Topology and behavior aware failure prediction for Blue Waters

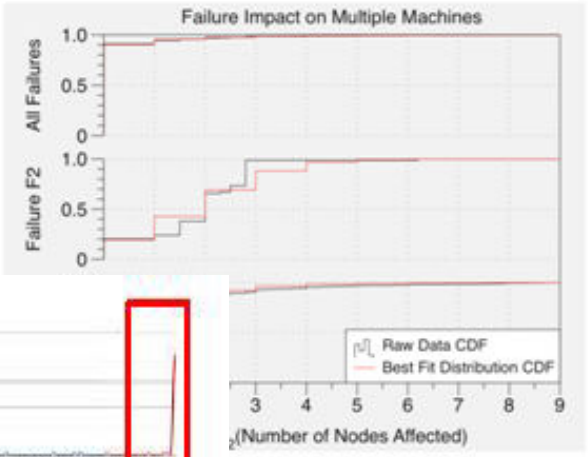
Ana Gainaru

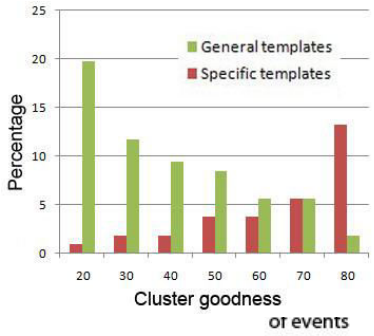
Franck Cappello, Marc Snir, Bill Kramer

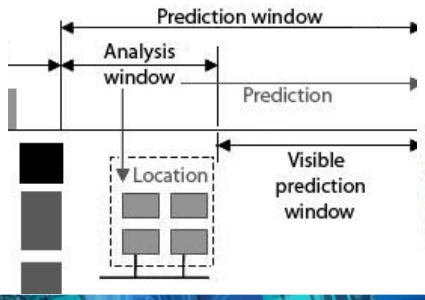
Timeline

3rd JLPC June 2010 | 4th JLPC Nov 2010 | 5th JLPC June 2011 | 6th JLPC Nov 2011 | 7rd JLPC June 2012 | 8th JLPC Nov 2012 | 9th JLPC June 2013 | Now

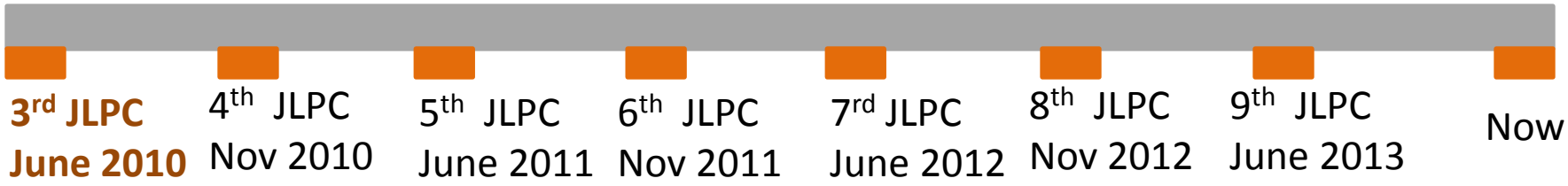
- 8 JLPC workshops
- Failure analysis and prediction



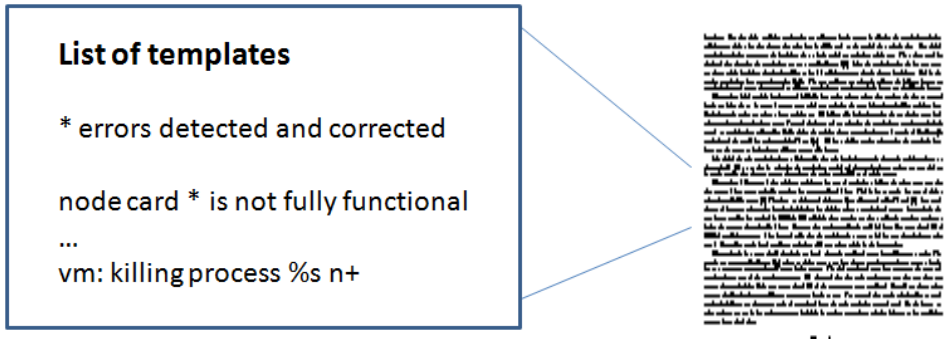




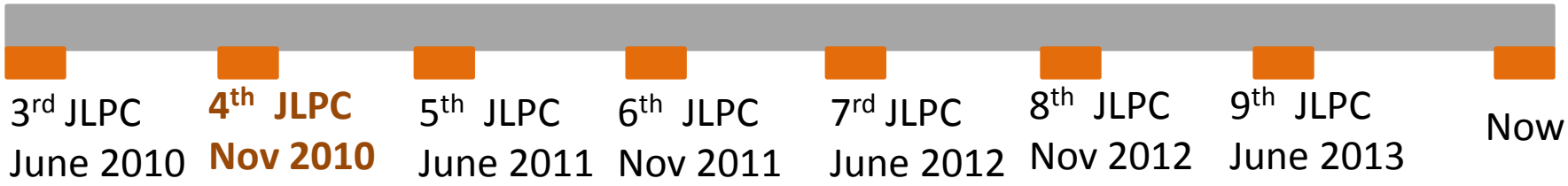
Timeline



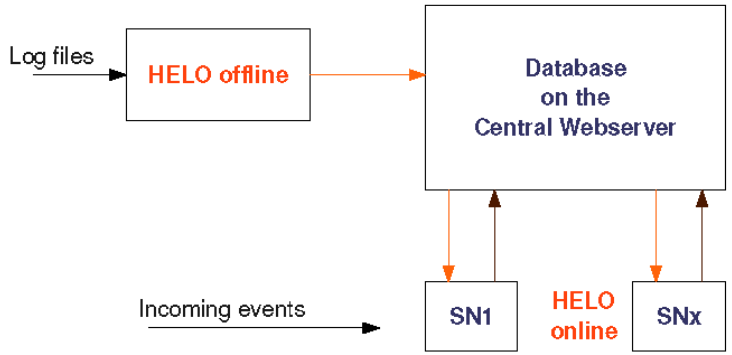
- Event log classification tool for large-scale systems
  - HELO tool
  - Classify messages
    - Create events
  - Paper at Europar 2011



