Towards Understanding Interconnect Failures in HPC Systems

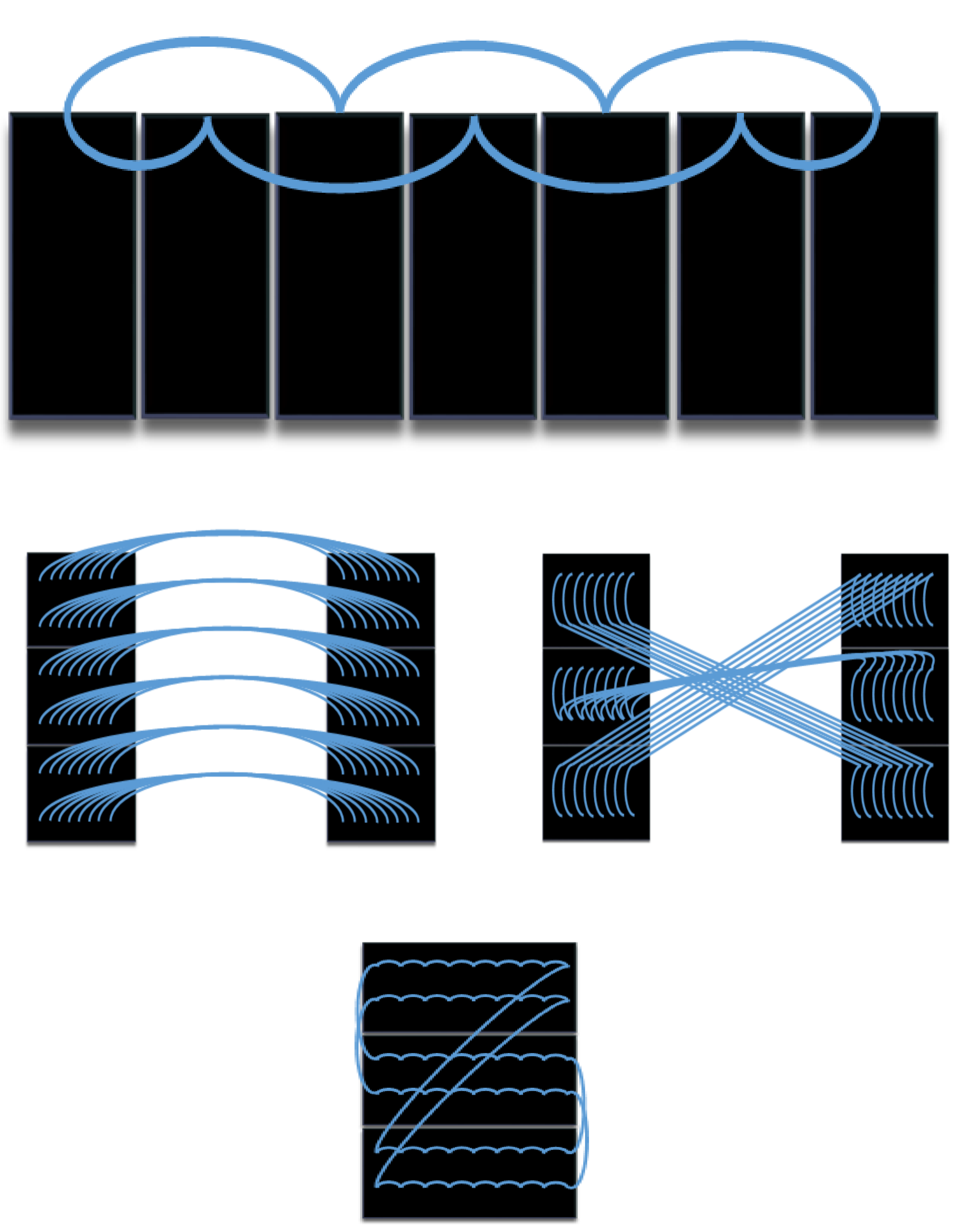
Mohit Kumar[1], Devesh Tiwari[2], Saurabh Gupta[3], Weisong Shi[1], Song Fu[4]
Wayne State University[1], Northeastern University[2], Oak Ridge National Laboratory[3], University of North Texas[4]

