Holistic Root Cause Analysis of Node Failures in Production HPC

Extended Abstract

Anwesha Das, Frank Mueller
North Carolina State University
{adas4,fmuelle}@ncsu.edu

ABSTRACT

Production HPC clusters endure failures incurring computation and resource wastage. Despite the presence of various failure detection and prediction schemes, a comprehensive understanding of how nodes fail considering various components and layers of the system is required for sustained resilience. This work performs a holistic root cause diagnosis of node failures using a measurement-driven approach on contemporary system logs that can help vendors and system administrators support exascale resilience.

Our work shows that lead times can be increased by at least 5 times if external subsystem correlations are considered as opposed to considering the events of a specific node in isolation. Moreover, when detecting sensor measurement outliers and interconnect related failures, triggering automated recovery events can exacerbate the situation if recovery is unsuccessful.

1 INTRODUCTION

Recent research on log mining based failure characterization [6, 7], prediction [3, 11] and recovery [2] have revealed helpful insights to address failures in supercomputers. As researchers are designing energy efficient exascale nodes [14], current computing platforms require robust failure handlers to keep up with system scale and density. Proactive fault tolerant solutions [3, 11] when supported by root cause diagnosis can improve lead times and help in responding to both manifested or imminent failures effectively.

Our work is novel in that it considers system environment conditions along with inter-component dependencies to increase lead times to failures enhancing node failure prediction schemes.

2 BACKGROUND AND RELATED WORK

[8, 9] discuss cases where interconnect failures (lane/link) and overheating cause system-wide outages (SWOs), job failures and failures during recovery. [9] report early indicators of interconnect faults and SWOs due to overlapping interconnect and filesystem fault recovery events. Such observations affirm the need for a holistic study to enhance failure prediction schemes [3, 11]. [2, 5, 7] study spatial/temporal correlations of failures, derive their logical correlations, and propose dynamic checkpointing schemes (C/R) after detecting system degradation. Other root cause diagnosis techniques [4, 10, 12, 16] either point out failure location or perform causal dependency analysis without holistic considerations or are application-centric. In contrast, we incorporate subsystem dependencies and quantify increase in lead times to node failures.

3 PROBLEM AND MOTIVATION

The current state-of-the-art lacks in the following aspects:
1. The various layers (software [6], hardware [15], application [13]) of these large-scale systems are mostly studied independently without exploiting their correlations during root cause analysis [4].
2. Diverse components of the system affect each other (e.g., interconnect [9], GPU [15], DRAM [1]). Focusing on a specific component in isolation provides a local view, without having a global perspective. Over a period of time faults recur or unprecedented events happen which requires re-investigation, adding unnecessary overhead.
3. Once a failure manifests, corrective actions need to be enhanced. Prevalent solutions (lazy C/R, migration) may not always improve resilience. A deeper understanding of how failures happen can aid in choosing the appropriate action for long-term system health.
This work investigates root causes of node failures considering software, hardware and application malfunctioning across diverse components with recommendations for mitigation approaches.

4 APPROACH AND UNIQUENESS

We consider system-wide environmental logs and blade/cabinet characteristics along with the node-specific internal events during the *unhealthy* time frame. We move from node to blade to cabinet to understand fault conditions and derive early indicators of impending failures. The controller logs coupled with event router messages provide deviations (higher/lower than the normal range) in sensor measurements (e.g., fan speed, temperature) to warn about health problems. Encompassing such features with node-specific events aids in holistic root cause analysis and derivation of lead times as high as 15 minutes. Correlating the state of shared resources with events internal to nodes help decipher the cause of failures.

5 RESULTS AND CONTRIBUTION

Results suggest that lead times can increase by 10 to 15 minutes (e.g., 2 mins to 12 mins) considering external subsystem correlations as opposed to focusing on node-specific events only [3]. The false positive rate with external correlations is lower compared to the lead times to failures considering only node-specific events (e.g., 18.35% to 8.58%). Erroneous patterns do not always cause failures, they are usually coalesced with additional hardware problems. Processes consuming excessive resources cause network errors making nodes unreachable. Processor interrupts in a single node can hamper sensor readings of an entire blade causing the air velocity to be automatically reduced by the firmware in the cabinet. While not all node failures can be predicted with significantly increased lead times, many of them can be flagged ahead of time if external environment conditions are diagnosed. These findings hint at potential actions such as reducing the number of reboots/restarts.

6 CONCLUSION

Accurate holistic root cause diagnosis is an indispensable step toward failure mitigation in practice. Proactive fault tolerant solutions

with estimated lead times may serve as a short-term cure. Since the root causes are not fixed, the same failures may recur unless we clearly understand how node failures happen. Better awareness has the potential to enhance recovery approaches. Analyzing the trade-offs of potential actions (proactive/reactive) when a node failure is imminent can have long-term benefits in lowering the occurrence of future failures.