Timeline



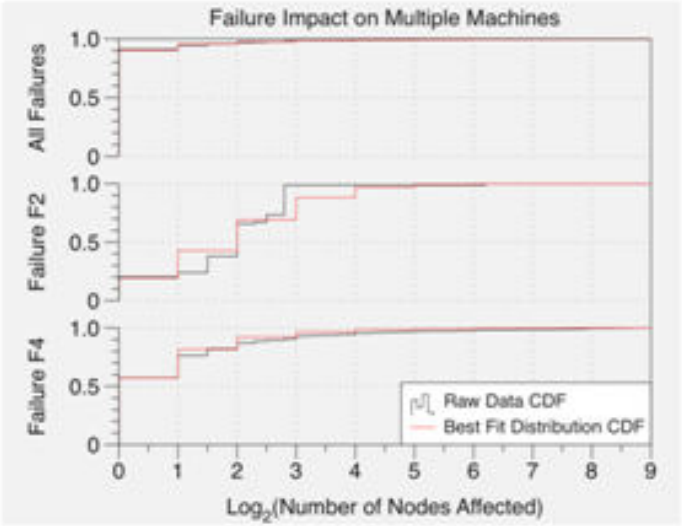
- Framework for Event Log Analysis in HPC
  - Work done with NCSA
    - Parallelize HELO on IBM service nodes
  - First example of correlation
    - Found in Blue Gene/L
  - Demo at SC 2010



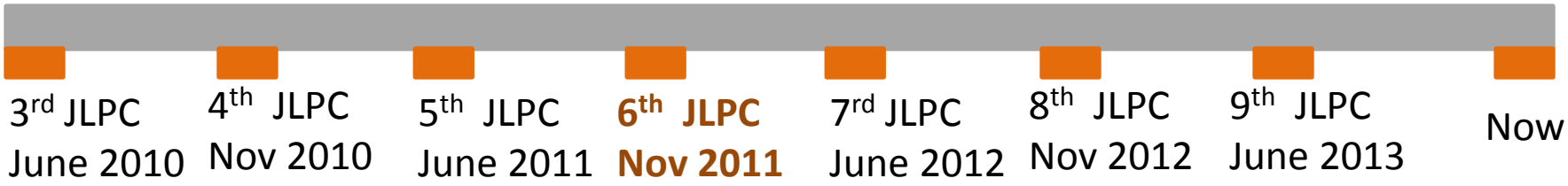
Timeline

3rd JLPC June 2010 | 4th JLPC Nov 2010 | **5th JLPC June 2011** | 6th JLPC Nov 2011 | 7rd JLPC June 2012 | 8th JLPC Nov 2012 | 9th JLPC June 2013 | Now

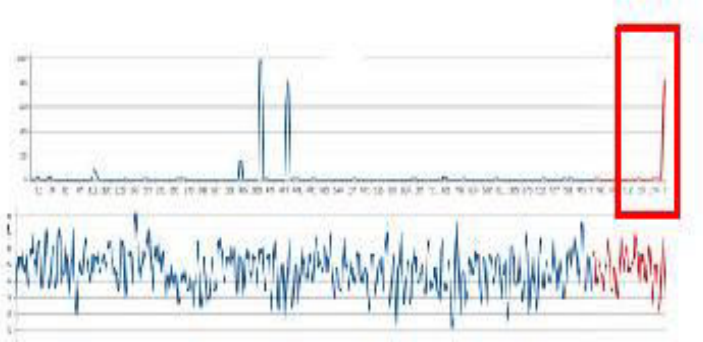
- Modeling and Tolerating Heterogeneous Failures in Large Parallel Systems
  - Work done with Eric Heien and Derrick Kondo
    - Analyze failures on NCSA's Mercury
    - Different failures have different behaviors
  - Paper at SC 2011



