Analysis of Node Failures in High Performance Computers Based on System Logs

[Extended Abstract]

Siavash Ghiasvand
Technische
Universität Dresden, ZIH
01069 Dresden, Germany
siavash.ghiasvand@tu-dresden.de

Florina M. Ciorba
University of Basel,
Department of Mathematics
and Computer Science
4001 Basel, Switzerland
florina.ciorba@unibas.ch

Ronny Tschüter
Technische
Universität Dresden, ZIH
01069 Dresden, Germany
ronny.tschueter@tu-dresden.de

Wolfgang E. Nagel
Technische
Universität Dresden, ZIH
01069 Dresden, Germany
wolfgang.nagel@tu-dresden.de

Failure rates in high performance computers rapidly increase due to the growth in system size and complexity. Hence, failures became the norm rather than the exception [5]. The efficiency of recovery mechanisms, e.g., checkpoint-restart, is dependent on the mean time between failures (MTBF). In the near future, the MTBF of HPC systems is expected to be too short, and that current failure recovery mechanisms will no longer be able to recover the systems from failures [1]. Early failure detection is a new class of failure recovery methods that can be beneficial for HPC systems with short MTBF. Detecting failures in their early stage can reduce their negative effects by preventing their propagation to other parts of the system [2].

Time correlated failures have been investigated in large-scale distributed systems [6]. The event logs examined, featured strong daily patterns and high auto-correlation. Event logs of heterogeneous servers have also been investigated and different forms of strong correlation structures were found including significant periodic behavior [4] .

A moving window was used to generate groups of spatially correlated failures from empirical data [3] . The authors showed that spatial failure correlation cannot be neglected in an accurate analysis of system downtimes.

The goal of this work is to share the knowledge gained via our observations with the community and to contribute to the foundation of failure detection techniques.

Delivering high performance is a key property of HPC systems. To prevent any performance penalty due to actively probing of nodes' status, we employ a passive monitoring approach. The native Linux system logs (referred to as syslogs) are the source of our monitoring information. The granularity of monitored components plays an important role in interpreting the system behavior. Herein we focus on the failures observed at the node level.

A *node outage* is defined as the case when no syslog entry from a particular node can be observed. Three causes of node outages are considered: (1) site-wide power outages, (2) planned maintenance, and (3) other reasons.

Node failure correlations can be made along three dimensions:

Temporal: the time gap between consecutive failures falls below a certain threshold,

Spatial: the failed nodes share a physical resource, and

Logical: the failed nodes share a logical resource.

Taurus[1] is an HPC system designed to handle highly parallel applications. The native Linux syslog daemon runs on all compute nodes. Regardless of the content of syslog entries, we consider each entry as a *heartbeat* of its respective generating node. Thus, one can monitor the health of the nodes and, as such, determine whether a node is dead (no heartbeat) or alive (has heartbeat).

To detect node failures, we first seek syslog entries which indicate a reboot of the node. Then we track back in the syslog entries to find the last available entry before the reboot line. Such an entry is considered the outage point. A *node outage* is defined as the last entry in the syslog entries of a node before a restart syslog entry appears. Although we know that the node might be still alive after its last syslog entry, since it is no longer responsive, we consider it as an outage.

For this study, we picked only those node outages which we believe were not caused by general power outages and scheduled maintenances, and, therefore, call them *node failures*.

This work constitutes a first step towards detection and correlation of node failures using the syslogs from our system.

[1] `https://doc.zih.tu-dresden.de/hpc-wiki/bin/view/Compendium/SystemTaurus`

Only based on a deeper understanding of the failure patterns and their correlations, effective failure recovery and prediction mechanisms can be devised. In general, based on our observations, on Taurus many failures are temporally correlated, spatially correlated, or both, depending on the time interval within our examined 8-month time period. The differences are mainly caused by the various usages of the system over time; this naturally leads to the need for logical correlation.

The logical correlation is not always easy to infer. In this study we learned that logical correlation can more easily be revealed by further examination of the spatially and/or temporally correlated failures. However this study is ongoing. To be able to detect early, diagnose, and correlate generic failures, further investigation and analysis is needed for a generic failure diagnosis and correlation methodology.

Upon completion of the study, such a generic failure diagnosis and correlation methodology could be used to detect and prevent failures in a shorter time and more efficiently than the nowadays techniques. Following this study, we will examine correlation patterns which could help detect failures in their early stage or even predict them.

The automation of the analysis approach is planned as future work. More accurate extraction of actual failures from all node outages is also necessary. Using the proposed analysis approach on the syslog entries of other HPC systems and comparing their results observed during this study can lead to more general insights about systems behavior and will help in early detection of failures.

1. REFERENCES

[1] F. Cappello, A. Geist, and W. Gropp. Toward Exascale Resilience: 2014 update. *Supercomputing Frontiers and Innovations*, 1(1):5–28, 2014.

[2] A. Gainaru, F. Cappello, M. Snir, and W. Kramer. Failure prediction for HPC systems and applications: Current situation and open issues. *International Journal of High Performance Computing Applications*, 27(3):273–282, July 2013.

[3] M. Gallet, N. Yigitbasi, B. Javadi, D. Kondo, A. Iosup, and D. Epema. A model for space-correlated failures in large-scale distributed systems. In P. DâĂŹAmbra, M. Guarracino, and D. Talia, editors, *Euro-Par 2010 - Parallel Processing*, volume 6271 of *Lecture Notes in Computer Science*, pages 88–100. Springer Berlin Heidelberg, 2010.

[4] R. Sahoo, M. Squillante, A. Sivasubramaniam, and Y. Zhang. Failure data analysis of a large-scale heterogeneous server environment. In *Dependable Systems and Networks, 2004 International Conference on*, pages 772–781, June 2004.

[5] M. Snir, R. W. Wisniewski, J. A. Abraham, S. V. Adve, S. Bagchi, P. Balaji, J. Belak, P. Bose, F. Cappello, B. Carlson, A. A. Chien, P. Coteus, N. A. Debardeleben, P. Diniz, C. Engelmann, M. Erez, S. Fazzari, A. Geist, R. Gupta, F. Johnson, S. Krishnamoorthy, S. Leyffer, D. Liberty, S. Mitra, T. S. Munson, R. Schreiber, J. Stearley, and E. V. Hensbergen. Addressing failures in exascale computing. *International Journal of High Performance Computing*, 2013.

[6] N. Yigitbasi, M. Gallet, D. Kondo, A. Iosup, and D. Epema. Analysis and modeling of time-correlated failures in large-scale distributed systems. In *Grid Computing (GRID), 2010 11th IEEE/ACM International Conference on*, pages 65–72, Oct 2010.