INTRODUCTION

- HPC deliver performance in Petaflops:
  - Fast computing devices
  - Interconnect
  - Back-end storage system
- Interconnect have a major impact :
  - Resilience mechanism
  - Congestion resolution mechanism
- Problem
  - No state-of-practice experience report on interconnect errors and congestion events
- Interconnect resilience and network congestion events on the Titan supercomputer

BACKGROUND

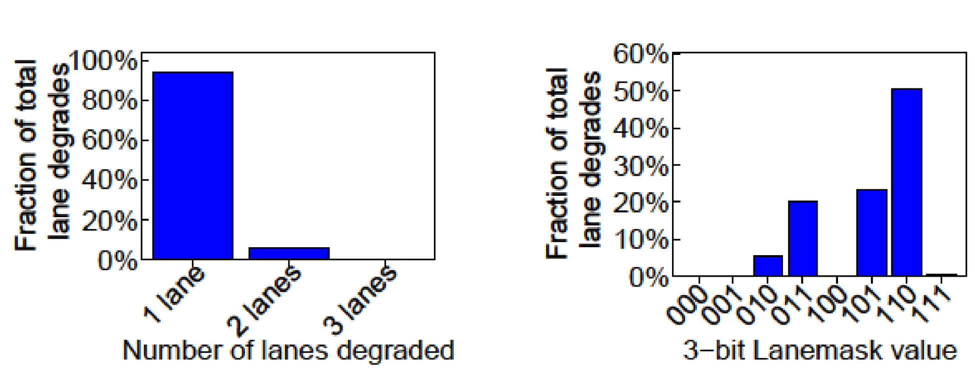
- Titan supercomputer
  - 200 cabinet- 25 row and 8 column
  - 18,688 nodes
  - 27.1 Petaflops peak performance
- Each cabinet consists of three cages
- Each cage has eight blades.
- Each blade consists of two application specific integrated circuits(ASIC)
- Network architecture
  - 3D torus topology using Cray Gemini
  - Each ASIC has 10 torus connection
  - Each torus connection has four links
  - Each link is composed of 3 single-bidirectional lanes.