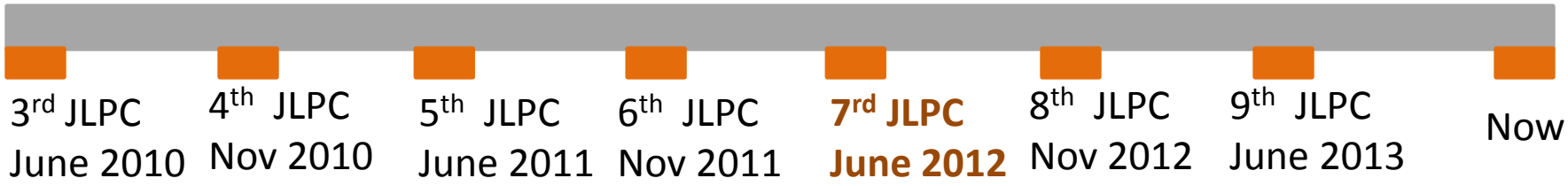
Timeline



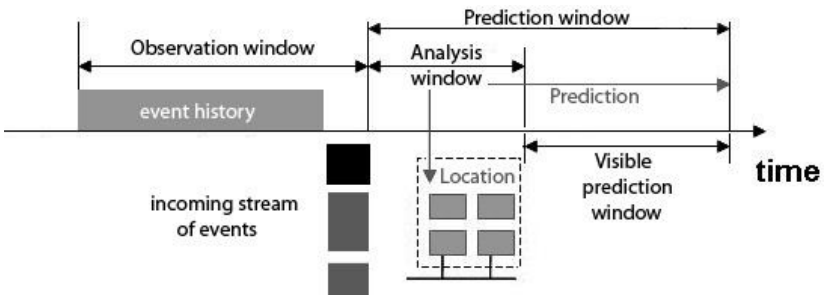
- Signal Analysis for Modeling the Normal and Faulty Behavior of Large-scale Systems
  - Signal analysis modules (ELSA)
    - Used to detect anomalies
    - Experiments done on the LANL system (public traces)
  - Paper at IPDPS 2012



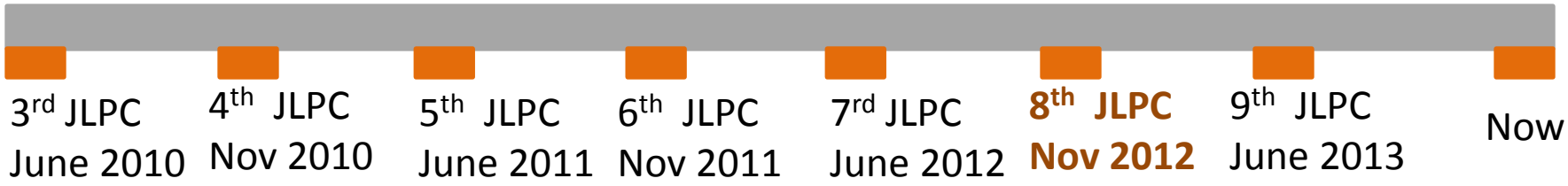
Timeline



- A detailed analysis of fault prediction results and impact for HPC systems
  - Combine signal analysis with data mining
    - Break down on different event types
    - Experiments done on the Blue Gene/L (public traces)
  - Paper at SC 2012

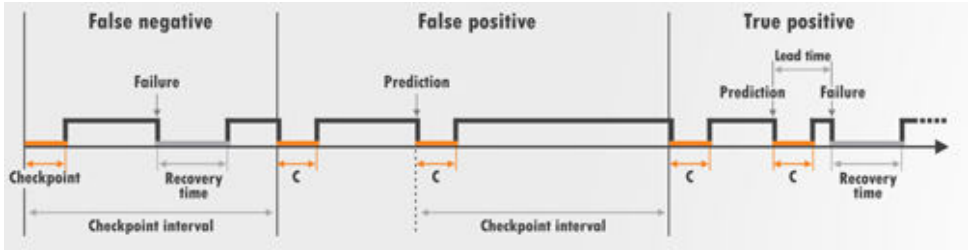


Timeline

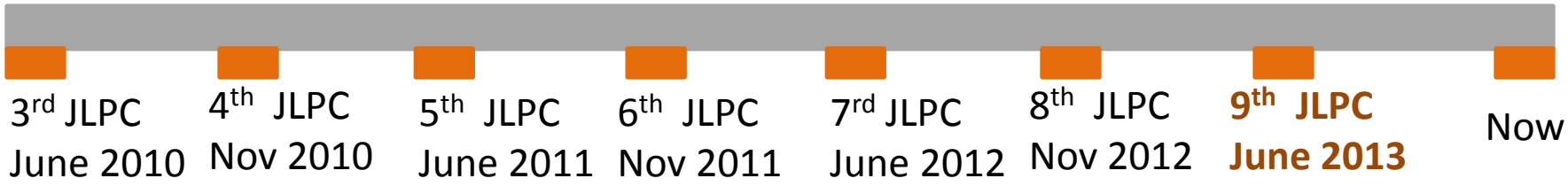


- Coupling failure prediction, proactive and preventive checkpoint
  - Combine ELSA with FTI
    - Measure the overhead on Tsubame 2.0 with the Gadget 2 application
    - Mathematical model

    by Slim Bouguerra
  - Paper at IPDPS 2013

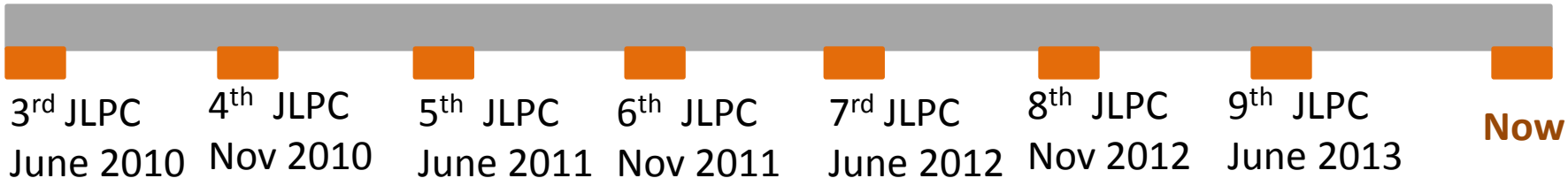


Timeline



- Challenges in predicting failures on the Blue Waters system
  - Online failure prediction
    - Results on BlueGene/L:
      - 50% recall 80% precision 10s lead time 3 months of training
    - Blue Waters: ~20% recall

Timeline



- Topology and behavior aware failure prediction for Blue Waters
  - Optimizations to increase the prediction results
  - Focus:
    - Multi-node failures
    - Application failure prediction
  - Submitted to IPDPS 2014

