CSL: COORDINATED SCIENCE LAB

NCSA

Analysis of Gemini Interconnect Recovery Mechanisms: Methods and Observations

Saurabh Jha*, Valerio Formicola*, Catello Di Martino†, William Kramer*, Zbigniew Kalbarczyk*, Ravishankar K. Iyer*

***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

2016
Petascale

400K – 3M
Cores

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

2020-2025
Exascale?

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

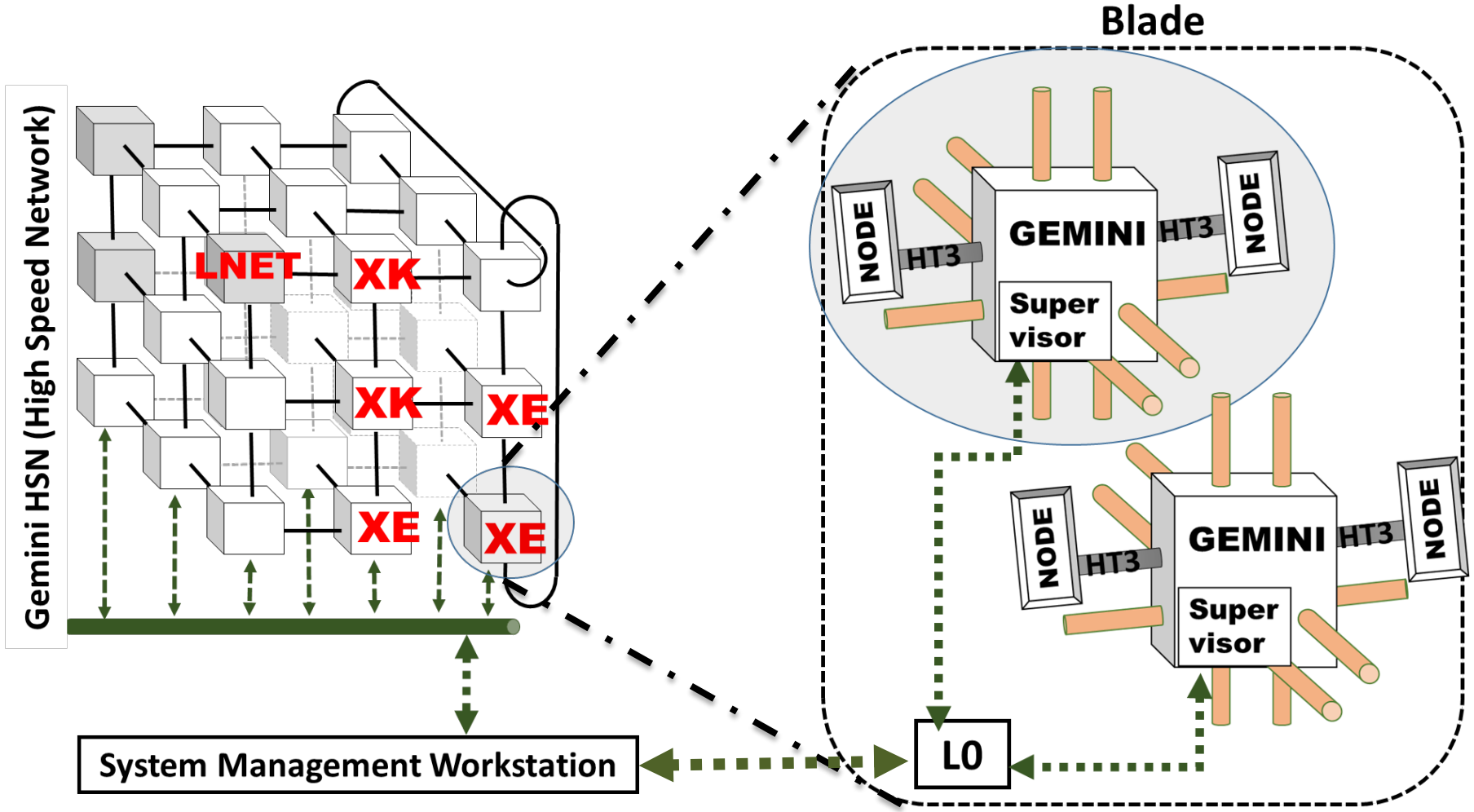
** For detailed results and other interesting insight please refer to:*

