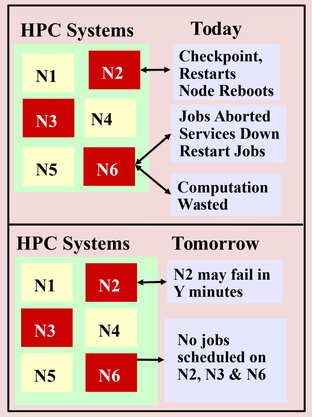
Desh: Deep Learning for HPC System Health Resilience

Anwesha Das[1], Frank Mueller[1], Charles Siegel[2], Abhinav Vishnu[3]
[1]North Carolina State University [2]Cray Inc. [3]Applied Micro Devices Inc.

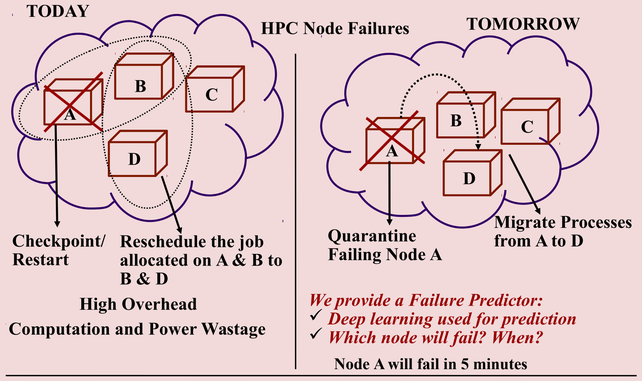
Motivation

- Currently HPC systems use reactive approaches for failure recovery
- Checkpoint/Restarts incur significant overhead requiring global co-ordination
- Anomaly detection schemes analyze faults after failure manifestation
- Combination of reactive and **proactive** fault tolerant solutions are required with **lead times** to failures
- **Reduce** computation and power wastage in large-scale compute clusters



Goals

Proactive Fault Tolerance



Background

Failures in HPC Systems

Petascale

- Longer MTBFs (Mean Time Between Failures, hours)
- Less Component Count
- Less Scalable Solutions
- Optimized Checkpoint/Restart

Exascale

- Shorter MTBFs (minutes)
- Higher Component Count (~10^6)
- More Frequent Failures
- Scalable Solutions
- Reduce computation and energy wastage

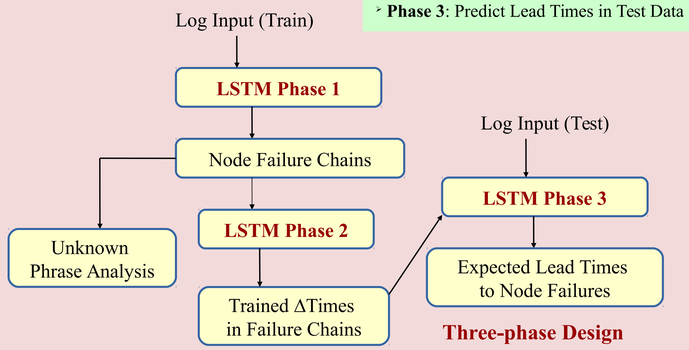
Goals:

- **Predict short lead times, pin-point the failure location**
- **Explore scalable unsupervised log mining techniques**

Solution Paradigm

Desh Architecture

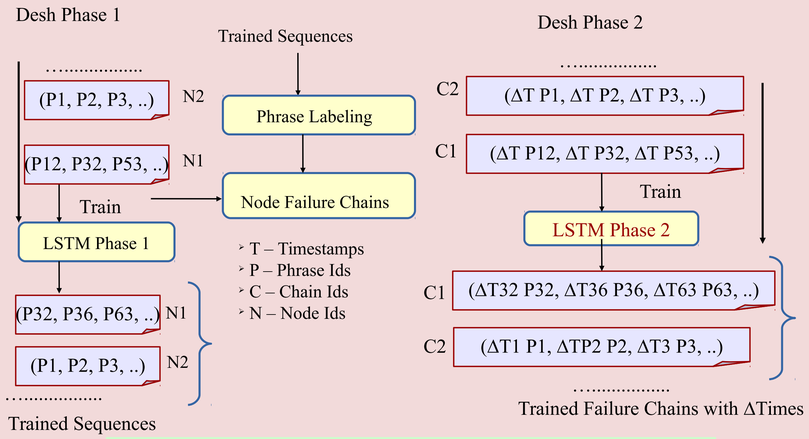
- **Phase 1**: Learn Failure Chains
- **Phase 2**: Learn Time Differences in Failure Chains
- **Phase 3**: Predict Lead Times in Test Data



Solution:

- **Learn the sequence of events over time from multiple nodes**
- **Learn the time differences between the events in the seen failure chains**
- **Predict expected lead times during testing based on the trained failure chains**

