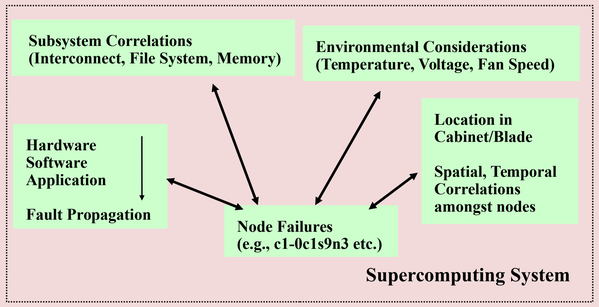
Holistic Root Cause Analysis of Node Failures in Production HPC

Anwesha Das, Frank Mueller --- North Carolina State University

Background and Motivation

Node Failures:

- Waste compute capacity and energy in HPC systems
- User job disruptions and **S**ystem **W**ide **O**utages (SWOs)



Investigate:

- ✓ What are the exact root causes?
- ✓ How sensor measurement deviations affect the functioning of blades?
- ✓ Does exploiting spatio–temporal locality of nodes enhance lead times?
- ✓ Do the symptoms of failures apply in non–failure cases as well?

Past Root Cause Diagnosis [2,3]: Subsystem correlations not considered, lead time analysis not well researched [4]

What is missing in the existing state-of-the-art ?

- *Hardware, software and application layers studied independently*
- *Focused study on a specific component provides a local view, global perspective remains unknown*

Challenges ?

- *Holistic considerations require knowing the implications of lower level vendor-specific log messages*
- *Symptom identification from several components and parameters to infer meaningful root cause*
- *Quantify lead time enhancements without increasing false positives or false negatives*

System-wide External Logs + Node Specific Internal logs

- ✓ Understand subsystem correlations on node failures (e.g., inter-, Intra-node dependencies)
- ✓ Understand the impact of sensor measurement deviations on cabinet/blade/chassis
- ✓ Holistic consideration of hardware, software and application events on diverse components for root cause diagnosis of node failures
- ✓ Enhance failure prediction schemes [1] by increasing lead times to failures

Research Goals

How well are spatial/temporal correlations indicative of the root cause?

By how much can the lead times increase if external factors are considered?

Node-related Logs

04:32:58.100694 socket * reports MCERR *
04:33:59.685959 cb node unavailable: * found in unavailable event

How to correlate?

Formulate Correlation
Identify Subsytem Malfunctioning
Spatial Temporal Analysis

System-wide Logs

Blade/Chassis-related events
Environmental Logs (Voltage, Temperature etc.)
SMW (Centralized workstation) Messages

- ➢ Evaluate increase in lead times
- ➢ Analyze inaccuracy in predictions
- ➢ Recommend mitigation approaches for long-term system health

Root Cause Diagnosis

Caveats and Pitfalls in Correlations

- *Inferring correctly why what happened is tricky:*
  - *Several software Traps can be a consequence of a hardware bug*
  - *Resource crunch caused by jobs can trigger too many interconnect faults*
- *Spatio/Temporal correlation needs more research:*
  - *File System, Interconnect errors affect nodes at similar times*
  - *Processor corruptions in a single day can be caused by jobs scheduled on nodes, spatially apart*
  - *How to infer real cause based on multiple tangible events?*

Solution Design

- Statistical analysis of root causes for node failures in HPC
- Estimate inaccurate predictions (false positives and false negatives) over a sample time-frame
- What are the conditions of trivial faults not leading to failures?
- Quantify increase in lead times w.r.t. the case when environmental influences are not considered
- Uncover insights to suggest mitigation approaches (proactive/reactive) for longevity of healthy conditions

- ➢ *Correlation based on not just symptoms but related implications of events*
- ➢ *Track down the root causes and their propagation across layers/components*
- ➢ *Will prevalent mitigation approaches fix the diagnosis result? Analyze what is the solution for the interpreted root cause?*

Production Logs Studied:

HPC System	Log Size	Duration	Cluster Scale
Cray XC Cluster1 (C1)	5 GB	2 weeks	5586 nodes
Cray XE Cluster2 (C2)	8 GB	2 weeks	6400 nodes

Cabinet
Blade
Node

→ **Move from fine-grained to coarse-grained, from lower spatial granularity to higher**

Node Events specific to nodes

→ **Unhealthy time frames with anomalous node shutdowns (failures)**

Sensor Measurements
Job events
SMW/Event Logs

→ **Identify external influences, derive indicative patterns for subsystem correlations**

Comprehensive understanding of how nodes fail?