1. *Analysis of Gemini Interconnect Recovery Mechanisms: Methods and Observations , Cray User Group 2016*
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

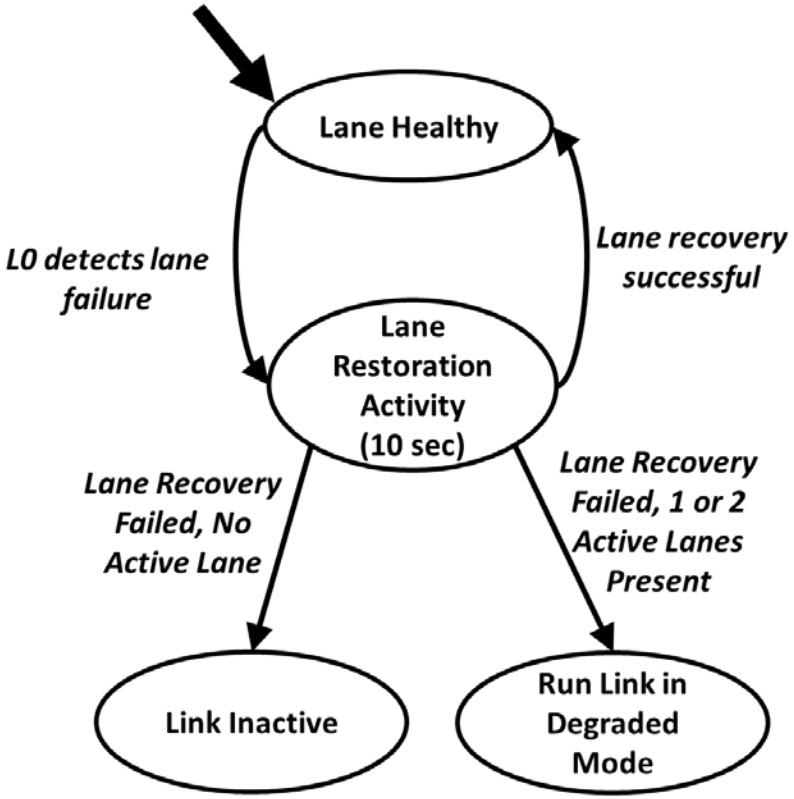
- Hardware:
  - Multiple Torus connections in X/Z direction
  - 2 redundant links and 3 redundant lanes per link
  - Packets protected by 16-bit CRC
  - Memory regions protected by SEC-DED (except router table buffers)
- Recovery Procedures
  - Lane Recovery
  - Link Recovery
  - Manual Recovery (Warm Swaps)