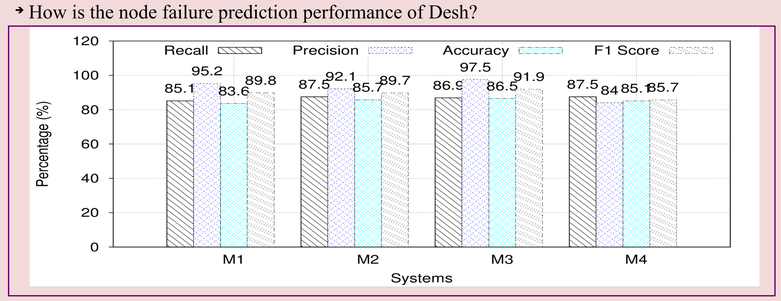
Solution Design



- **Node-wise logs sequentially fed to LSTM**
- **Learn sequence of events leading to a failure and the ΔTs in failure chains**

Results

How is the node failure prediction performance of Desh?

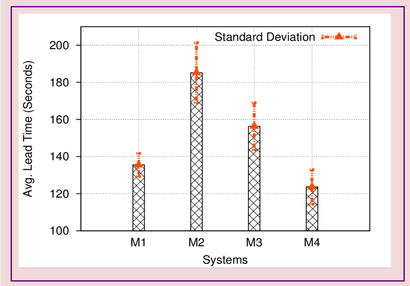


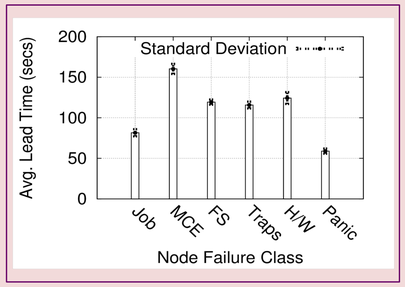
Implications:

- Overall high Accuracy and F1 score
- Recall ≥ 85%, Precision ≥ 84%, Accuracy ≥ 83%, F1 score ≥ 85%
- 83.6 ≤ Accuracy ≤ 86.5% (Matters more than Precision !!)
- Less False Positives and False Negatives

Results

How does the lead times vary across diverse failure classes and systems ?



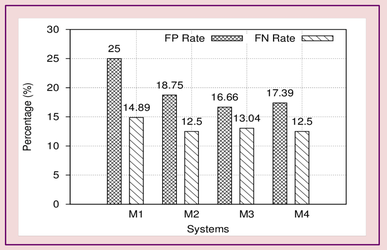


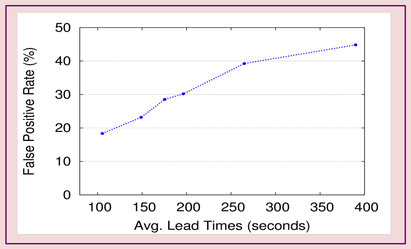
Implications:

- > 2 minutes avg. lead times (LHS)
- High average standard deviation (± 16.5 seconds) (LHS)
- System: More variety of failure classes
- Lead times of a single Failure Class are similar (reproducible) with low standard deviation
- MCE and H/W errors → Higher lead times (RHS)
- Kernel Panics → Short lead times (RHS)
- Lead times of a specific system has high standard deviation, every system has different proportion of failure classes causing this variation

Results

How much are the incorrect predictions? How does the FP rate vary with increasing lead time?





Implications:

- FP Rate: 16.66 to 25% across all the systems
- FN Rate: 12.5 to 18.75% across all 4 systems
- For lead times of 105 to 196 seconds: FP Rate is 18 to 30%
- With higher lead times of 6.48 mins (389 secs) FP Rate rises to 44 %
- 2 minutes lead time achievable with acceptable False Positive Rate

Conclusion

- Desh: 2 to 3 minutes lead times to node failures
  - Accuracy ≥ 83%, F1 score < = 89.99%
  - FP Rate across systems:16.66% to 25%
  - Proactive Actions: Node cloning (90 secs), Process-level job migration (13 to 24 secs), Quarantining unhealthy nodes
- Lead Time Sensitivity with Failure Classes
  - Lead times to *Kernel Panics* caused failures are short (58.87 secs)
  - *MCE* and *Hardware-caused* failures have high lead times (124 to 160 secs)
  - Lead time variations *higher across systems* than failure class
- Unknown Phrase Analysis
  - Erroneous phrase benign in one chain, fatal in the other in the context of causing failed nodes
  - Severity levels of a single message insufficient, indicator of component failures; event chain analysis required

Acknowledgments: This work was funded in part by subcontracts from Lawrence Berkeley and Sandia National Laboratories, Air Force Office of Scientific Research grant FA9550-12-1-0442, and NSF grants 1217748 and 0958311. Any views expressed in this paper are those of the authors and do not necessarily reflect the views of AMD, Cray, NSF, or any other national labs.