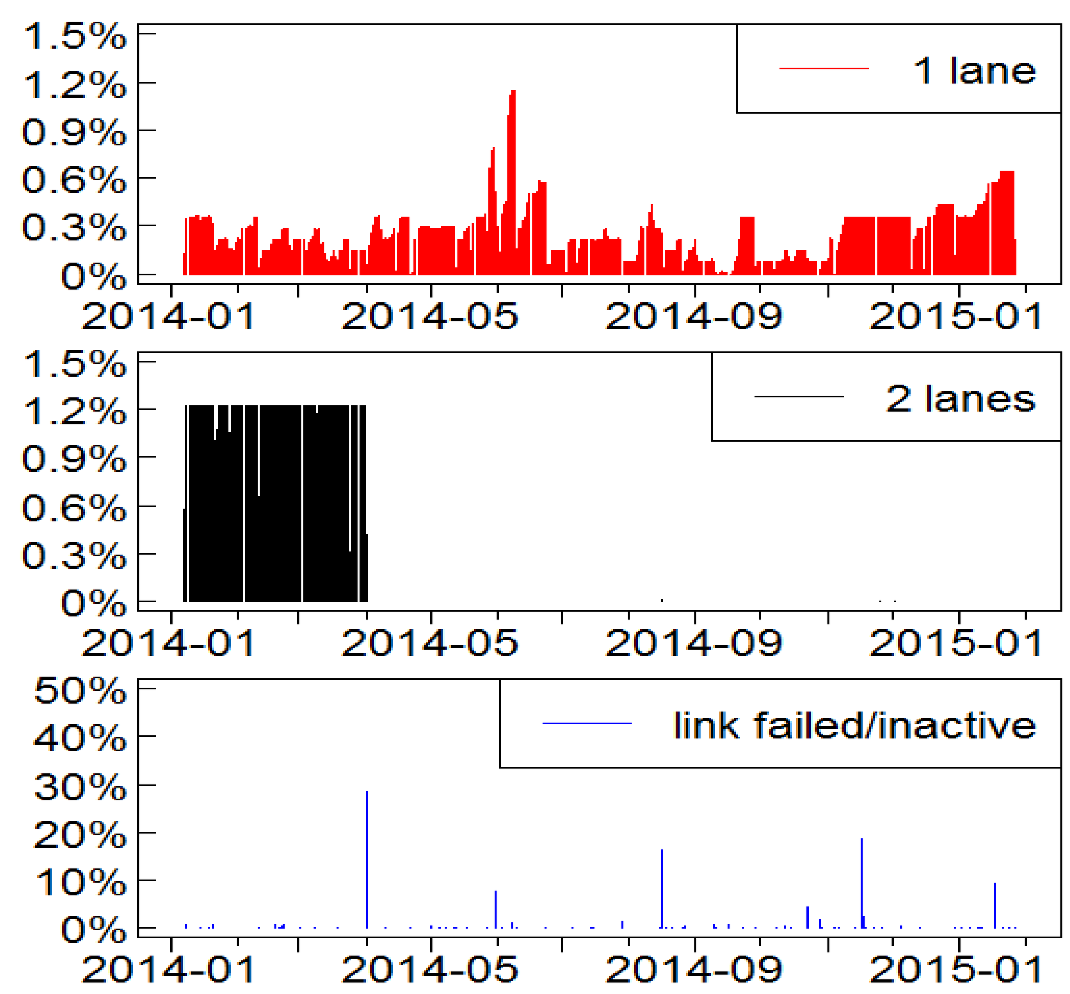
- Link can withstand failure of up to two lanes
- Dataset collected from Jan 2014 to Jan 2015
  - xtnetwatch
  - xtnlrd

LANE DEGRADE ANALYSIS

- Lane degrade events overall frequency
  - 90% cases only one lane in a link is degraded
  - Single lane failure varies significantly

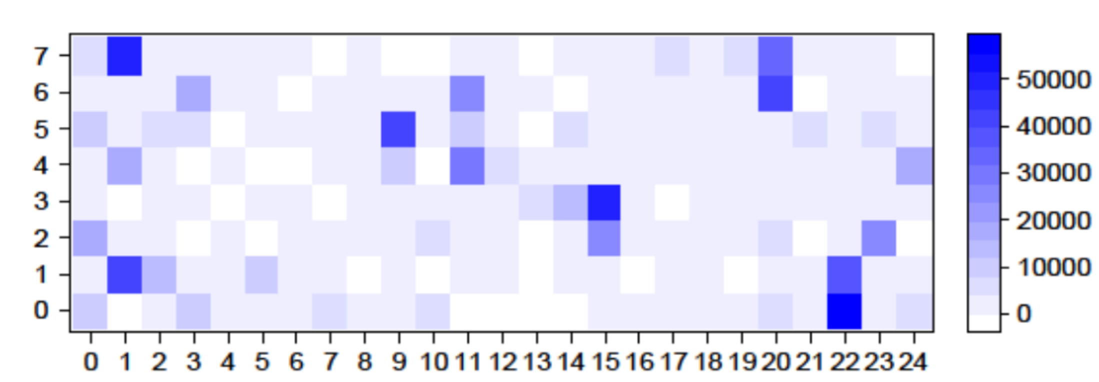


- Lane degrade events frequency over time
  - Lane degrades are not limited to a specific time period
  - 1-lane degrade events doesn't lead to 2-lane degrade events