Blue Waters (studied) Logs

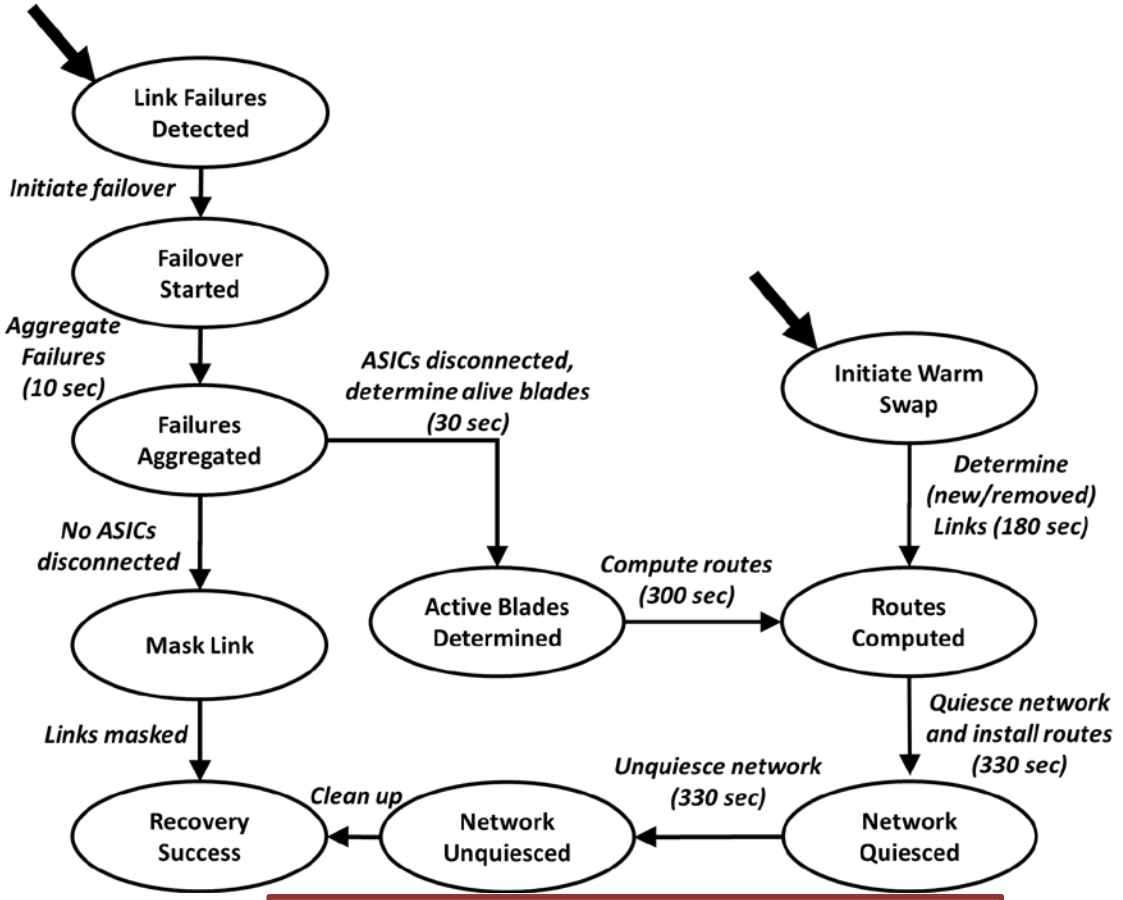
- Time : [819 days] - January 1, 2013 to March 31, 2015

Data Source	Events Registered	Dataset Size
Raw Syslogs	75,760,682,632	13 TB
Manual Failure Reports	4,184	1.4 MB
Coalesced Workload from LogDiver[2]	20,600,030	8 GB