Quick reminder

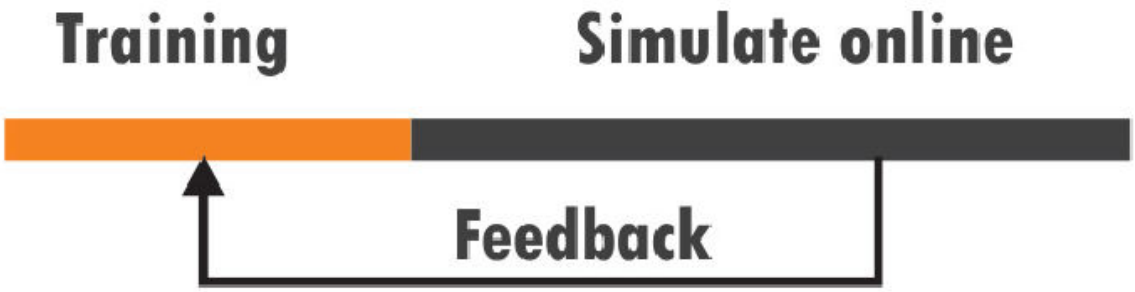


Figure 1. Failure prediction: simulate online

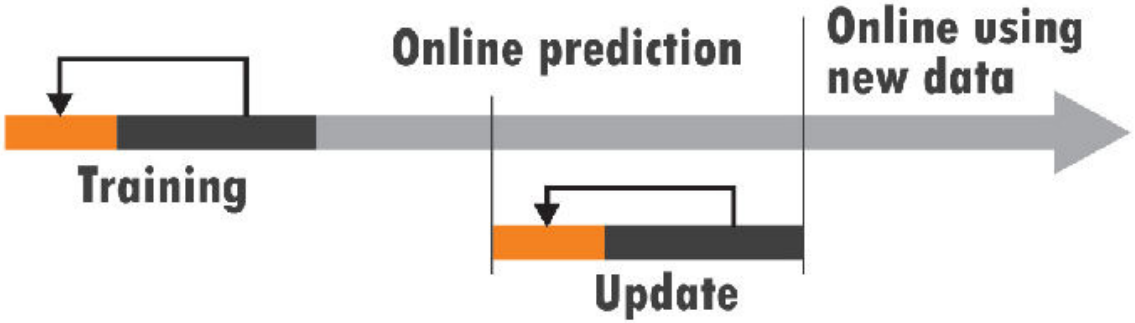


Figure 2. Online failure prediction

Quick reminder

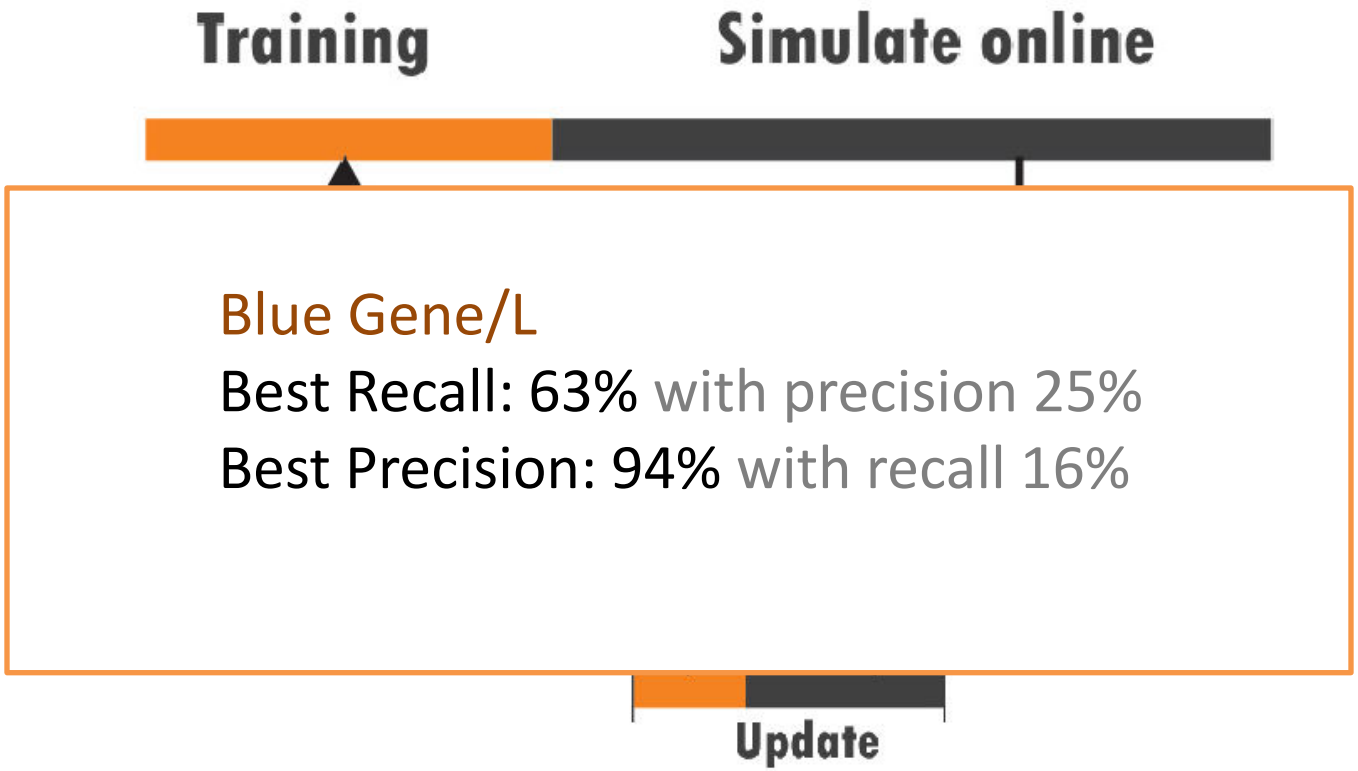


Figure 2. Online failure prediction

Online prediction on BW

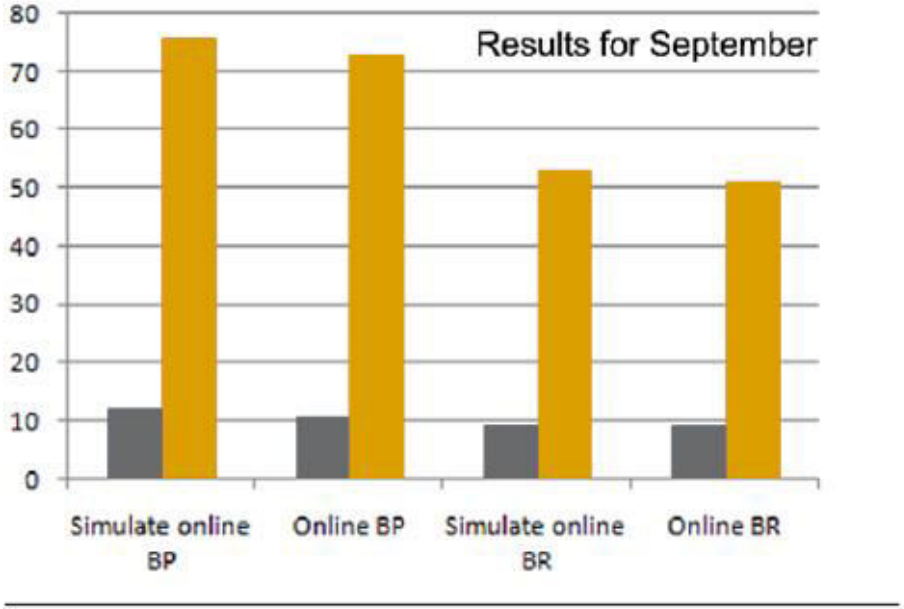
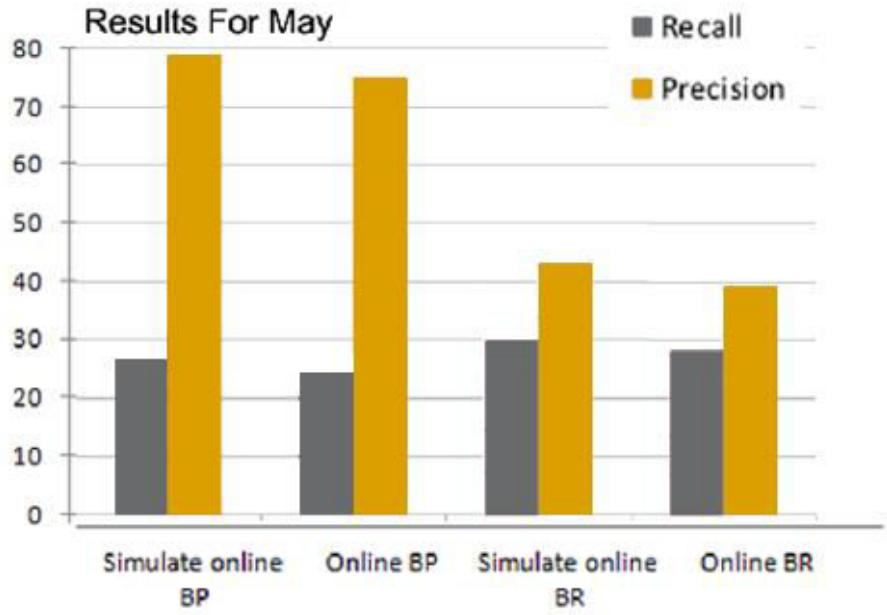


Figure 6. Precision and recall for the Blue Waters



- In August 2013, Blue Waters was upgraded with 12 additional Cray XK racks, each with 96 nodes

Limitations

- Location propagation
  - Over 90% of our predictions for multi-node failures do not succeed in discovering all the nodes in the fail set

- Locations on Blue Waters:
  - c2-1c2s2n0
  - For multi-nodes that are incorrectly predicted
    - Predict the slot/cage/cabinet

