

Understanding the Spatial Characteristics of DRAM Errors in HPC Clusters

The 7th Workshop on Fault Tolerance for HPC at eXtreme Scale (FTXS) 2017

Ayush Patwari (Purdue University), Ignacio Laguna (LLNL)
Martin Schulz (LLNL), Saurabh Bagchi (Purdue University)

June 26, 2017

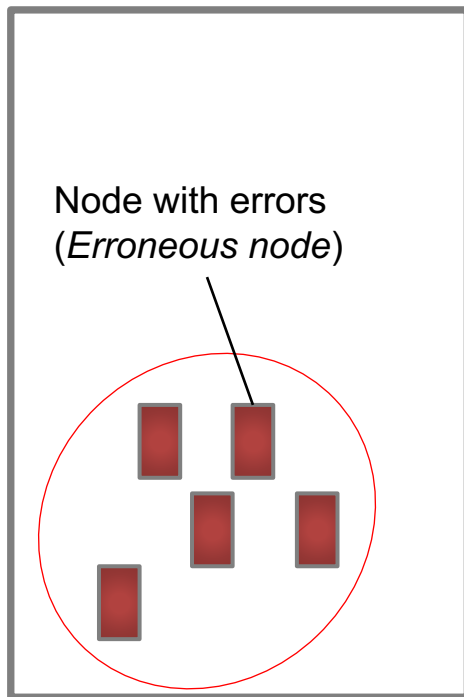


Understanding the Characteristics of DRAM Errors in HPC Clusters is Important to Address Resilience Challenges

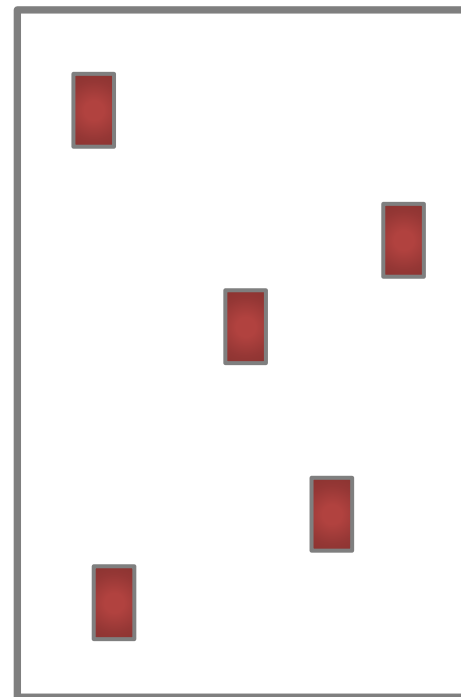
- DRAM errors are a common source of failures
- Exascale machines estimated to have hundreds of petabytes of memory
- Previous studies focus on within node/rack analysis of DRAM errors
- We need more studies on spatial characteristics
 - Physical layout of the clusters
- Useful for temperature management, job scheduling, failure prediction

Is there Any Spatial Correlation Between Erroneous Nodes in a Cluster?

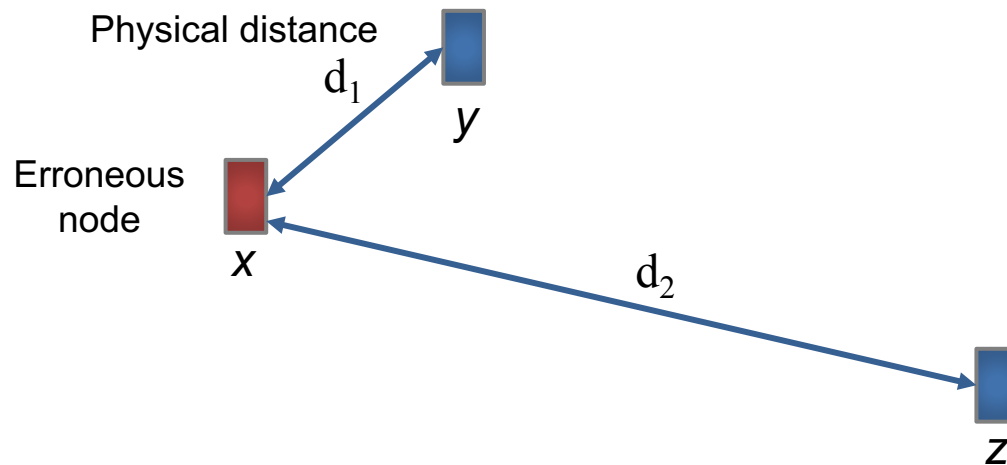
Nodes Rack (#1)



