Analysis of Gemini Interconnect Recovery Mechanisms: Methods and Observations


*University of Illinois at Urbana Champaign/NCSA
† Bell Labs

Saurabh Jha
sjha8@Illinois.edu
Department of Computer Science
University of Illinois, Urbana – Champaign
CUG 2016
CSL.ILLINOIS.EDU
2016
Petascale

2020-2025
Exascale?

400K – 3M
Cores

Mean Time
Between System-
wide Failure ~1-2
Weeks

10 -100M cores
billions of threads

Resiliency going to be a major issue!
Path to Understanding Interconnect Resiliency Challenges

- **Measure** failure rates and mean time between failures
- **Model** interconnect failures and *interconnect recovery operations*
  - Extended LogDiver\(^2\) with interconnect analysis tool to re-create recovery scenarios by generating recovery-sequence clusters \(^1\)
- **Build** failure propagation paths and dissect root causes for failures
  - Analysis of recovery-sequence clusters helps to build failure propagation paths and dissects root causes
- **Quantify** impact
  - System-wide outages
    - 27.7% of system-wide outages caused by network-related recovery operations
  - Application failures
    - 20% of applications running during the unsuccessful failover procedure failed
    - 0.2% of applications running during the successful recovery procedures failed

*For detailed results and other interesting insight please refer to:
2. LogDiver: A Tool for Measuring Resilience of Extreme-Scale Systems and Applications*
Outline

• Gemini Overview
• Dataset
• Network Analysis Methodology & Tool
• Example Use Case & Observations
• Conclusions
Gemini Overview in Blue Waters

- **Size**
  - XE: 22,640 CPU Only nodes
  - XK: 4,224 GPU+CPU nodes

- **Gemini**
  - 3D Torus
  - Topology: 24x24x24
  - 48 Port Router
  - 6 links: X+, X-, Y+, Y-, Z+, Z-
  - 9.6 GB/sec
  - 10 Torus Connection per router
Gemini Resiliency Features

• Hardware:
  – Multiple Torus connections in X/Z direction
  – 2 redundant links and 3 redundant lanes per link
  – Packets protected by 16-bit CRC
  – Memory regions protected by SEC-DED (except router table buffers)

• Recovery Procedures
  – Lane Recovery
  – Link Recovery
  – Manual Recovery (Warm Swaps)
Blue Waters (studied) Logs

• Time: [819 days] - January 1, 2013 to March 31, 2015

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Events Registered</th>
<th>Dataset Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Syslogs</td>
<td>75,760,682,632</td>
<td>13 TB</td>
</tr>
<tr>
<td>Manual Failure Reports</td>
<td>4,184</td>
<td>1.4 MB</td>
</tr>
<tr>
<td>Coalesced Workload from LogDiver[2]</td>
<td>20,600,030</td>
<td>8 GB</td>
</tr>
</tbody>
</table>
Recovery Operations Described As State Transition Diagrams
Analysis Workflow Steps

1. Filtering and tagging
2. Clustering
3. Correlating with system-wide outages
4. Correlating with application failures
Data Analysis Pipeline (LogDiver+)
Topologically-aware State Transition Based Clustering Algorithm

Temporal (State-Machine) Status

C1
- Faults
- Failures

C2
- Faults
- Failures

Failover Start

Failover Transition

Failover Finish

Success

Spatial View

System /SMW

Cabinet /1

Blade/1

ASIC/Node/1

Link/1

Log Stream

1. Packet Misrouted (Node/1)
2. Corrupt Routing Tables (ASIC/3)
3. Link Failed (Link/1)
4. ASIC Failed (ASIC/3)
5. Link Failover Begin (SMW)
6. Aggregate Failures (SMW)
7. Network Quiesced (SMW)
8. Network Unquiesced (SMW)
9. Failover Finished (SMW)
10. Failover Success (SMW)

Cabinet /2

Blade/2

ASIC/Node/2

Link/2

Cabinet /3

Blade/3

ASIC/Node/3

Link/3

Link/4
Completion Status of Recovery Procedures

Event Counts
- Lane Recoveries 253,000
- Link Recoveries 318
- Warm Swaps - 559

Overall Success Percentage of Interconnect Recovery Procedures
Impact of Software Upgrades on Recovery Completion Status

Indicates Major Software Upgrade in the Gemini Recovery Code
System Impact

• ~27% of the system-wide outages (28/101) were related to network recovery operations

- Overlapping Lustre Recovery: 17.9%
- Failure During Recovery, 3: 10.7%
- HSN Deadlock Due To Corrupt Routing Tables, 4: 14.3%
- Handshake/Timeout Issue, 16: 57.1%
Application Impact

- Application impact was analyzed by disambiguating exit status of applications using ALPS logs and syslogs via LogDiver.
  - User related exit reasons were ignored, e.g. Segmentation fault

- Irrespective of completion status (success/failed) of Gemini recovery operations, applications may fail
  - 20% of applications running during the unsuccessful failover procedure failed
  - 0.2% of applications running during the successful recovery procedures failed
Successful Link Failover Operation

Failure Propagating and Impacting Application state

ASIC failed
Routing table corruption

Can Lead to loss of control packets Causing applications to fail

Begin link failover
Rerouting required
Quiesce network

Node unavailability can cause application failure, need for feedback to application

Unquiesce network
Rerouting success
Link failover success

Pre-quiescence phase
Quiescence phase
Post-quiescence phase

T + 82 s
T + 127 s
T + 230 s

Can Lead to loss of control packets Causing applications to fail

Node unavailability can cause application failure, need for feedback to application
Conclusions

• Built LogDiver+ and demonstrated its capabilities for understanding and measuring the impact of network-related failures.
• Mined and analyzed failure propagation paths and reasons for the failure of the recovery and what-if analysis.
• Measured the impact of network failures on system and applications.
Future Roadmap

• Real-time resiliency measurements
  – Deployment of LogDiver+ at NERSC, LANL, SNL

• In-depth analysis of Aries networks on Mutrino (SNL) and Trinity (LANL)

• Use statistical learning to extract actionable intelligence
Any Questions...  
Just Ask!