Non-trivial implications of vendor specific low-level system logs ← **Consultation with system administrators and cluster management team**

Results

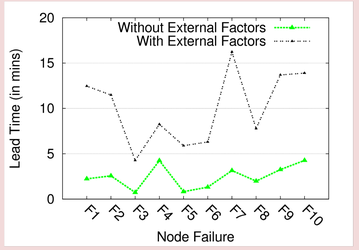


Figure 1

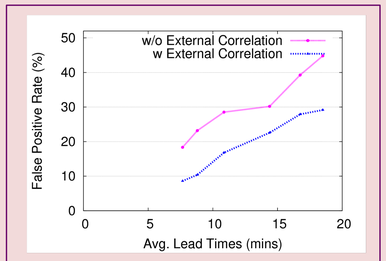


Figure 2

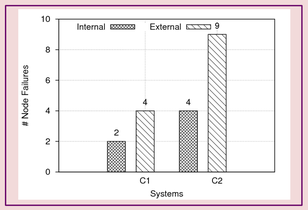
Observations:

- *Figure 1 shows that lead times of node failures increase by ~5 times compared to node failure analysis in isolation*
- *Figure 2 shows that with ~5 times increased lead times, the false positive (FP) rate do not rise with external correlations (they are lower than the FP rate with only node-specific events)*

Results (cont.)

Observations:

- *Internal Causes (console/message/consumer logs)*
  - *Application-triggered, do not have early external indicators*
  - *Lead time enhancements not possible*
- *External Causes (Controller/Environment/Event)*
  - *Fan speeds & voltage often operate below the min threshold*
  - *Temperature threshold violations common, but not the main culprit of failures*
  - *Link errors affect blades partially, transient faults exist, manifested failures may have unknown causes (intangible in logs, e.g., Solar Flares)*
  - *Lead time enhancements possible based on early indicators*
- *Live migrations or periodic checkpoint-restarts may not improve resilience when temperature and voltage conditions are not restored in the cabinets*



- *In systems C1 and C2, > 50% node failures have early indicators related to external causes, in a sample time-frame of 10 days*
- *Internal causes:*
  - *Job triggered resource exhaustion*
  - *Processor corruptions followed by kernel oops*

Conclusions

- Root cause diagnosis **enhances lead times to node failures by ~5 times**
- With external environmental correlations, **false positive rates are lower** w.r.t. the cases without any external correlation.
- Several failures have unknown root causes (intangible in logs):
  - Generic algorithm impractical, automation got the goal
  - Measurement-driven statistical analysis, insights to potential causes
- Holistic understanding of how failures happen enhance awareness of what actions to take for long-term system health.
- Results suggest that more than 20% of the sensor reading deviation messages do not lead to eventual failures.
- Further investigation
  - Pin–point conditions when typical software traps and hardware faults do not result in failures.
  - Analyze inaccuracy in failure predictions with system-wide environmental considerations.
  - Quantify inter-node correlations in the context of resource sharing and components influencing them (file system, interconnect).

References:

[1] A Das, F Mueller, C Siegel, and A Vishnu. 2018. Desh: Deep learning for system health prediction of lead times to failure in HPC. In HPDC, Tempe, AZ, USA.

[2] Z Zheng, L Yu, Z Lan, and T Jones. 2012. 3–Dimensional root cause diagnosis via co-analysis. In ICAC, San Jose, CA, USA.

[3] X Fu, R Ren, S. A. McKee, J Zhan, and N Sun. 2014. Digging deeper into cluster system logs for failure prediction and root cause diagnosis. In IEEE CLUSTER, Madrid, Spain.

[4] S Jha, J. M. Brandt, A. C. Gentile, Z Kalbarczyk, G. H. Bauer, J Enos, M. T. Showerman, L Kaplan, B Bode, A Greiner, A Bonnie, M Mason, R. K. Iyer, and W Kramer. 2017. Holistic Measurement-Driven System Assessment. In IEEE CLUSTER, Honolulu, HI, USA.

Acknowledgments: This work was supported in part by DOE subcontracts from Lawrence Berkeley National Lab and Lawrence Livermore National Lab, and NSF grants 1525609 and 0958311.