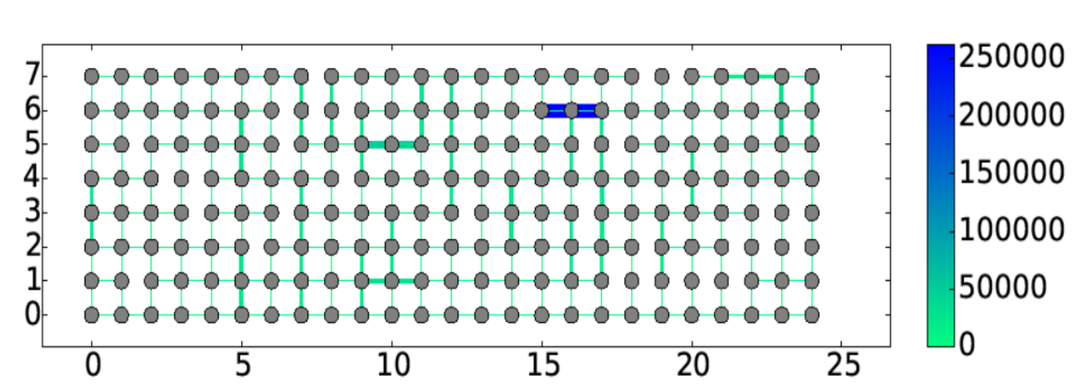


- Lane degrade events spatial distribution
  - Hot spots for links within the cabinet are not the same as the hot spots for links crossing cabinet boundaries

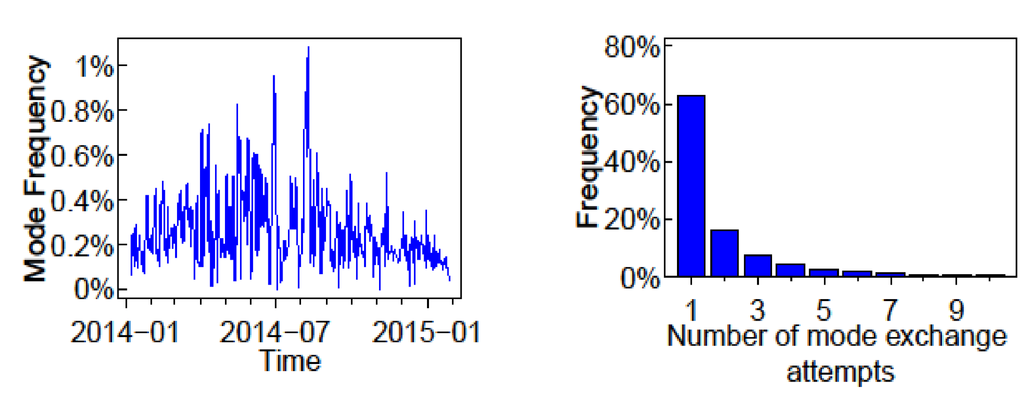
Per-cabinet distribution of lane degrades