Nodes Rack (#20)



A Mathematical Foundation to Model Spatial Correlations Would be Useful



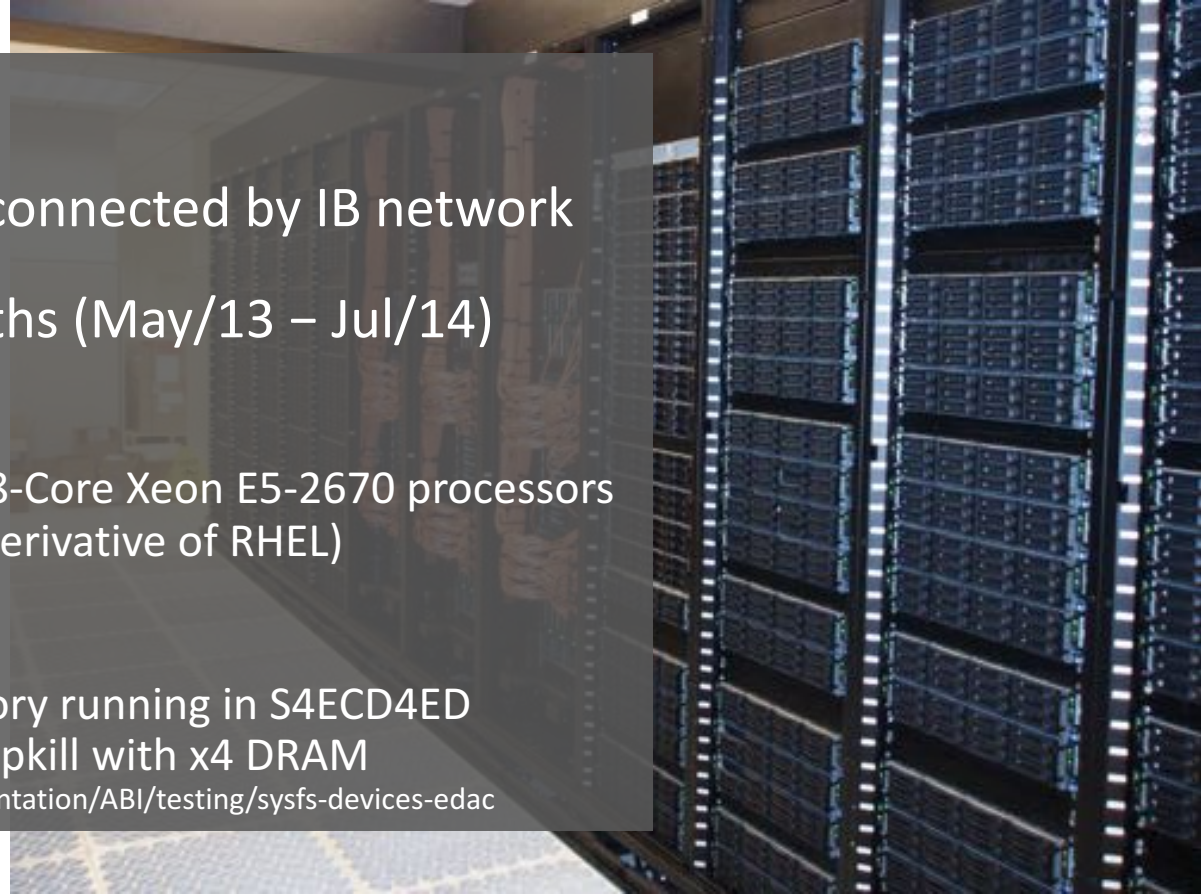
What is the **probability** that, given that a node x had errors, a node y at a distance d_1 , will have errors in the future?

$$P(\text{node } y \text{ will have errors} \mid \text{node } x \text{ had errors}) = ?$$

