Essential Feedback Loops
systems and apps, both man and machine

Blue Water / TeraGrid
Fault-Tolerance Workshop
Albuquerque, New Mexico
March 19, 2009

Jon Stearley <jrstear@sandia.gov>
(With input from Bob Balance and Sue Kelly)
Questions for speaker:

• What can you tell us about your fault/error situation currently?
  – What rate of errors?
  – What types of errors do you see?
  – How does one know if there is a fault?

• What are you worried about in the future?
  – What do you think is needed to help?
  – What do you think it is reasonable for apps people to do?
### Redstorm (both sides) causes of Unscheduled Downtime (lifetime)

<table>
<thead>
<tr>
<th>CAUSE</th>
<th>CATEGORY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HW</td>
</tr>
<tr>
<td>Cache Parity Error</td>
<td>4</td>
</tr>
<tr>
<td>Color Change</td>
<td>0</td>
</tr>
<tr>
<td>FC Connectivity</td>
<td>1</td>
</tr>
<tr>
<td>Hardware</td>
<td>49</td>
</tr>
<tr>
<td>I/O: Hang</td>
<td>0</td>
</tr>
<tr>
<td>I/O: Indeterminate Error</td>
<td>0</td>
</tr>
<tr>
<td>I/O:Channel error</td>
<td>0</td>
</tr>
<tr>
<td>I/O:DDN:</td>
<td>83</td>
</tr>
<tr>
<td>I/O:DDN:Offline</td>
<td>7</td>
</tr>
<tr>
<td>I/O:DDN:Reboot</td>
<td>14</td>
</tr>
<tr>
<td>I/O:Hang</td>
<td>0</td>
</tr>
<tr>
<td>I/O:LSI:</td>
<td>4</td>
</tr>
<tr>
<td>I/O:Lustre</td>
<td>0</td>
</tr>
<tr>
<td>I/O:Qlogic HBA</td>
<td>13</td>
</tr>
<tr>
<td>L0 Fault</td>
<td>0</td>
</tr>
<tr>
<td>Mesh</td>
<td>66</td>
</tr>
<tr>
<td>Mesh:Cable</td>
<td>5</td>
</tr>
<tr>
<td>Mesh:Deadlock</td>
<td>39</td>
</tr>
<tr>
<td>Mesh:FIFO Overrrun</td>
<td>0</td>
</tr>
<tr>
<td>Mesh:LUT error</td>
<td>0</td>
</tr>
<tr>
<td>Mesh:Link Failed</td>
<td>26</td>
</tr>
<tr>
<td>Mesh:Link Inactive</td>
<td>104</td>
</tr>
<tr>
<td>Module:Powerdown</td>
<td>13</td>
</tr>
<tr>
<td>Module:Powerdown:VRM</td>
<td>3</td>
</tr>
<tr>
<td>Module:Powerdown:Verty</td>
<td>33</td>
</tr>
<tr>
<td>Mtce</td>
<td>0</td>
</tr>
<tr>
<td>Portals</td>
<td>0</td>
</tr>
<tr>
<td>Power:Cabinet:EPO</td>
<td>11</td>
</tr>
<tr>
<td>Power:Cable</td>
<td>1</td>
</tr>
<tr>
<td>Power:Outage</td>
<td>0</td>
</tr>
<tr>
<td>Procedural</td>
<td>0</td>
</tr>
<tr>
<td>Procedural:Operator Error</td>
<td>0</td>
</tr>
<tr>
<td>SDB Crashed</td>
<td>0</td>
</tr>
<tr>
<td>Software</td>
<td>0</td>
</tr>
<tr>
<td>TCP/IP:CRC</td>
<td>0</td>
</tr>
<tr>
<td>Test:System</td>
<td>0</td>
</tr>
<tr>
<td>Unknown</td>
<td>0</td>
</tr>
<tr>
<td>Upgrade:HW</td>
<td>0</td>
</tr>
<tr>
<td>Upgrade:SW</td>
<td>0</td>
</tr>
<tr>
<td>Verty Failure</td>
<td>4</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>480</td>
</tr>
<tr>
<td><strong>%</strong></td>
<td>72%</td>
</tr>
</tbody>
</table>
Downtimes

Jobs Running/Day

12/11/07
12/18/07
12/25/07
1/1/08
1/8/08
1/15/08

Scheduled
Unscheduled

12/11/07
12/18/07
12/25/07
1/1/08
1/8/08
1/15/08

Launches

Large counts indicate user testing sessions
Moral of the story…

• Lots of faults and fault types.

• Yes it is a big problem now…
  Yes it will be a bigger problem later.

• But expert drivers get the job done!
Questions for speaker:

• What can you tell us about your fault/error situation currently?
  – What rate of errors?
  – What types of errors do you see?
  – How does one know if there is a fault?

• What are you worried about in the future?
  – What do you think is needed to help?
  – What do you think it is reasonable for apps people to do?
What are you worried about in the future?

• Silent corruption / soft errors (Michalak, Bronevetsky, …)

And the lack of…
What do you think is needed to help?

• Well-defined, Standardized Metrics! (Stearley, Daly)
A Slide From John Daly:

Defining a Productive Work Rate in Terms of How the System is Spending its Time

Operational Utilization = \frac{Run Time}{Operations Time} = \frac{t_r}{t_{op}}

Runtime Efficiency = \frac{Solve Time}{Run Time} = \frac{t_s}{t_r}

Adapted with permission from Jon Stearley (based on SEMI-E10)

Productive Work Rate = Efficiency \cdot Utilization
A Slide From John Daly:
Operations Rate Only Tells Part of the Story:
Red Storm From The Application’s Perspective

Daily Availability for the 5000 Node Job with 7-Day Average MTBI and Efficiency
(Cumulative Availability = 60% and Cumulative Efficiency = 63%)

Available Time
System interrupts
Application Interrupts
Efficiency
MTBI (System Only)
MTBI (Sys + App)
What do you think is needed to help?

- Well-defined, Standardized Metrics!
- Architecture-independent integrated subsystems with useable licenses
  - Don’t try to solve everything with your monolith!
  - Eg CIFTS, OVIS, Sisyphus, TriPOD Monitoring suite, ...
- Good fault reporting.
  - Don’t just exit(), log what you were trying to do!
  - Need a better reporting mechanism.
    - Systems people have systems data, apps people have apps data - we need more common ground.
What do you think is reasonable for apps people to do?

- Practice good fault reporting.
- Learn best-practices for app fault-tolerance
  - Metrics and checkpoint optimization (Daly)
  - Checkpoint compression (Gibson, Schroeder)
- Be aware of fault-tolerant algorithm research (eg Chen et al)
- Collaborate with systems people towards fault tolerance (eg Glosli et al)
  - You can’t solve the problem alone, and neither can we.
Essential Feedback Loops
systems and apps, both man and machine

The End

Jon Stearley <jrstear@sandia.gov>
• What do all these exit codes mean?

• Which ones indicate system faults?

• What are correlated factors?
  – Username?
  – Application name?
  – Node?
  – Network switch? Cable?
  – Syslogs? App logs? …

(we need better ways to investigate such questions…)
Red Storm Jumbo Mode, 12/8/07 – 3/11/08
System CPU Hours by Job Size

Classified Opus Computation

Unclassified All Computations

Classified All Computations

All Computations

16384-25920
8192-16383
4096-8191
2048-4095
1024-2047
<1024
Red Storm