Location propagation

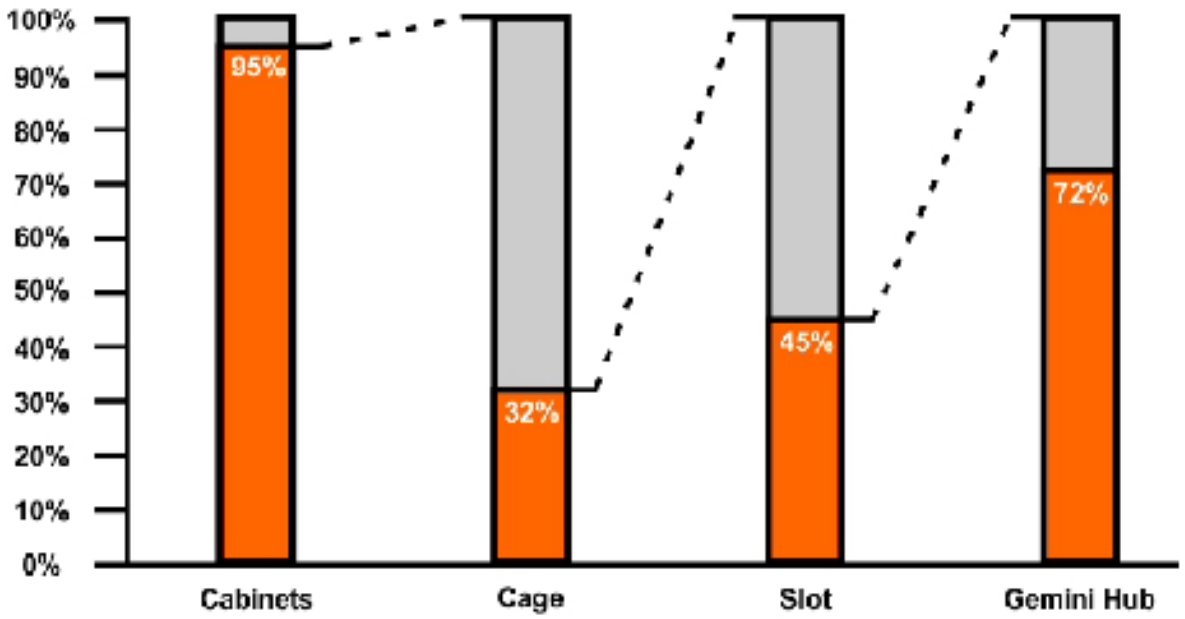
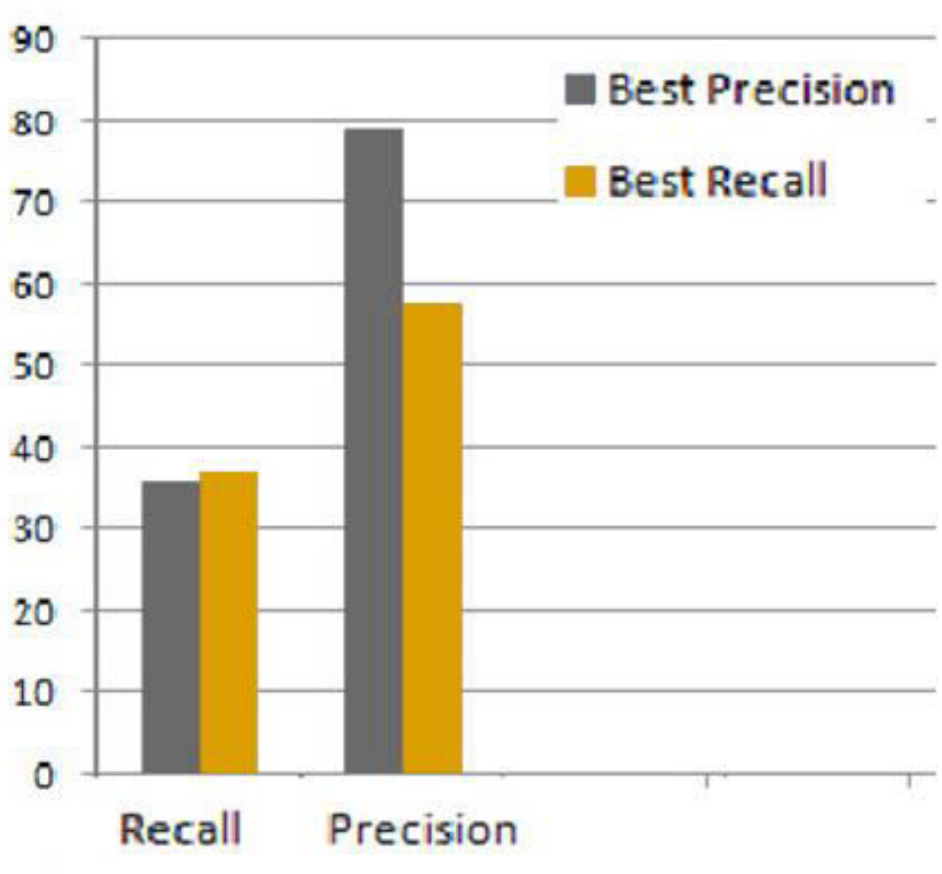