DRAM Error Data Gathering in the LLNL Cab System

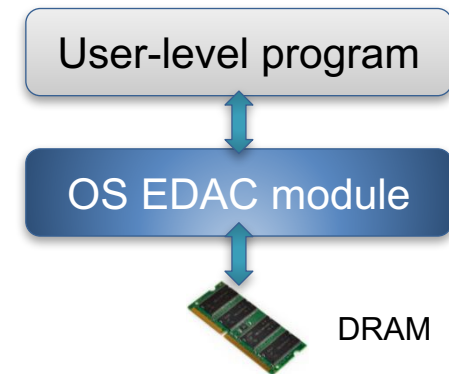
- Cab Cluster at LLNL
- 1,296 compute nodes, connected by IB network
- Period of time: 14 months (May/13 – Jul/14)
- Node configurations:
 - Each node has two Intel 8-Core Xeon E5-2670 processors
 - TLCC operating system (derivative of RHEL)
- Memory:
 - x4 DDR3 1600MHz memory running in S4ECD4ED
 - S4ECD4ED: similar to Chipkill with x4 DRAM

Ref: <https://www.kernel.org/doc/Documentation/ABI/testing/sysfs-devices-edac>



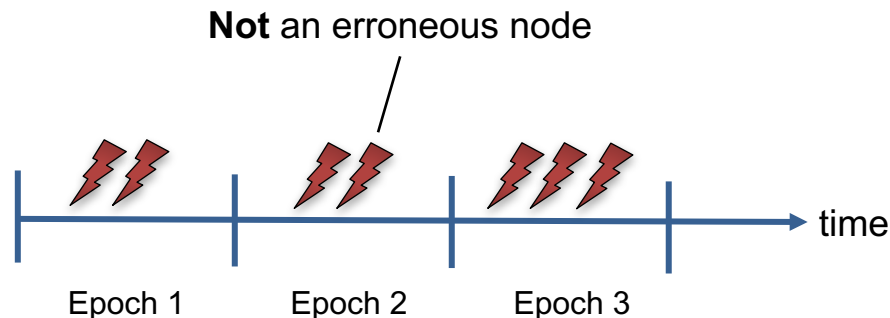
User-level Jobs to Gather ECC Errors Data

- We use EDAC (Error Detection and Correction)
 - Module in Linux for handling hardware-related errors
 - Ref: <https://01.org/linuxgraphics/gfx-docs/drm/driver-api/edac.html>
- EDAC counters are stored at /proc and are available to user-level programs
- 256-node jobs are submitted to collect EDAC information
 - 2 jobs submitted per day
- Job scheduler doesn't give node preferences

