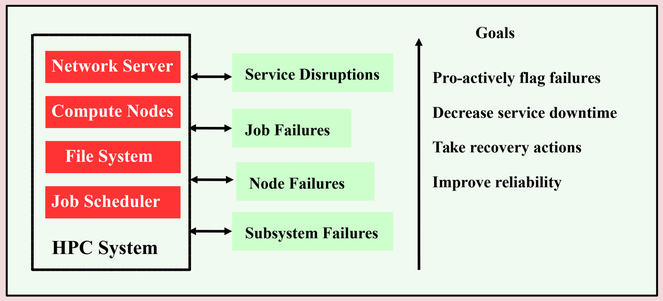
Desh: Deep Learning for HPC System Health Resilience

Anwesha Das[1], Abhinav Vishnu[2], Charles Siegel[2], Frank Mueller[1]
[1]North Carolina State University [2]Pacific Northwest National Lab

Motivation



HPC systems: ***Increasing component size, Evolving design complexity, System logs diverse, Log mining difficult***
Resilience: ***Large-scale processing, Insufficient understanding of failure indicators***
Requirement: ***Log investigation, clarity about log messages, their implications***
– Anomaly prediction, Service disruption prevention

Problem

System Logs

ERROR: Type:2; Severity:80; Class:3; Subclass:D; Operation: 2

SMM IPL failed to set SMRAM window to EFI_MEMORY_WB

AER: Multiple Corrected error received

mcelog: failed to prefill DIMM database from DMI data

What phrases aid failure indication ?

CorrectableMemErr Link CRC error (cnt: 4)

db_hook (pid 54378) stdout: No job records are eligible to be pruned.

How to capture time sensitive dependencies ?

Kernel crash occurs 20 seconds

Lustre failure messages for 10 minutes

Link control block failure in 5 minutes

How to scale log mining ?
How to predict failures with high lead time ?

Challenges - ***Can failed event truly indicate failure ? How to distinguish real failures from noise and benign events ? Is a scalable automated framework possible ?***
Goal - ***Investigate deep learning techniques such as LSTM for HPC system failure prediction, Research methods to scale training phase of logs and predict sensible events.***

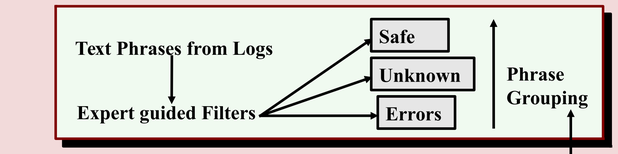
Background

- Past Research: Anomaly detection/prediction for older HPC systems
  - Past Logs: Comparatively more structured
  - Past Focus: Statistical Analysis, Inadequate stress on text semantics & lead times
- Contemporary HPC systems: New format, unstructured text logs
  - New scope: Natural Language Processing (NLP), Deep Learning [3] based Techniques
- Past Techniques:
  - Logistic regression, PCA (principle component analysis) [4], Event correlation, Probabilistic Model and Markov Chain based mechanisms
    – Feature extraction: Supervised or easier to do labeling
  - Support Vector Machines (SVMs) [1] & Sequence Mining [2] based mechanisms
    – Correlation extraction difficult for time-sensitive data dependencies

References:

1. Errin W Fulp, Glenn A Fink, and Jereme N Haack. 2008. Predicting Computer System Failures Using Support Vector Machines. WASL 8 (2008), 5–5.
2. Xiaoyu Fu, Rui Ren, Sally A McKee, Jianfeng Zhan, and Ninghui Sun. 2014. Digging deeper into cluster system logs for failure prediction and root cause diagnosis. In Cluster Computing (CLUSTER) 2014 IEEE International Conference on. IEEE, 103–112.
3. Adam Coates, Brody Huval, Tao Wang, David Wu, Bryan Catanzaro, and Ng Andrew. 2013. Deep learning with COTS HPC systems. In International Conference on Machine Learning. 1337–1345.
4. Zhiling Lan, Ziming Zheng, and Yawei Li. 2010. Toward automated anomaly identification in large-scale systems. IEEE Transactions on Parallel and Distributed Systems 21, 2 (2010), 174–187.

