr>
<tr>
<td>8  GPU stopped processing</td>
<td>X</td>
</tr>
<tr>
<td>9  Driver error programming GPU</td>
<td>X</td>
</tr>
<tr>
<td>10 Unused</td>
<td>X</td>
</tr>
<tr>
<td>11 Invalid or corrupted push buffer stream</td>
<td>X</td>
</tr>
<tr>
<td>12 Driver error handling GPU exception</td>
<td>X</td>
</tr>
<tr>
<td>13 Graphics Engine Exception</td>
<td>X</td>
</tr>
</tbody>
</table>

- Errors are documented
- [http://docs.nvidia.com/deploy/xid-errors/](http://docs.nvidia.com/deploy/xid-errors/)
Common Hardware Failures

- Easy to detect
  - IT DON’T WORK

If its broke, fix it…
GPU has Fallen Off the Bus

Yo JimBob, looks like another GPU fell off the bus!
GPU Has Fallen Off the Bus

- Very annoying
- Could be Hardware or Software
- Bugs Filed
  - 789858, 790527, 804390, 808113, 808114, 814107, 818374
- Corrective Action
  - Remove and clean GPU
GPU Killed by Application

**Symptom**
- User application hangs
- User application fails to make progress
- Compute node dies
- Did I mention Fallen off the Bus?

**Corrective Action**
- Run a GPU optimized HPCG
  - Nodes will crash and die
- If no hardware failure, clean and reseat GPU
Slow GPUs

- **Symptom**
  - Users report application performance issues
  - Reports of slow nodes

  “one of my jobs running on daint seems to be much slower than the other equivalent jobs. I suspect a broken node.”

- **Corrective Action**
  - Can reproduce with DGEMM – 40% reduction in performance

- **Root Cause 99%**
  - Bug 822829
  - Link width reduced from 16x to 8x
  - Detectable with nvidia-smi

```
<table>
<thead>
<tr>
<th>GPU Link Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link Width</td>
</tr>
<tr>
<td>Max : 16x</td>
</tr>
<tr>
<td>Current : 8x</td>
</tr>
</tbody>
</table>
```
System Regression Test Suite

- GPU Related System Checks
  - GOM setting
  - GPU exists
  - Link Width set to 16x
  - ~10 minutes

- GPU Killer
  - ~20 minutes

- Regression Test Suite
  - DGEMM
  - Representative Applications
  - Fully automated
  - ~20 minutes
Metric of Success

- What is the true measure of success?
- Quicker system validation
  - Previously: 3 hours post boot
  - Now: < 1 hour post boot
    - 10,544 node hours returned to service
- Suspect nodes removed from service
  - Detected before users
  - Improved user experience
  - Higher application success rate
What’s Next

- NVIDIA CUDA Toolkit 6.5
  - Improvements in error detection and recovery
  - Could eliminate certain failures
  - Could better identify failure conditions
  - Scheduled for May 11\textsuperscript{th}

- Better Error Tracking
  - Need to track all XID error codes
  - Better error detection, improves reliability
Conclusion

- User Experience Successfully Improved
- Continue to make improvements
Thank you for your attention.