Figure 9. Location propagation results

Location propagation

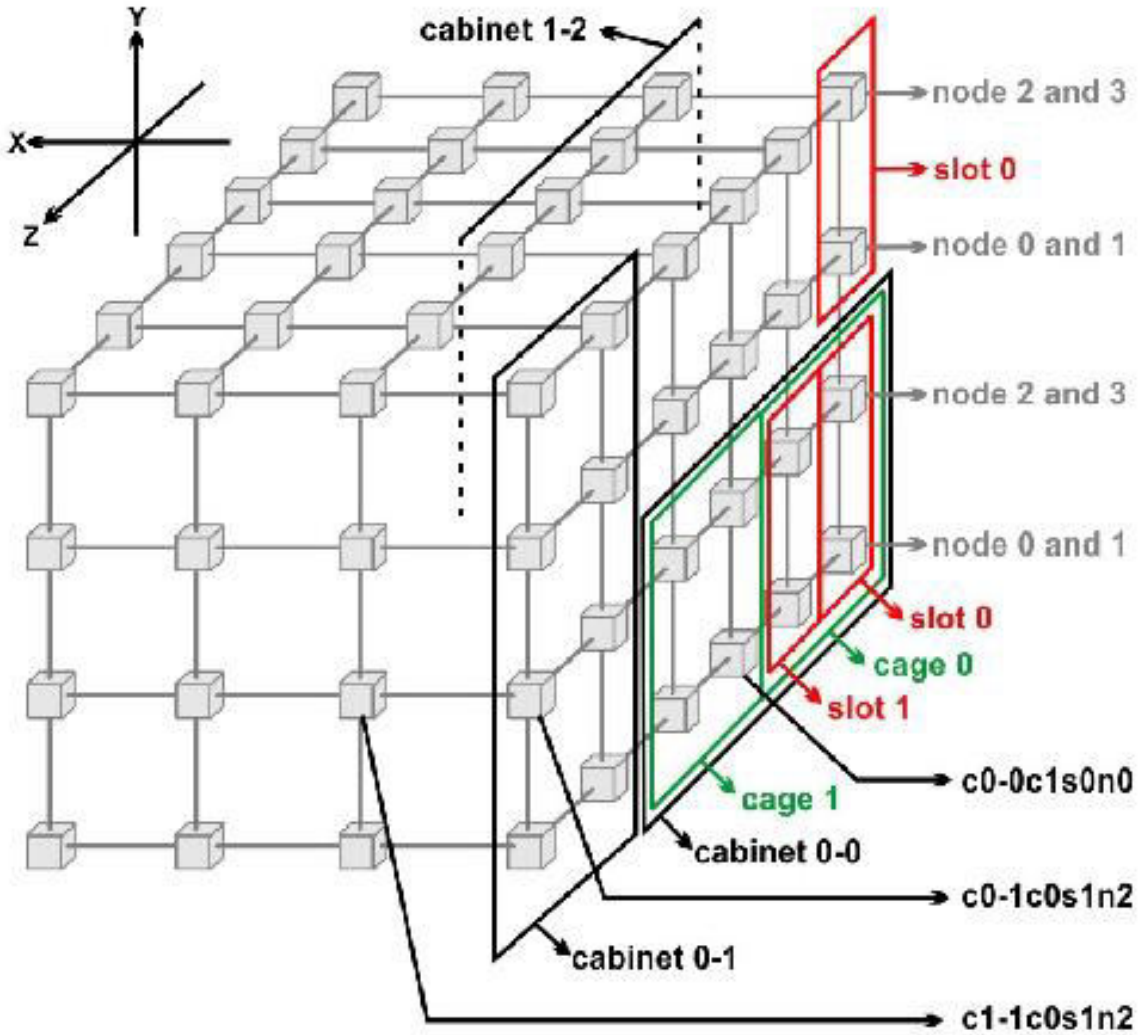
- c2-1c12s2n0
  - 4 nodes in one slot
  - 8 slots in one cage (32 nodes)
  - 3 cages in one cabinet (96 nodes)
- Over-predicting failing nodes



Topology aware

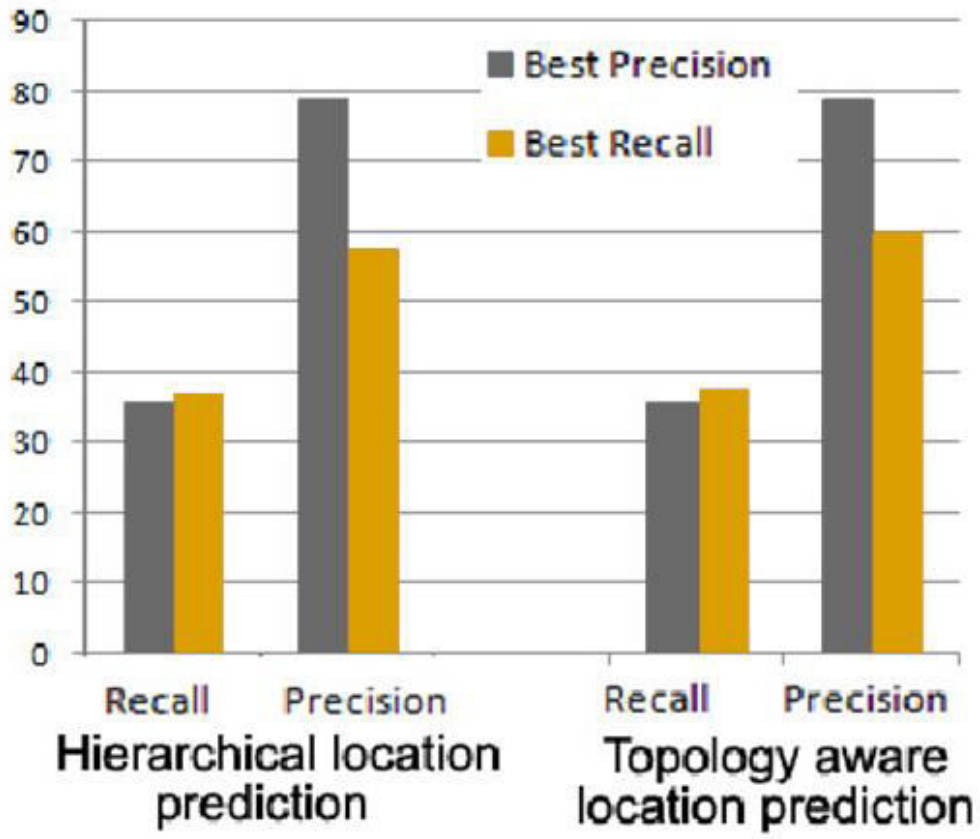
23x24x24 3D
torus network

Total 276 cabinets.