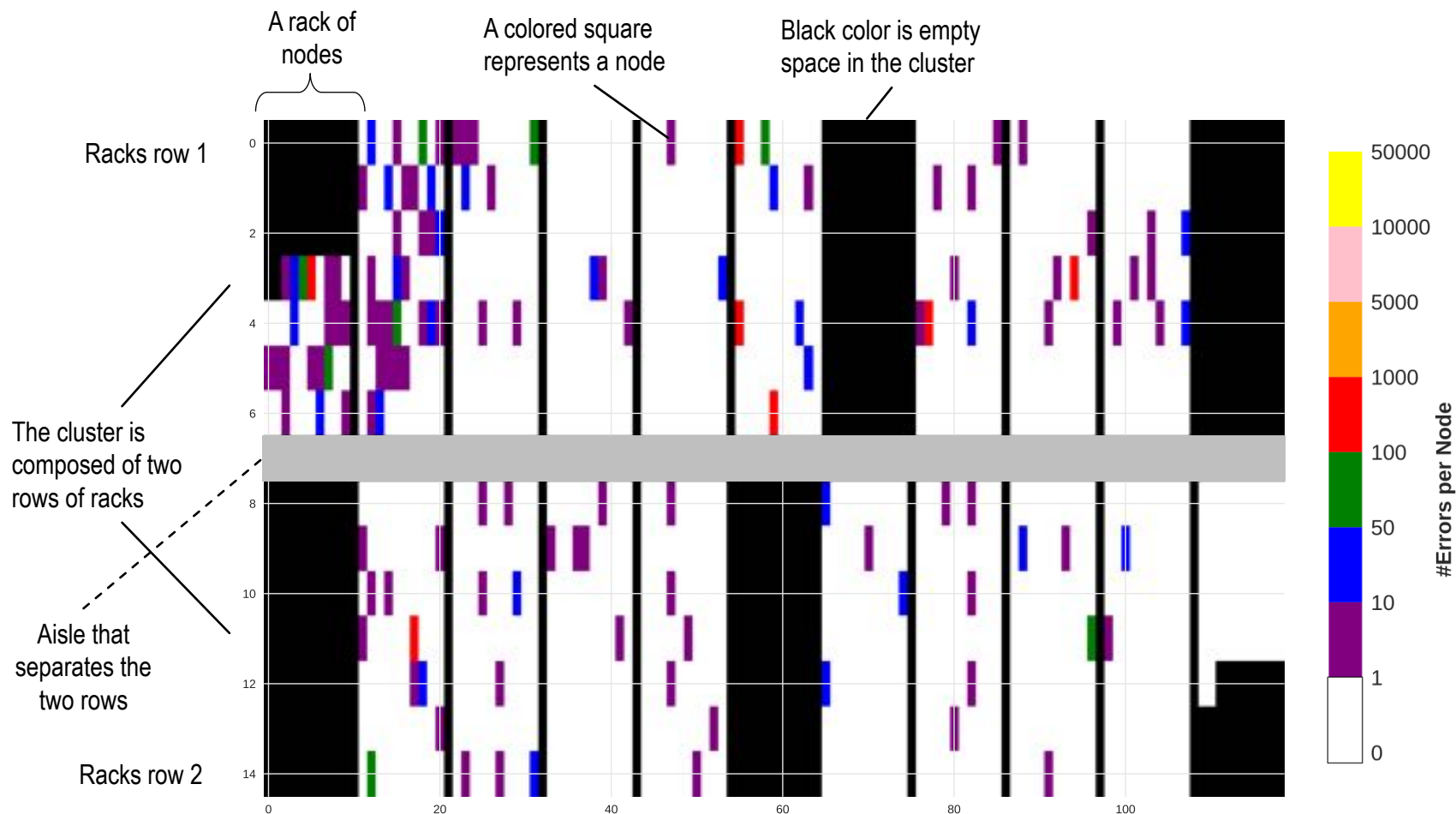


Terminology of the Study


- **Correctable Error (CE):** caused by transient, hard, or intermittent faults, which can be corrected using techniques like ECC.
- **Epoch:** defines a period of time in which errors are analyzed (one month)
- **Erroneous Node:** a node on which at least one correctable error was observed during the epoch being considered

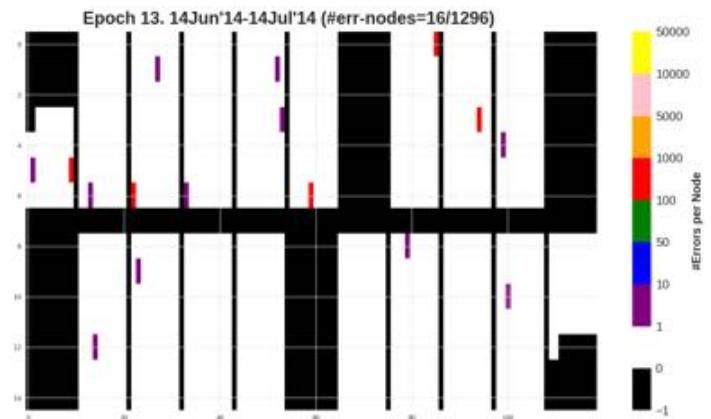
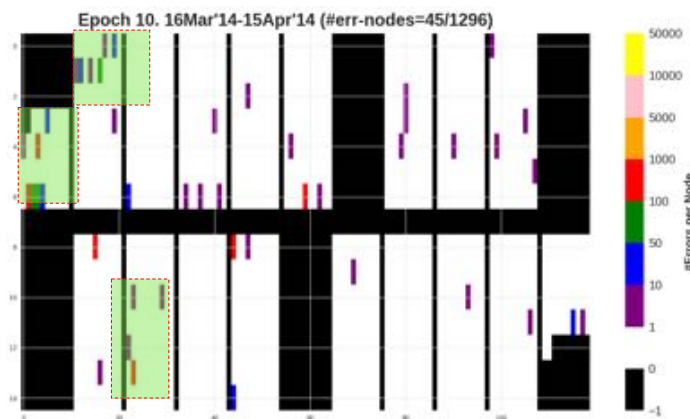
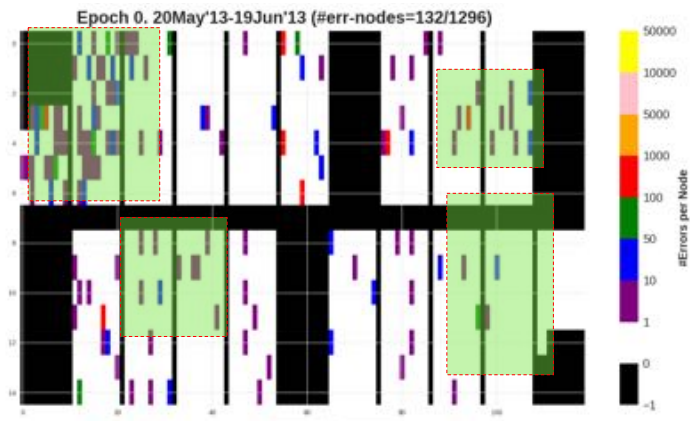