Recovery Operations Described As State Transition Diagrams



Lane Recovery

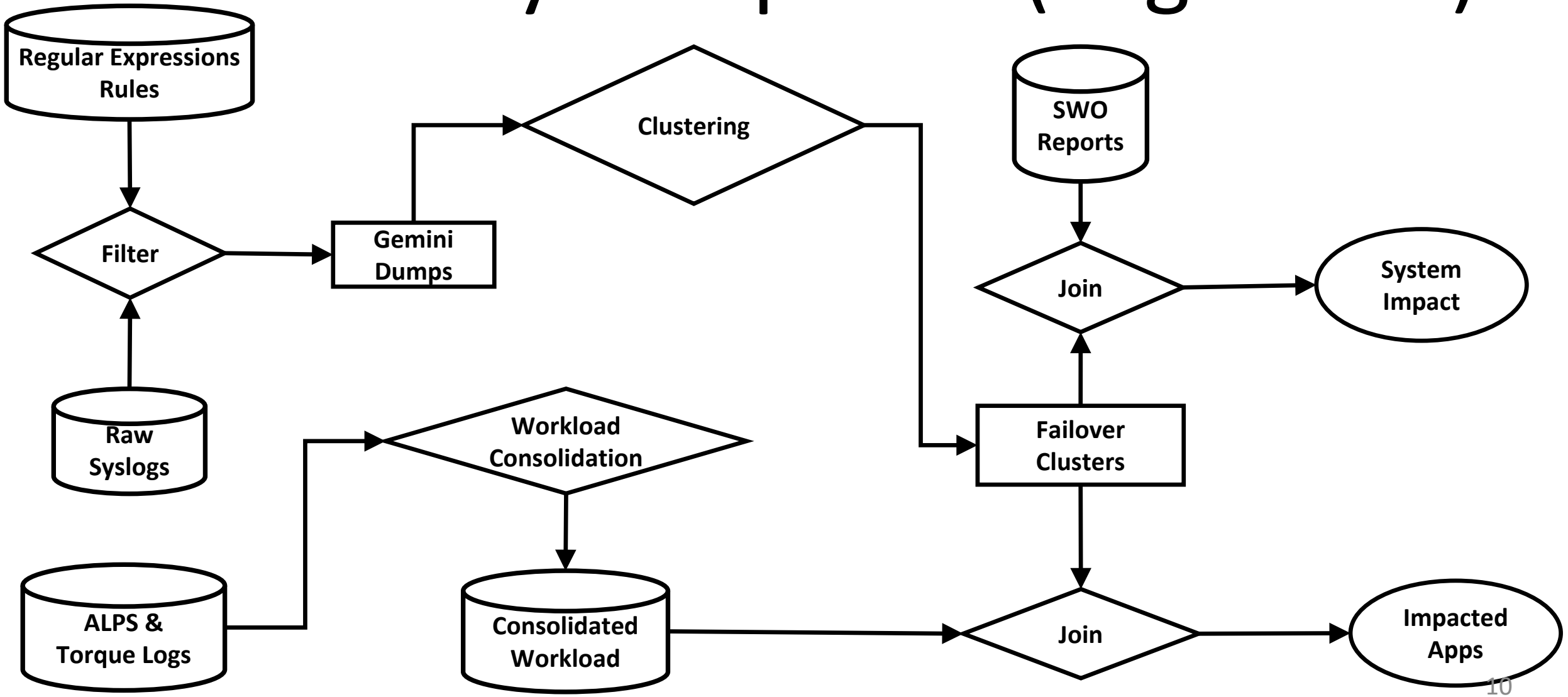


Link Failover / Warm Swap

Analysis Workflow Steps

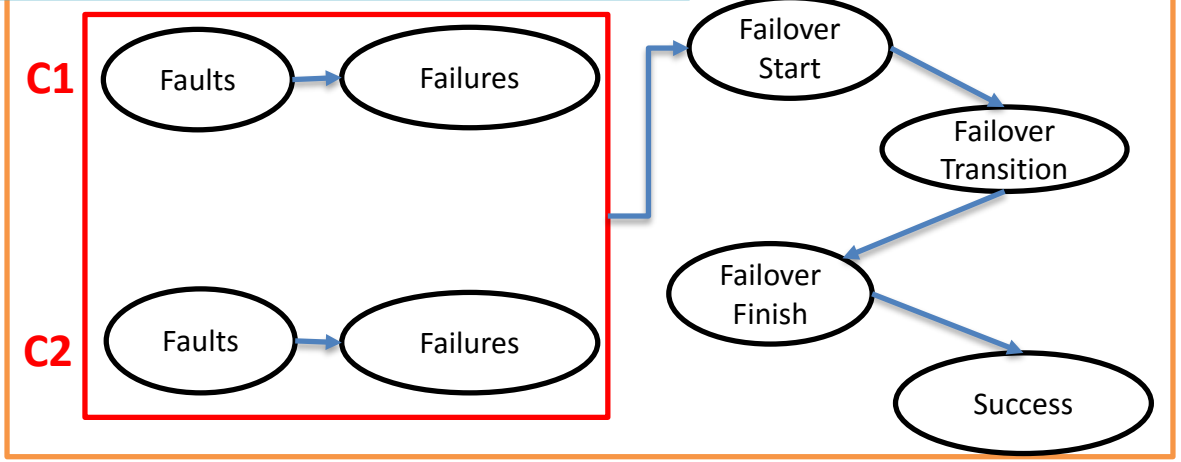
1. Filtering and tagging
2. Clustering
3. Correlating with system-wide outages
4. Correlating with application failures

Data Analysis Pipeline (LogDiver+)