Across-cabinet distribution of lane degrades

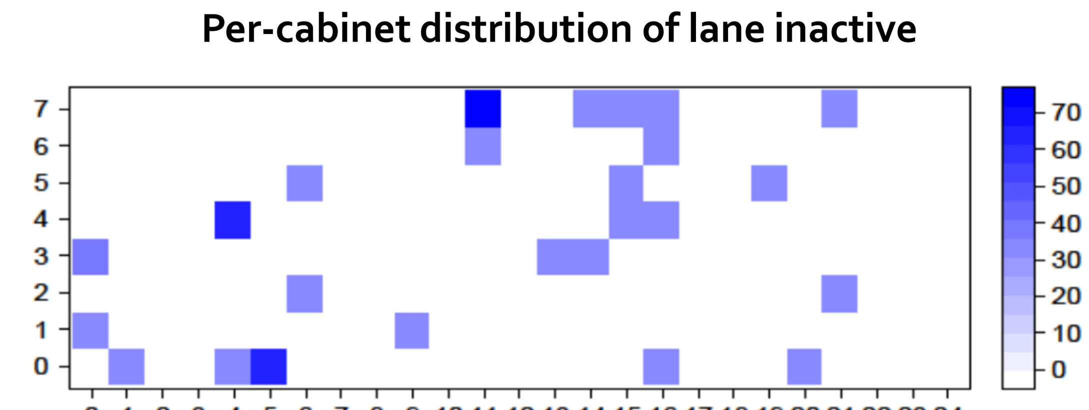


- Mode exchanges attempts – Max. 256
  - 85% of the lanes restored in three or less attempts
  - Mode exchange attempts events frequency is similar to lane degrade events

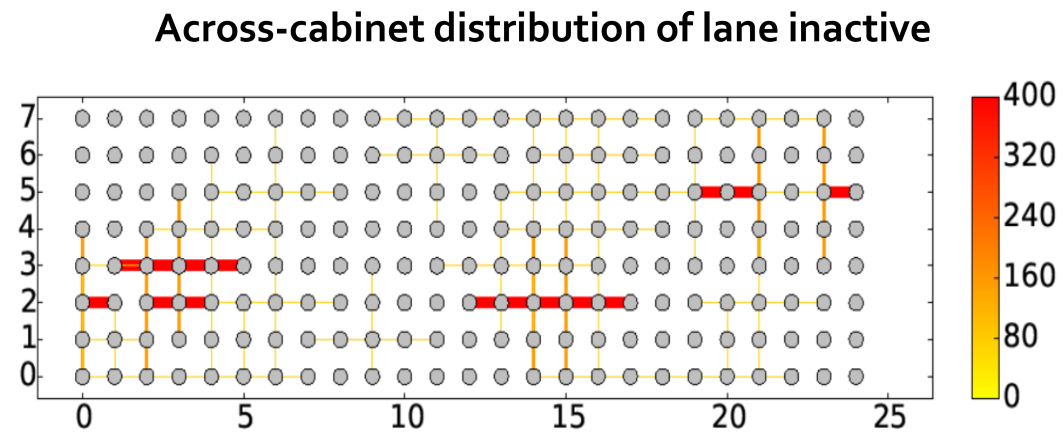


- Link Inactive
  - Hot spots for links inactive can not be predicted by observing time and location of lane degrade events

Per-cabinet distribution of lane inactive



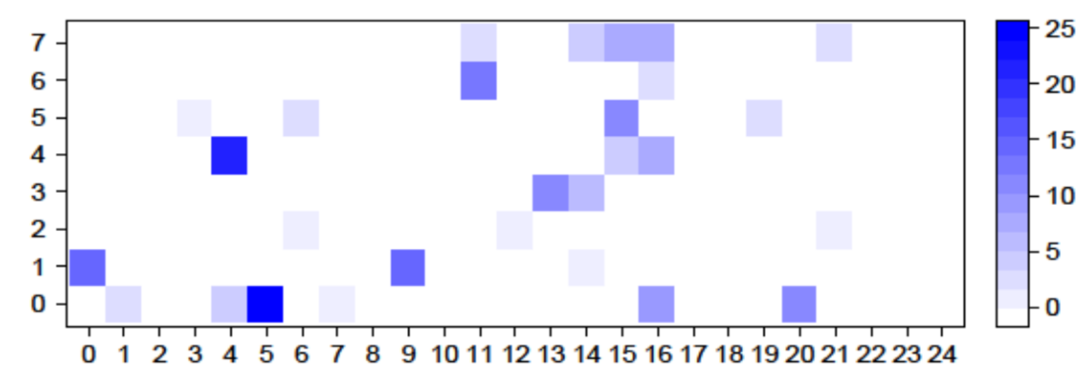
Across-cabinet distribution of lane inactive