Physical Layout of the Cluster

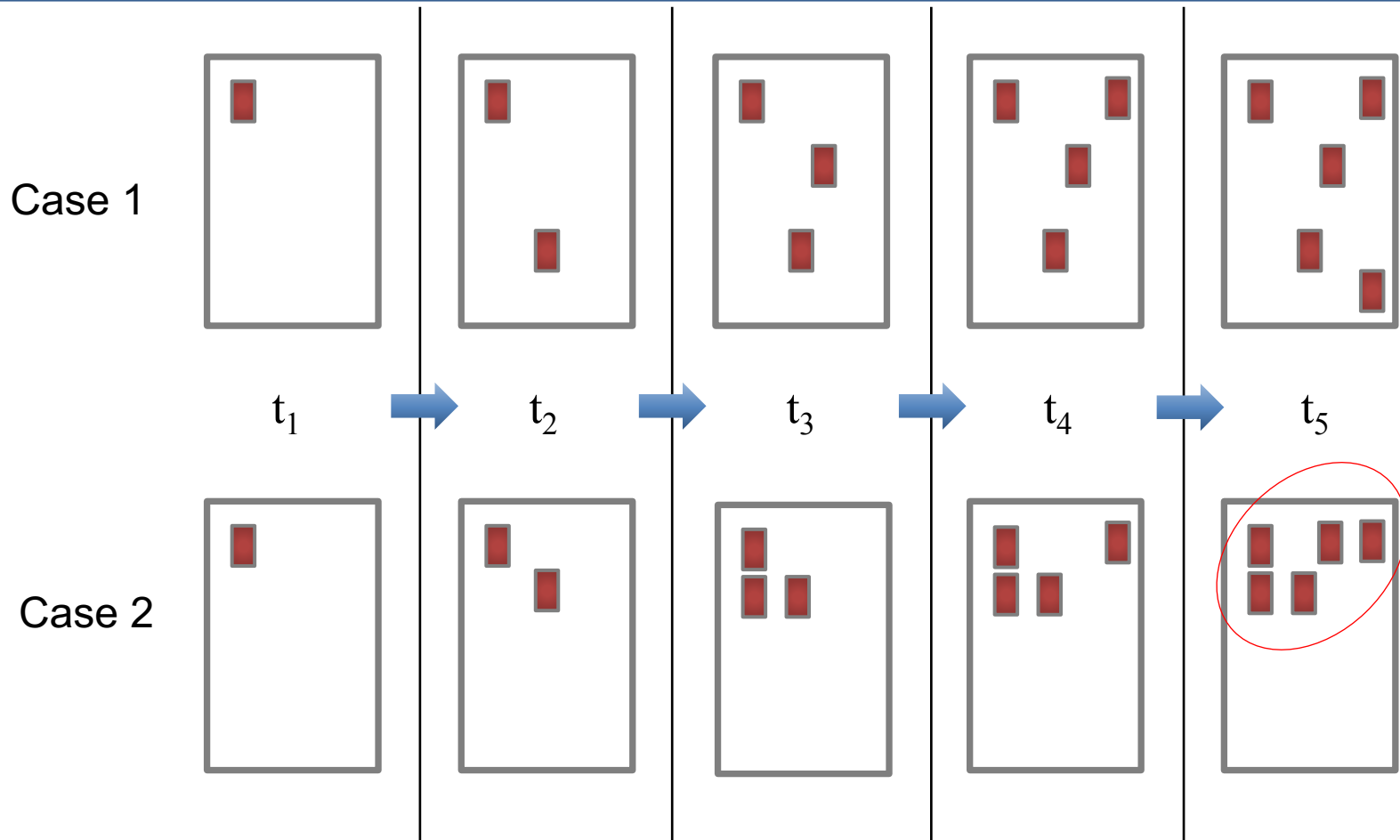


Examples of *Error Herding* (or Grouping)

 Error herd



Is the Grouping Occurring by Chance (Randomly)?