Topologically-aware State Transition Based Clustering Algorithm

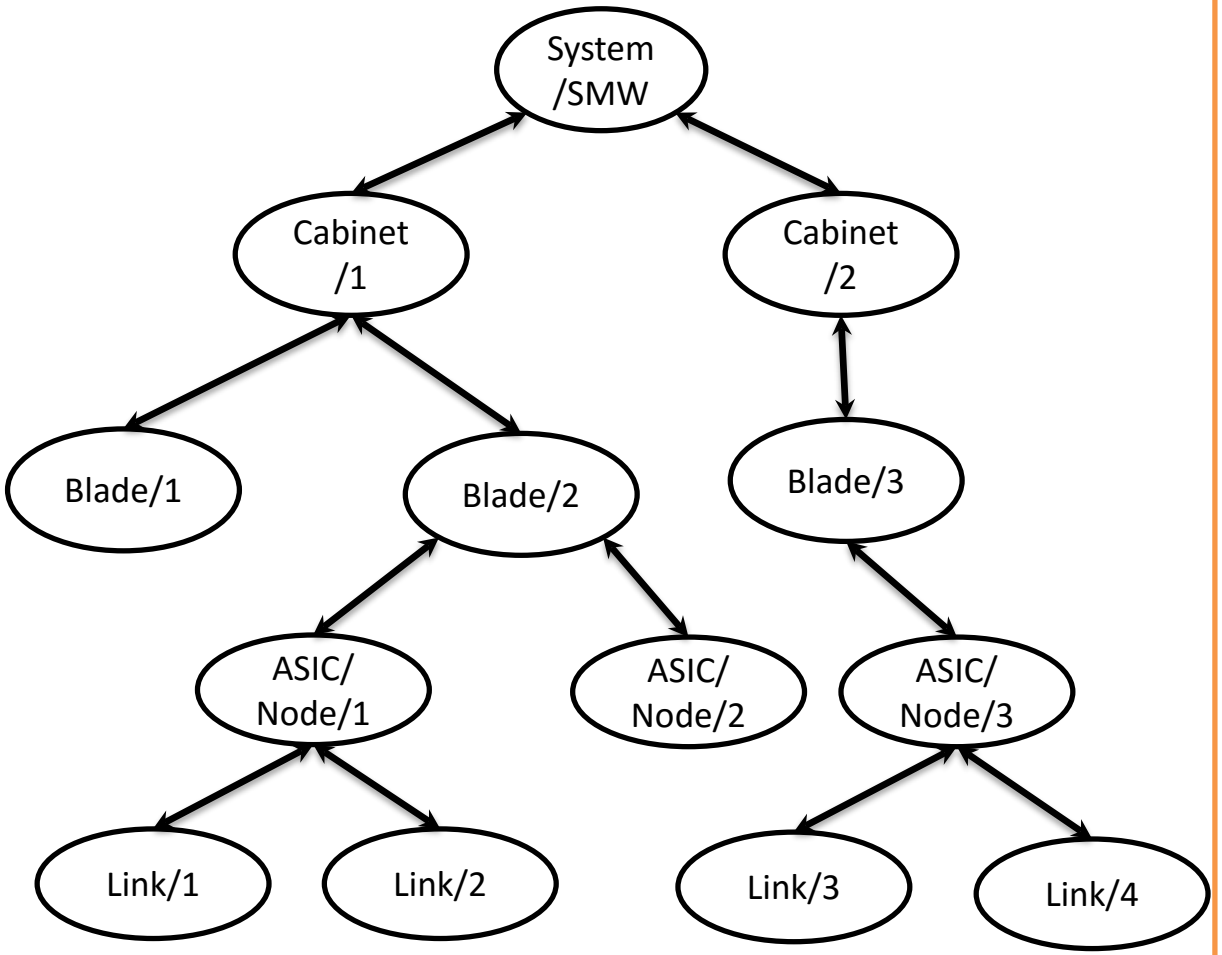
Temporal (State-Machine) Status



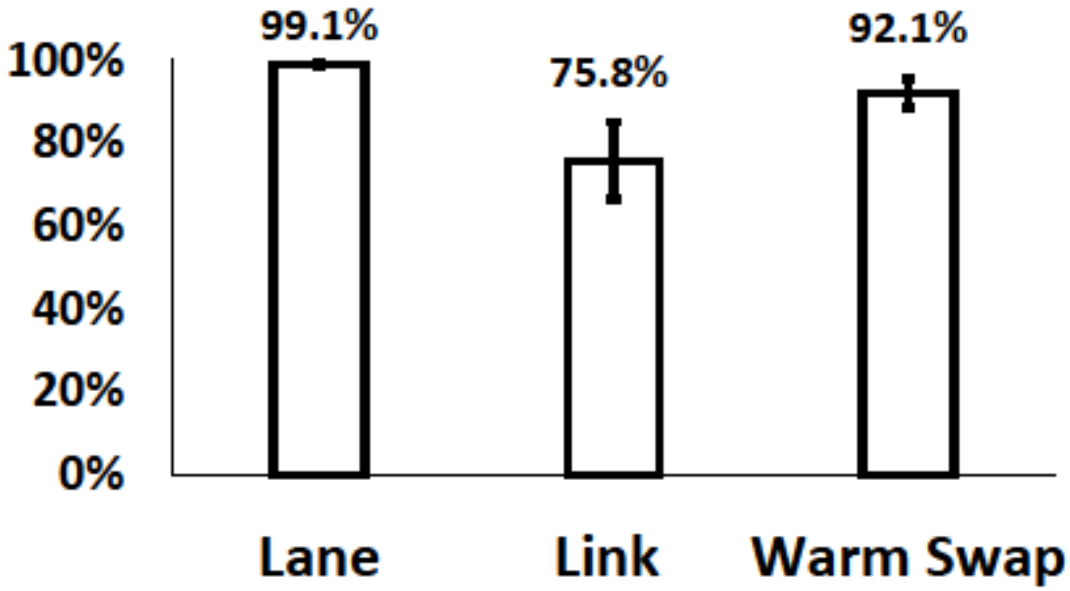
Log Stream