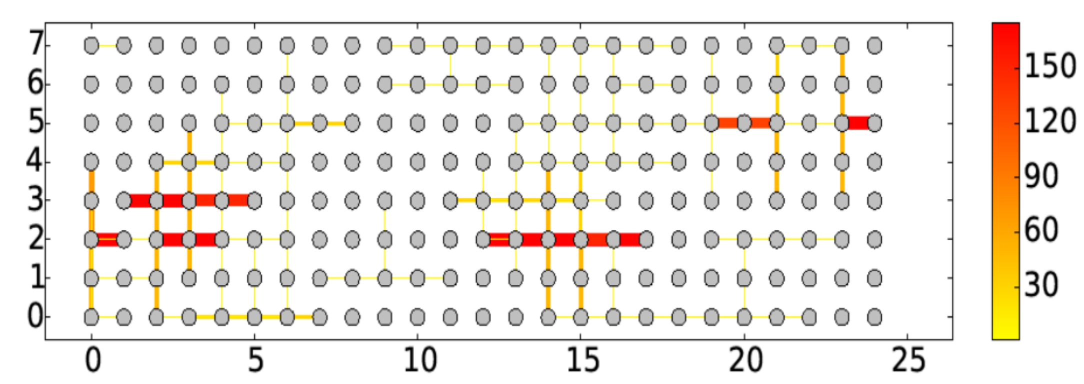
ERROR ANALYSIS

- Magnitude of errors is small
- High correlation with lane inactive
- Bad Send EOP errors