Both cases can occur by chance

We Perform Statistical Analysis to Understand if Grouping Occurs by Chance

- **Null hypothesis:** erroneous nodes are due to random occurrence (→ there is no actual spatial groups in the data)
- **Alternative hypothesis:** the opposite of null hypothesis

Question:

How do get samples of randomly distributed erroneous nodes?

We generate samples of erroneous nodes randomly using the same error rate of the gathered data

Methodology to Test the Null Hypothesis

1. Define a neighborhood of $M \times N$ sq. units
 - 1 unit = 0.2 feet (minimum distance between two nodes in physical layout);
 - We use $M = 4$, $N = 2$
2. Determine the number of **erroneous_nodes** in the neighborhood for each node
3. Generate a frequency distribution of **num_of_nodes** versus **erroneous_neighbors**
4. Generate a random sampling of **erroneous_nodes** for the epoch using the same error rate of the epoch
5. Perform a chi-squared test to test the null hypothesis
 - Significance level of 5%

Results of the Hypothesis Testing

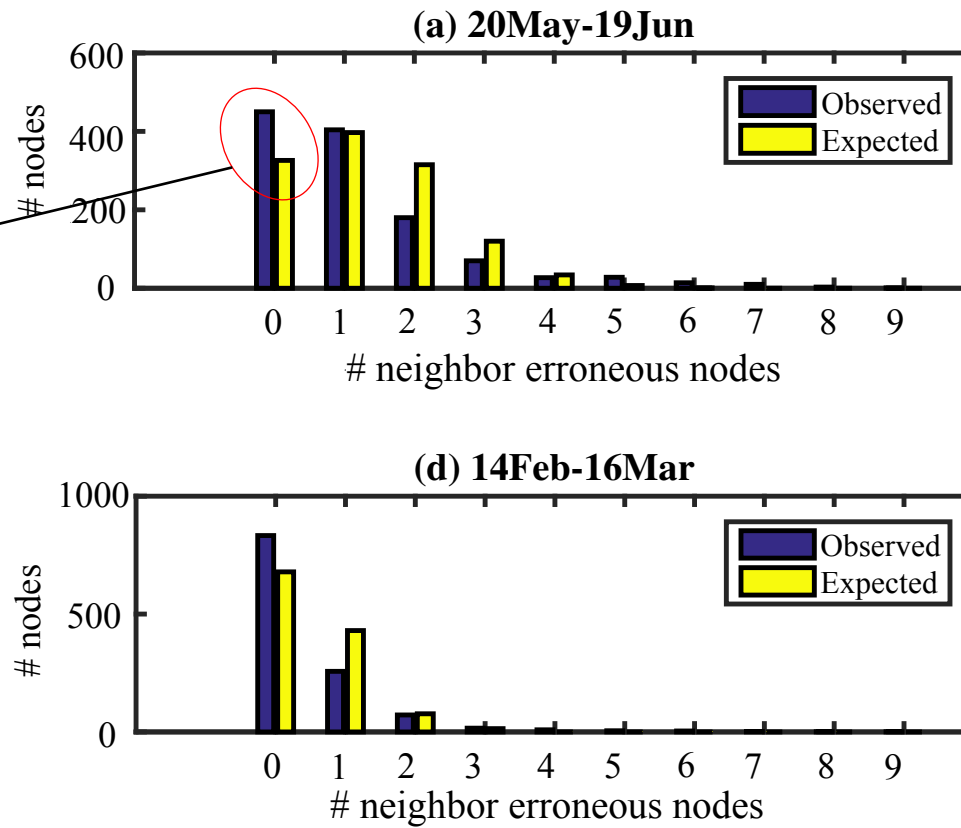
- Only 7 epochs had enough data samples for testing
 - 30 samples minimum
- In 5 out of 7 epochs, the p-value was less than 0.05
- We reject the null hypothesis in these 5 cases