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)

Spatial View



Completion Status of Recovery Procedures



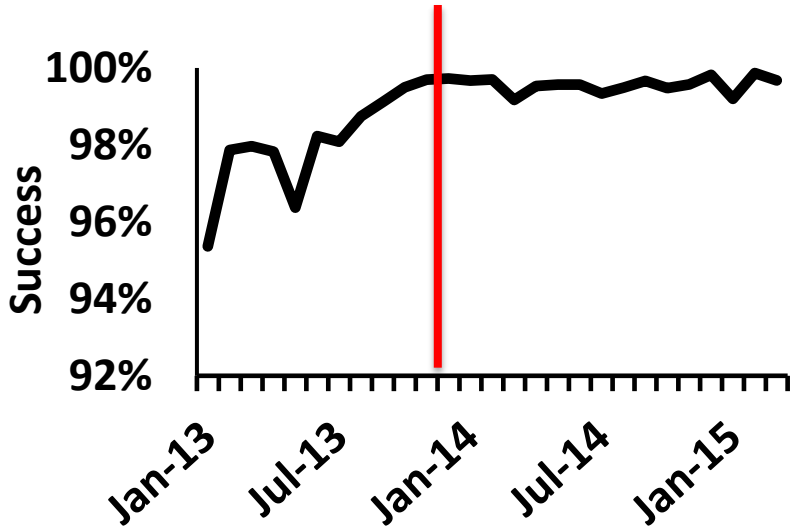
Overall Success Percentage of Interconnect Recovery Procedures

