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

Siavash Ghiasvand [§], Florina M. Ciorba [★], Ronny Tschüter [§], and Wolfgang E. Nagel [§]

[§] Technische Universität Dresden, Germany [★] University of Basel, Switzerland

1. Motivation and approach

- The mean time between failures (MTBF) in high performance computers is expected to become too short. [1]
- Current failure recovery mechanisms will no longer be applicable. [2]
- Understanding the correlation between failures is key for early failure detection. [3]
- Failures can be correlated along three dimensions: **temporal**, **spatial,** and **logical**.
- Correlations along the temporal and spatial dimensions can be leveraged to infer correlations along the logical dimension.
- Logical correlations can facilitate early failure detection (Figure 1). [4]

- System log entries of Taurus* from 01-09-2014 to 30-04-2015 are investigated.
- The correlation of node failures in time and space as well as reason is studied.
- Temporal and spatial information is extracted using in house Python/Bash scripts.
- Information required for logical correlation is extracted from the SLURM accounting database (via the *sacct* command). [4]

* https://doc.zih.tu-dresden.de/hpc-wiki/bin/view/Compendium/SystemTaurus

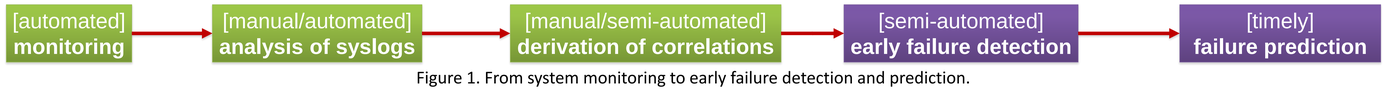

Figure 1. From system monitoring to early failure detection and prediction.

2. Initial analysis

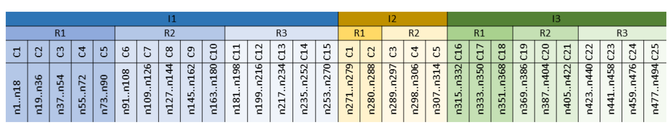

Figure 2. Topology of Taurus. I, R, and C denote island, rack, and chassis, respectively.

Figure 4. The maximum interval between observing two node outages on Taurus is 9 days.

- Figure 2 describes the topology of the HPC system under study.
- Figures 3-4 reveal the behavior of the system within the study period.

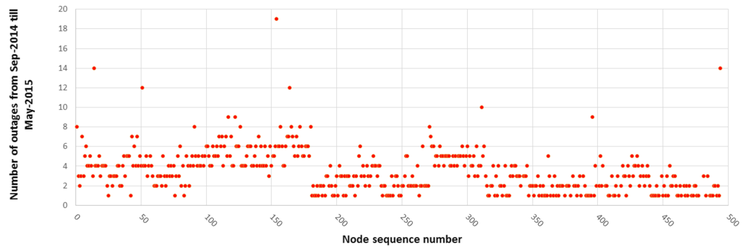

Figure 3. Total number of node outages (1669) for each of the 512 nodes between 01-09-2014 to 30-04-2015.

3. Logical correlation

- In Figures 5-7, the bottom row (reason) indicates correlations along the logical dimension, extracted directly from the system log.
- The logical correlation in these examples is also supported by and verified against the information obtained from the *sacct* command line tool.

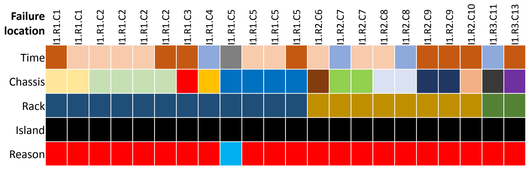

Figure 5. Node failures for 24 hours on 06-12-2014.

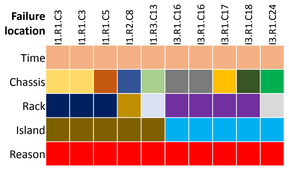

Figure 6. Node failures for 24 hours on 20-04-2015.

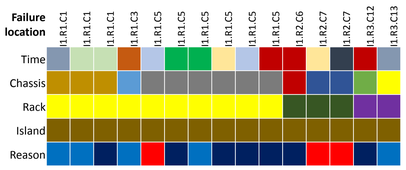

Figure 7. Node failures for 24 hours on 21-04-2015.

4. Outcome

- Early node failure detection is possible via understanding the logical correlation between observed failures.
- In general, logical correlation of failures directly from monitoring data is not feasible.
- In many cases, correlations can be made along the logical dimension based on the temporal and spatial failure correlations.
- Early failure detection is possible via
  - Directly extracting the logical correlations between failures.
  - Deriving the logical correlation from temporal and spatial correlations.

5. Conclusions and future work

- Early detection of failures is needed in rapidly growing high performance computers.
- Logical failures correlation is key for early failure detection.
- Although logical correlation is not straightforward to infer, temporal and spatial correlations are more easily detectable.
- Logical correlation can be derived from failures correlated in time and space.
- The “monitoring-to-failure detection” process (Figure 1) needs to be automated.
- Further investigation on Taurus2 (the latest HPC system at TU Dresden) is planned.
- The final goal is reliable prediction of certain types of failures.

References

[1] 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. Munson, R. Schreiber, J. Stearley, and E. V. Hensbergen, “Addressing Failures in Exascale Computing,” International Journal of High Performance Compututing, Apr. 2013.
[2] F. Cappello, A. Geist, and W. Gropp, “Toward Exascale Resilience: 2014 update,” Supercomputing Frontiers and Innovations, vol. 1, no. 1, pp. 5–28, 2014.
[3] 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, Appl., vol. 27, no. 3, pp. 273–282, Jul. 2013.
[4] S. Ghiasvand, F. M. Ciorba, R. Tschüter, W. E. Nagel , “Lessons learned from spatial and temporal correlation of node failures in high performance computers”, in International Conference on Parallel, Distributed and Network-Based Processing, Heraklion Creete, Greece, February 2016, *under review.*

Acknowledgment

This work is in part supported by the German Research Foundation (DFG) in the Cluster of Excellence “Center for Advancing Electronics Dresden” (cfaed) and in the Collaborative Research Center 912 “Highly Adaptive Energy-Efficient Computing”.

The authors also acknowledge Holger Mickler of Technische Universität Dresden for his support in collecting the monitoring information on the high performance computing systems at TU Dresden.