Conclusion

In most of the cases, the spatial grouping of erroneous-nodes is not due to random occurrence and rather it is an effect of other physical or operating factors

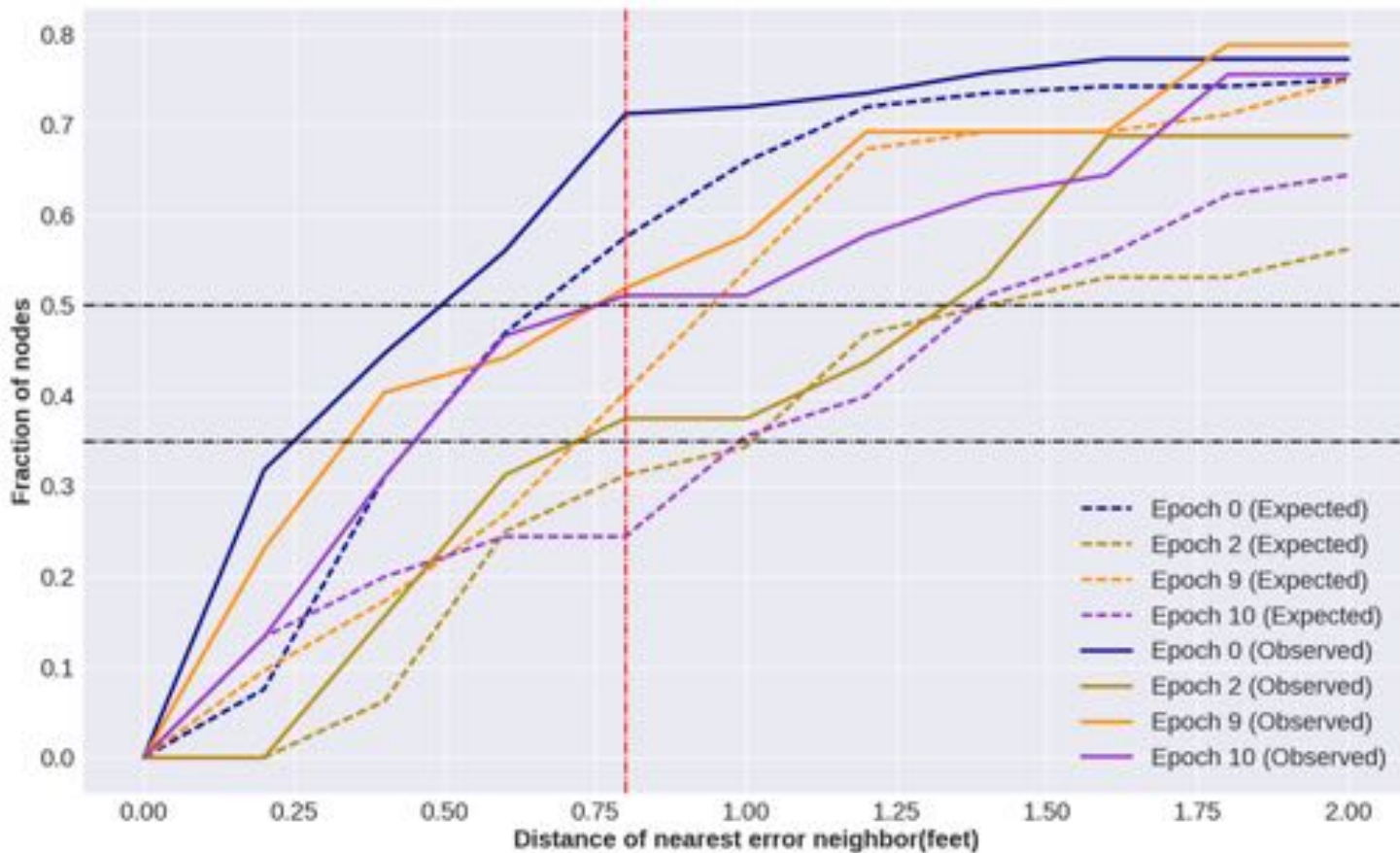
Distribution of Number of Neighboring Erroneous Nodes for the Nodes in Two Epochs