Event Counts

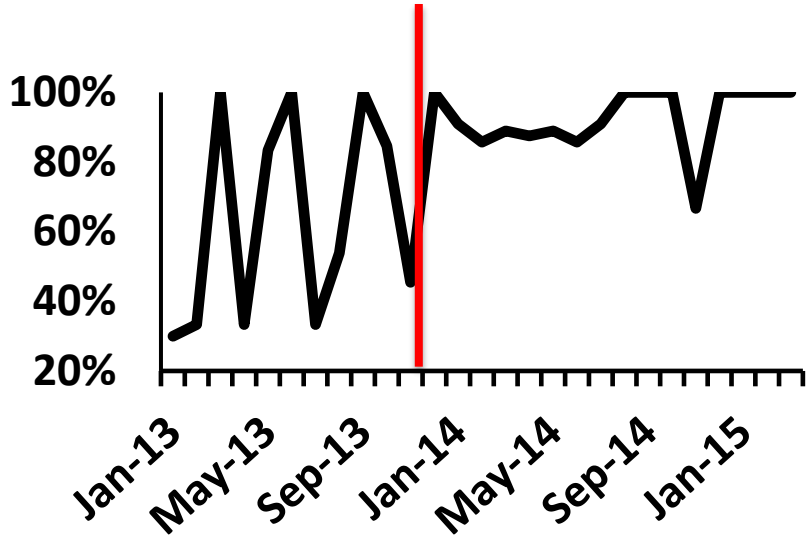
- Lane Recoveries 253,000
- Link Recoveries 318
- Warm Swaps - 559

Impact of Software Upgrades on Recovery Completion Status

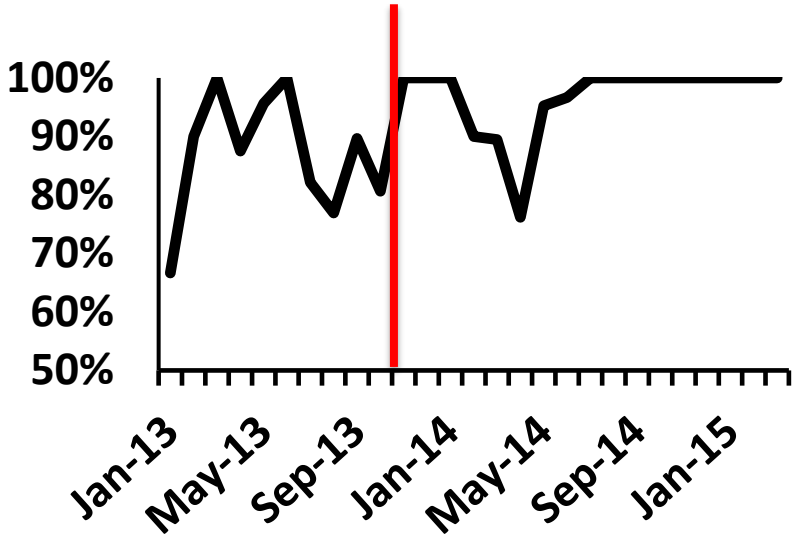
Indicates Major Software Upgrade in the Gemini Recovery Code