REFERENCES

[1] Leonardo Bautista-Gomez, Ferad Zyulkyarov, Osman Unsal, and Simon McIntosh-Smith. 2016. Unprotected computing: A large-scale study of dram raw error rate on a supercomputer. In *SC*. IEEE Press, 55.

[2] Leonardo Arturo Bautista-Gomez, Ana Gainaru, Swann Perarnau, Devesh Tiwari, Saurabh Gupta, Christian Engelmann, Franck Cappello, and Marc Snir. 2016. Reducing Waste in Extreme Scale Systems through Introspective Analysis. In *IPDPS*. 212–221.

[3] Anwesha Das, Frank Mueller, Charles Siegel, and Abhinav Vishnu. 2018. Desh: Deep learning for system health prediction of lead times to failure in HPC. In *HPDC*. 40–51.

[4] 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 *IEEE CLUSTER*. 103–112.

[5] Siavash Ghiasvand, Florina M. Ciorba, Ronny Tschüter, and Wolfgang E. Nagel. 2016. Lessons Learned from Spatial and Temporal Correlation of Node Failures in High Performance Computers. In *24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP*. 377–381.

[6] S. Gupta, T. Patel, C. Engelmann, and D. Tiwari. 2017. Failures in Large Scale Systems: Long-term Measurement, Analysis, and Implications. In *SC*.

[7] Saurabh Gupta, Devesh Tiwari, Christopher Jantzi, James H. Rogers, and Don Maxwell. 2015. Understanding and Exploiting Spatial Properties of System Failures on Extreme-Scale HPC Systems. In *DSN*. 37–44.

[8] Saurabh Jha, Jim M. Brandt, Ann C. Gentile, Zbigniew Kalbarczyk, Gregory H. Bauer, Jeremy Enos, Michael T. Showerman, Larry Kaplan, Brett Bode, Annette Greiner, Amanda Bonnie, Mike Mason, Ravishankar K. Iyer, and William Kramer. 2017. Holistic Measurement-Driven System Assessment. In *IEEE CLUSTER*. 797–800.

[9] Saurabh Jha, Valerio Formicola, Catello Di Martino, Mark Dalton, William T Kramer, Zbigniew Kalbarczyk, and Ravishankar K Iyer. 2017. Resiliency of HPC Interconnects: A Case Study of Interconnect Failures and Recovery in Blue Waters. *IEEE Transactions on Dependable and Secure Computing* (2017).

[10] Baris Kasikci, Benjamin Schubert, Cristiano Pereira, Gilles Pokam, and George Candea. 2015. Failure sketching: a technique for automated root cause diagnosis of in-production failures. In *SOSP*. 344–360.

[11] Jannis Klinkenberg, Christian Terboven, Stefan Lankes, and Matthias S. Müller. 2017. Data Mining-Based Analysis of HPC Center Operations. In *IEEE CLUSTER*. 766–773.

[12] Catello Di Martino, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer, Fabio Baccanico, Joseph Fullop, and William Kramer. 2014. Lessons Learned from the Analysis of System Failures at Petascale: The Case of Blue Waters. In *DSN*. 610–621.

[13] Catello Di Martino, William Kramer, Zbigniew Kalbarczyk, and Ravishankar K. Iyer. 2015. Measuring and Understanding Extreme-Scale Application Resilience: A Field Study of 5, 000, 000 HPC Application Runs. In *DSN*. 25–36.

[14] Alvise Rigo, Christian Pinto, Kevin Pouget, Daniel Raho, Denis Dutoit, Pierre-Yves Martinez, Chris Doran, Luca Benini, Iakovos Mavroidis, Manolis Marazakis, et al. 2017. Paving the way towards a highly energy-efficient and highly integrated compute node for the Exascale revolution: the ExaNoDe approach. In *Euromicro Conference on Digital System Design (DSD)*. IEEE, 486–493.

[15] Devesh Tiwari, Saurabh Gupta, James H. Rogers, Don Maxwell, Paolo Rech, Sudharshan S. Vazhkudai, Daniel A. G. de Oliveira, Dave Londo, Nathan De-Bardeleben, Philippe Olivier Alexandre Navaux, Luigi Carro, and Arthur S. Bland. 2015. Understanding GPU errors on large-scale HPC systems and the implications for system design and operation. In *HPCA*. 331–342.

[16] Ziming Zheng, Li Yu, Zhiling Lan, and Terry Jones. 2012. 3-Dimensional root cause diagnosis via co-analysis. In *ICAC*. 181–190.