Expected data points: obtained from random sampling



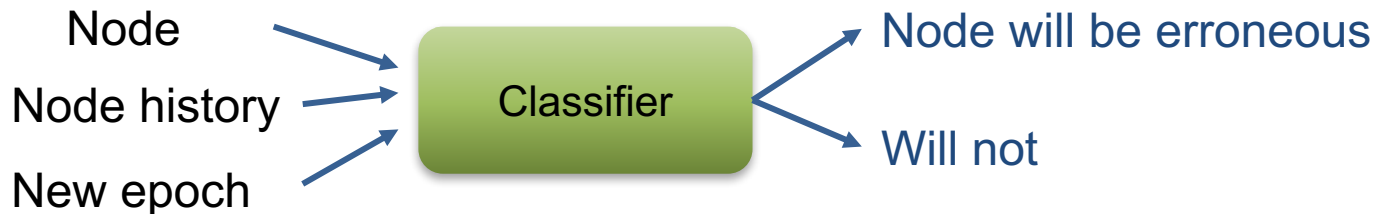
Significant differences is evidence that null hypothesis can be rejected

CDF Plots for Fraction of Nodes Having an **Erroneous Neighbor** Within a Given Distance



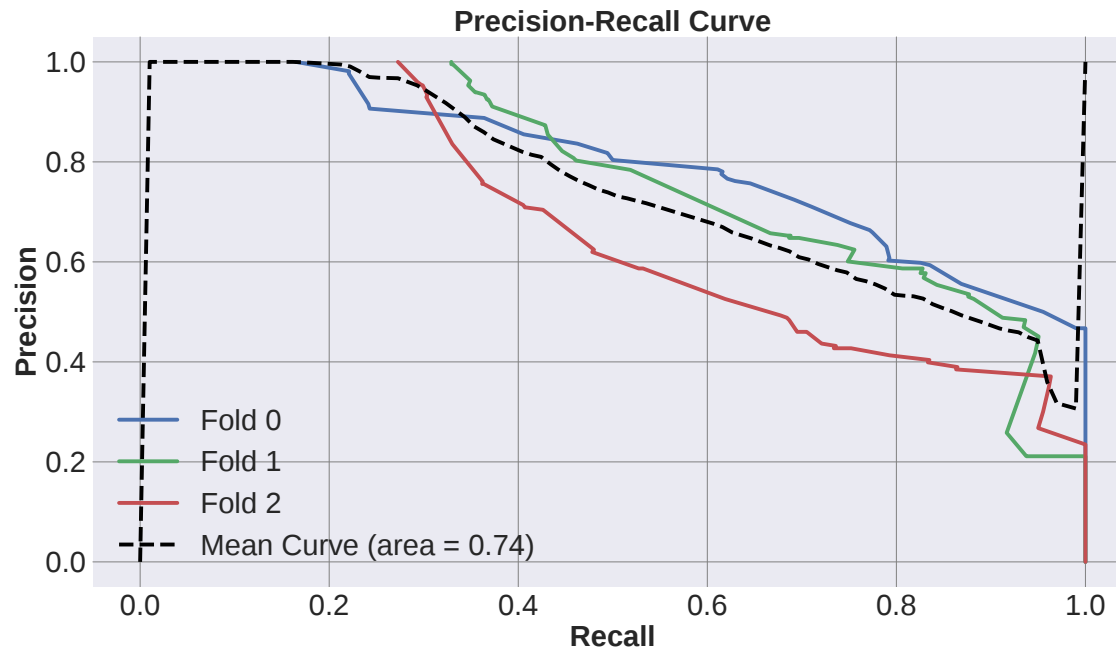
Prediction of Erroneous Nodes Using Machine Learning

- Being able to predict future erroneous nodes would be of great value
 - Job allocation, job migration, and overall system management
- Classification model that can predict if a node will be erroneous at a given timestamp t
 - Given the history of the state of its neighborhood, for time $[t-1, t-2, \dots, t-n]$
- We use 6 numeric features



Precision vs. Recall Curve for the Classifier

- 3 folds of cross-validation for the classification task
- The values shown are only for the error class (minority)
- For the majority class precision and recall are always very close to 1



In Summary

- 1 We show insights into the spatial correlation of DRAM errors in an HPC cluster and show that this **phenomenon is not due to random occurrence**
- 2 We show that nodes that experience a high degree of correctable errors **can be spatially correlated** in certain racks of the cluster
- 3 **Future work:** correlate spatial groups and epochs with other cluster data temperature? workload?



Questions?