Lane



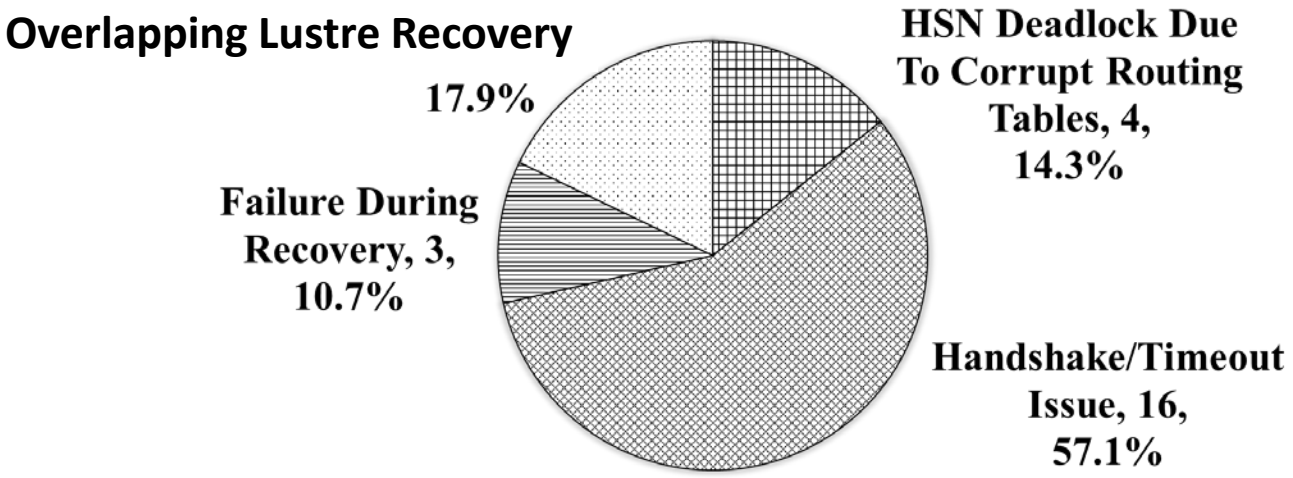
Link



Warm Swap

System Impact

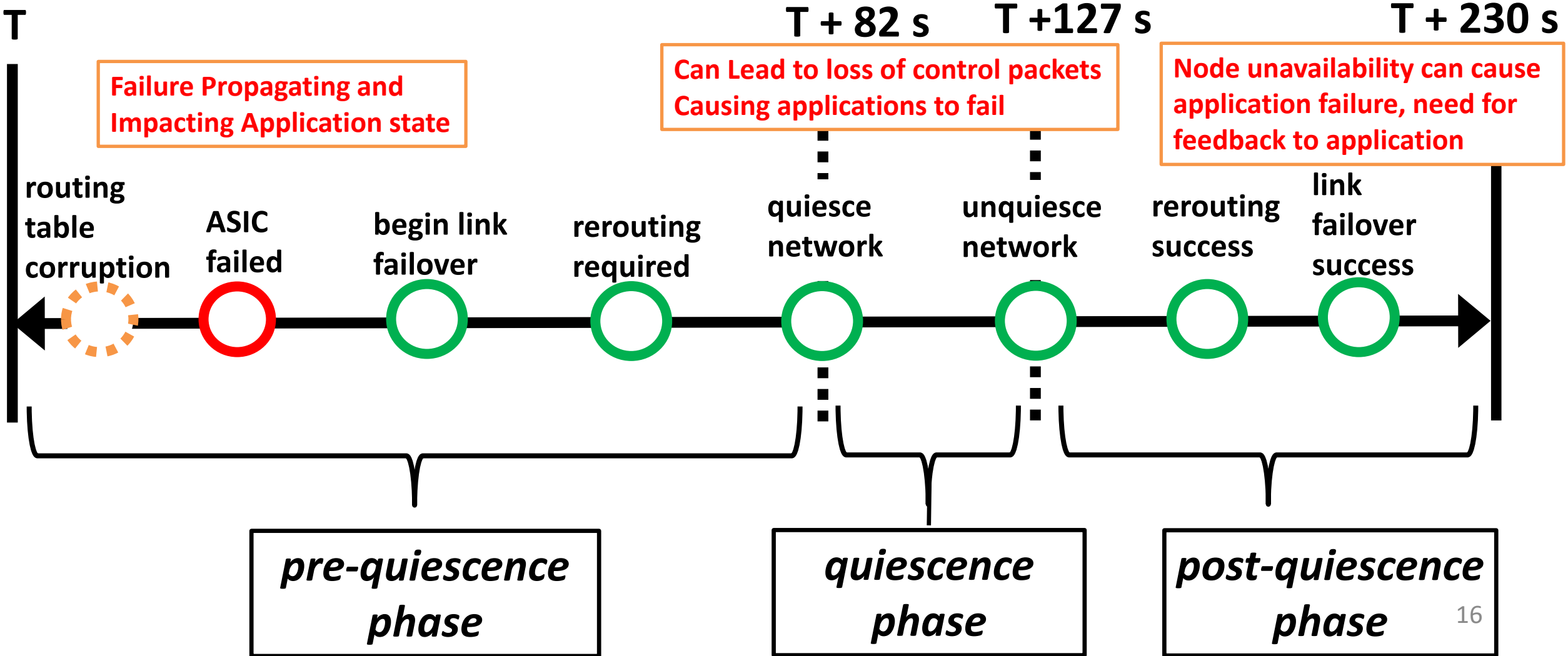
- ~27% of the system-wide outages (28/101) were related to network recovery operations



Application Impact

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



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

NERSC
CRAY THE SUPERCOMPUTER COMPANY
Los Alamos NATIONAL LABORATORY EST. 1943
Sandia National Laboratories
NCSA

Any Questions...
Just Ask!