Per-cabinet distribution of Bad Send EOP errors

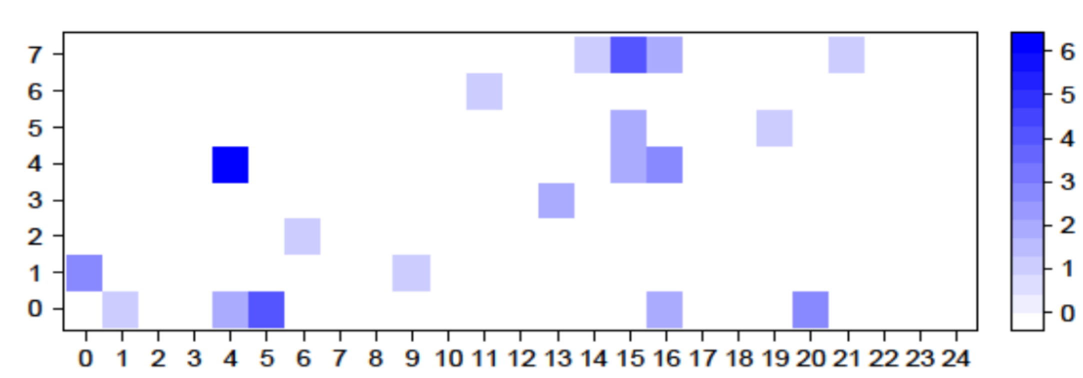


Across-cabinet distribution of Bad Send EOP errors

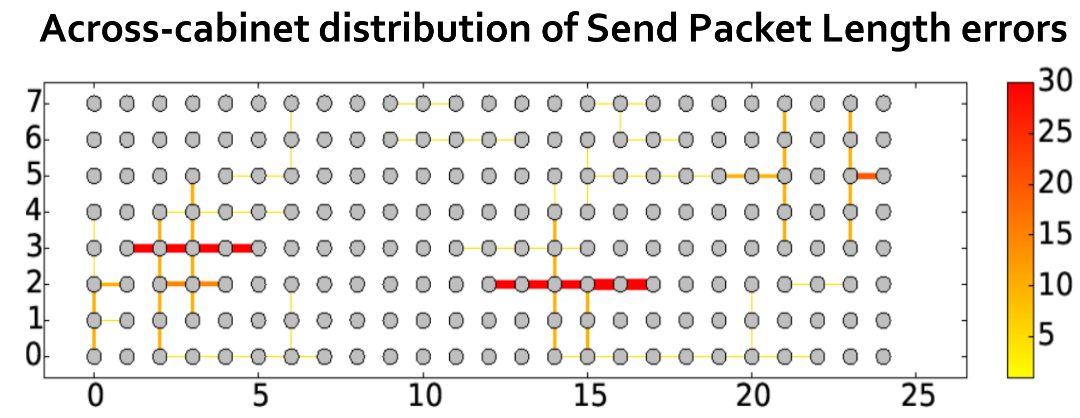


- Send Packet Length errors

Per-cabinet distribution of Send Packet Length errors

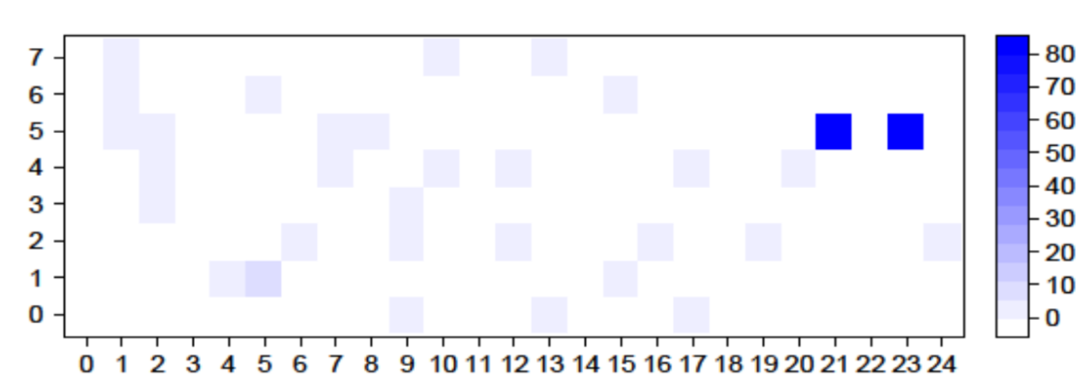


Across-cabinet distribution of Send Packet Length errors



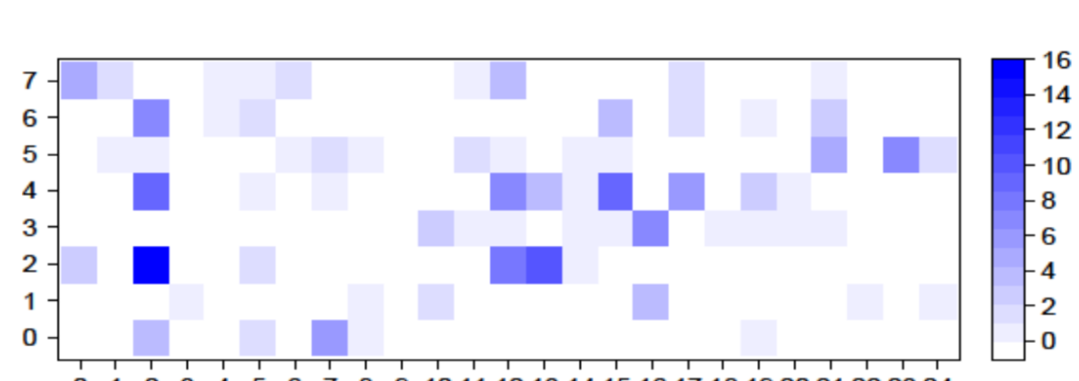
- Routing Table Corruption errors

Per-cabinet distribution of Routing Table Corruption errors



- HSN ASIC LCB lanes reinit failed errors

Per-cabinet distribution of ASIC lanes reinit failed errors



CONCLUSIONS

- Explored characteristics of lane degrade and major interconnect errors
- 80% cases lane inactive lead to interconnect errors
- Lane inactive events can't predict lane degrades events or mode exchanges attempts

CONTACT

Mohit Kumar, Ph.D. Candidate.
Department of Computer Science
Wayne State University
mohitkumar@wayne.edu