Topology aware

- Reduces the node over-estimation
  - By 15%
  - Future work – better patterns

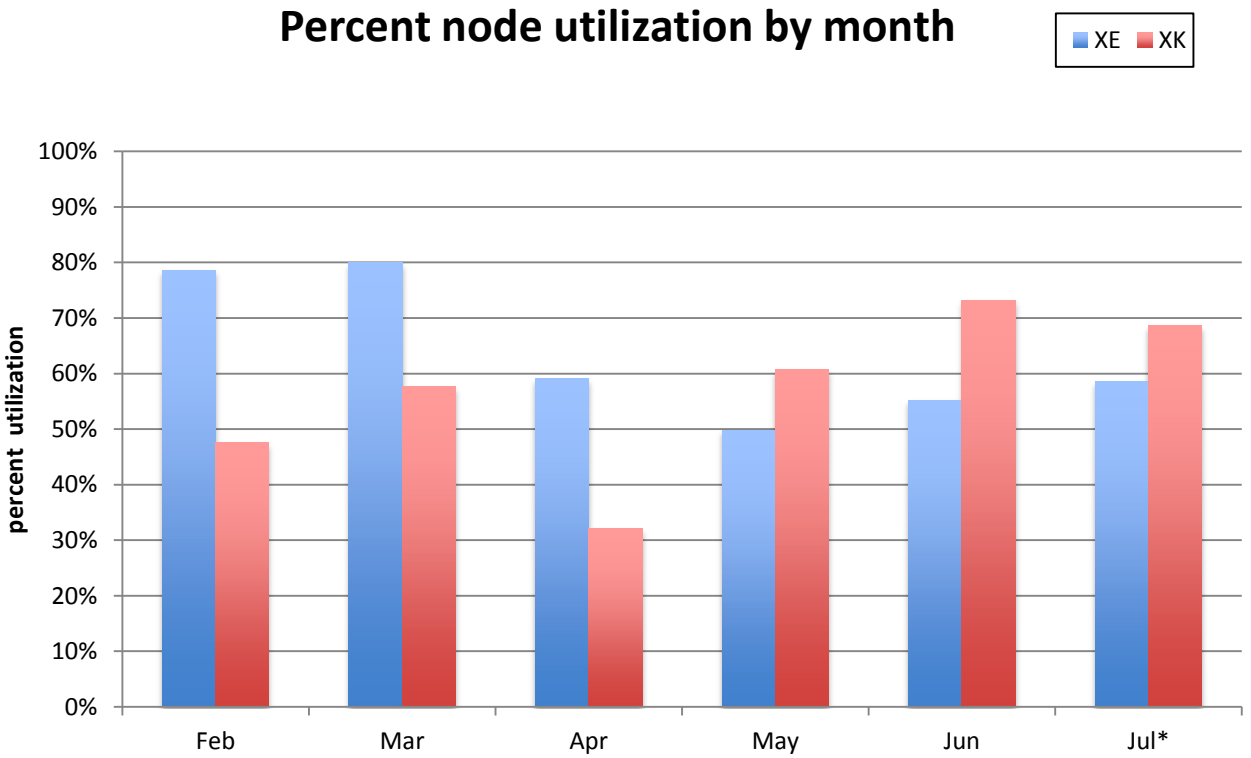


Application level

- Depending on system usage
- Depending on failure type
  - Crashed nodes do not affect jobs

- Lead time might be smaller/greater

Blue Waters utilization



- 237 Cray XE6 and 32 Cray XK7 (12 after August)

Application failure prediction

- Only around 44% of failures lead to at least one application crash.
  - 62% of the failure types predicted lead to application crashes
  - Corresponds to an increase in the recall of 5%
    - 40% when we use topology aware prediction
- Lead time depends on the application type

Application failure prediction

- Luster are the most frequent system failure
  - Only 5-10% lead to app crashes
  - ELSA was unable to predict location
- The first DIMM failure is not predicted
  - Subquential DIMM failures are captured

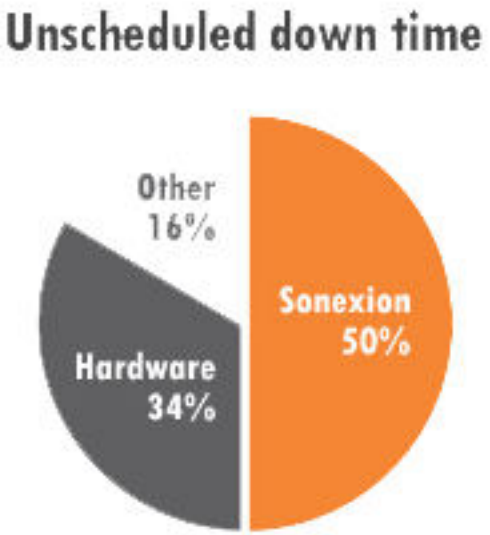


Table 2. Frequency of Special Characters

Failure type	Percentage	Recall	Application Crashes	Application Crash Recall
Luster MDT Failure	39.6%	7%	5%	0%
Luster OST Failure	16.3%	15%	13%	0%
DIMM Failure	15.7%	38%	11%	58%
Compute Blade	2.9%	62%	21%	64%
PBS Out-of memory	3.6%	44%	0%	0%

Application failure prediction

- Application and system level predictions are different
  - Most of the system failures are seen as performance degradation at the application level
  - Could predict app degradation?

- Better understanding of the topology of the system can increase app failure prediction

Conclusion

- System level prediction
  - Blue Waters is still young
  - Using topology and system information improves the accuracy
- Application level prediction
  - Understanding different error types
  - The recall value is better than for system level

Future work

- Understand app performance degradation
  - Analyzing IO patterns of an application we could predict file system degradation (and failures)
    - GPFS at Argonne
  - App migration on detecting/predicting degradation trade-off
- Increase the current results
  - For both system level and application level prediction

Additional Q&A

Thank you

Ana Gainaru

againaru@illinois.edu