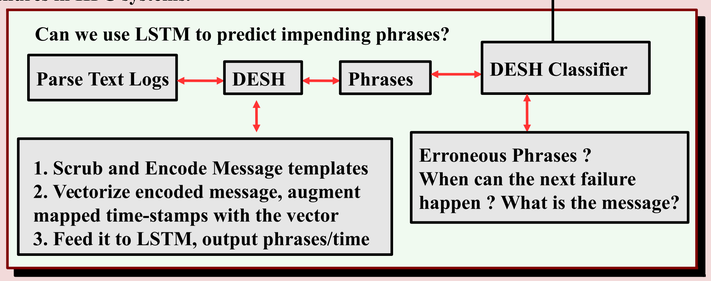
Solution Paradigm



Required ?

Efficient scalable solutions for failure prediction, proactive recovery

Deep learning: Efficient in vision and speech recognition,
Can it enhance reliability of HPC systems ?
Desh: Investigates LSTM (long-short term memory), to pinpoint failures in HPC systems.



Insights and Findings

- ***Unknown phrases: Not safe/error, further exploration required***
- ***Cray Logs: Desh is able to predict 13 to 20% of the erroneous phrases***
  – Next step: Quantify failures from unsupervised phrase prediction
- ***Filters: High cardinality of filters for grouping***
  – Discard noise (harmless information)?
  – Can we reduce cardinality?
- ***LSTM: Enhanced phrase structuring required***
  – Experiment with multiple-sized time bins

System Data Details

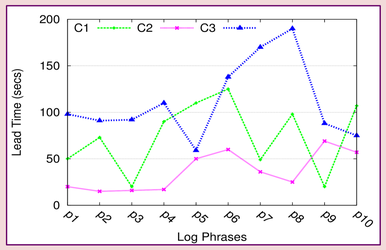
System	Data Size	Duration
Cray XC Cluster1 (C1)	22 MB	1 week
Haswell Cluster (C2)	100 MB	1 month
Cray XC Cluster2 (C3)	80 MB	3 weeks

FPR (False Positive Rate): Phrases which didn't appear in the test data, but Desh predicted, (depends on training set)
Lead Time: The correctly predicted phrases are cross validated in the data, to know how much ahead in time, the phrases actually occur (after the last trained phrase)

Results

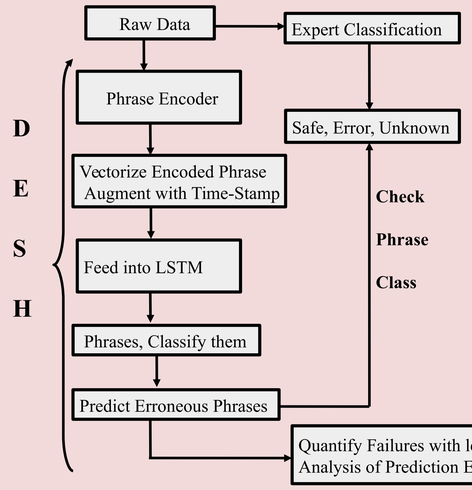
System	FPR (%)	Error (%)
Cray XC Cluster1 (C1)	5.2	20
Haswell Cluster (C2)	4.7	13
Cray XC Cluster2 (C3)	8.2	18

The false positive rate is low considering the diverse length phrases with static and dynamic contents across C1, C2 and C3.



Prediction: Desh predicts phrases which do occur in the future, More research required for analyzing time to quantify failures.
Time Variation: C2 Cluster data - less varied system files, message types, Lead times of C1 and C3 have higher mean deviation than C2 because of higher variety of log messages.

Desh Prototype



Why LSTM ?

- ***Can unlearn and relearn time-series data***
- ***Ability to capture long-term and short-term data correlations***
- ***Known to be efficient for large scale data processing***
- ***To understand whether Neural Networks unlike Markov Models can predict derived phrases learned from the training data.***

Conclusion

- ***Desh successfully classifies log messages based on semi-supervised failure prediction.***
- ***Desh predicts phrases efficiently using stateful neural networks (less than 200 secs/epoch for 80 MB data).***
- ***Desh predicts phrases which indeed occur in the test data ahead of the time.***
- ***Identified scopes to improve HPC system health considering phrase embeddings and semantics for better lead times.***

Future Work

- ***How little expert labeling can auto-classify the predicted phrases?***
- ***How to analyze the unknown class for understanding which phrases are mostly safe or part of an anomaly?***
- ***How to predict future time-series accurately to aid failure prediction with location information?***
- ***Comparative analysis of Desh with existing prediction techniques on multiple HPC cluster logs.***

Acknowledgments: Dr. Abhinav Vishnu, Dr. Charles Siegel and Dr. Frank Mueller for insightful guidance and helpful suggestions. The CSF and EMSL division of PNNL for cooperating with HPC cluster access and data sharing.