Toward Resilience in HPC: A Prototype to Analyze and Predict System Behavior

Siavash Ghiasvand§, Wolfgang E. Nagel§, and Florina M. Ciorba★

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

Motivation

- The failure rate of high performance computers increases rapidly.
- The mean time between failures (MTBF) is expected to become too short [1] and current failure recovery mechanisms will no longer be efficient. [2]
- Early failure detection is essential to prevent the destructive effects of failures. [3]
- Ideally failure prediction can be the solution for this challenge.
- By analyzing the system behavior we might be able to predict its behavior, which will enable us to predict failures or at least detect them in their early stage.

Collecting information

Our methodology is based on a cyclic workflow:

1. System monitoring
2. Analysis of monitoring data
3. Derivation of correlations
4. Early failure detection
5. Timely failure prediction

Monitoring system status

Information such as CPU load, power consumption, and network availability increases the failure detection accuracy.

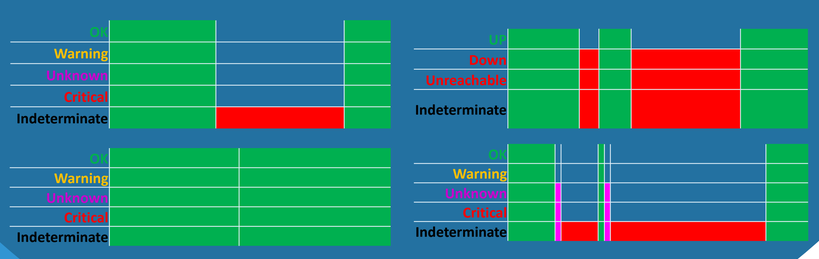

Figure 1. CPU load, power consumption, network availability and batch system availability of nodes

Correlating failures

As a first insight we recognized a meaningful pattern in node outages of Taurus. Nodes which were physically next to each other, had the same pattern of outages over the period of our study.

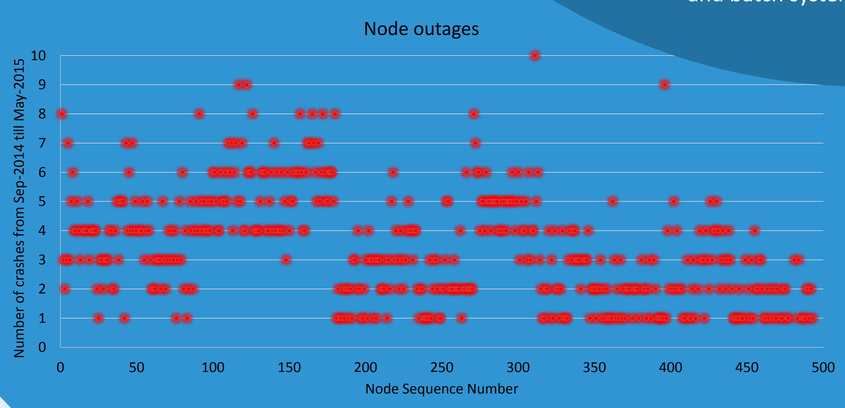

Figure 2. Total number of node outages for Taurus* between 01-09-2014 to 30-04-2015.

* https://doc.zih.tu-dresden.de/hpc-wiki/bin/view/Compendium/SystemTaurus#Phase_2

Patterns of anomalies

Figure 3. Patterns of anomaly, recognized as a sudden pick, sudden valley, or as a unusually low heartbeat rate

The golden interval

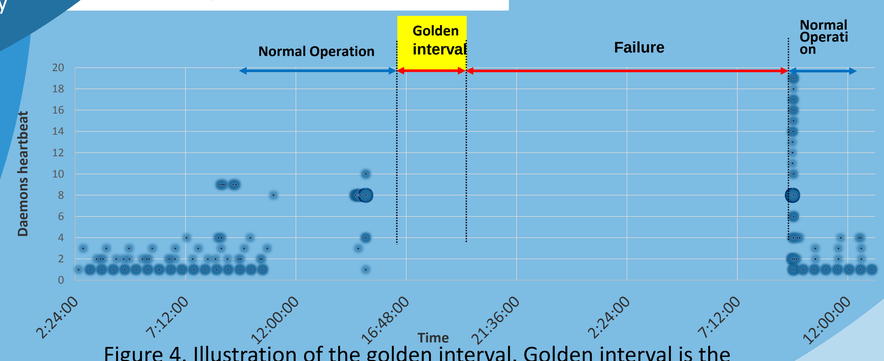

Figure 4. Illustration of the golden interval. Golden interval is the time window from detecting an anomaly until occurrence of a system failure

Temporal, Spatial and logical correlations

Early node failure detection can be achieved via two methods:

- identifying the reason behind the failures logical correlation) which is not always easy.
- identifying the temporal and/or spatial correlation of failures, which indirectly reveals the logical correlations.

Finally, we can state that all nodes which have the probability to have the same logical correlation will eventually encounter failures. [4,5]

(**1**) Failures with identical color in the time row occurred in less than a certain amount of time after a previous failure. (**2**) Failures with identical color in the chassis, rack, or island rows, occurred in the same chassis, rack, or island, respectively, as the failures preceding them in these locations. (**3**) Failures with identical color in the reason row, occurred due to the same reason as other failures on this day.

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

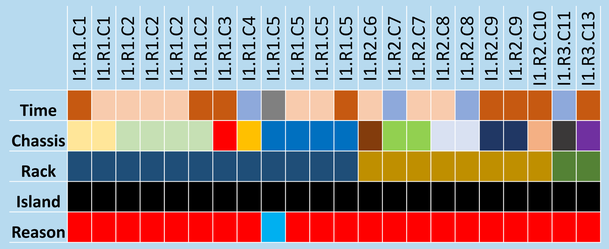

Figure 6. All node failures for 24 hours on 21-04-2015.

Conclusions and future work

With the fast growing failure rate, We believe that failure prediction is the solution for tomorrow's HPC systems. To make the failure prediction feasible we defined a 5-step workflow, out of which, the 3 first steps are already done and the 4th step is in progress.

As the next step, the "monitoring-to-failure detection" process needs to be automated. Further investigation on other HPC systems is also planned.

The final goal is introducing a prototype for analyzing and predicting certain types of failures.

References

[1] M. Snir, et al., "Addressing Failures in Exascale Computing," Int. J. High Performance Computing, 2013.

[2] F. Cappello, et al., "Toward Exascale Resilience: 2014 update," Supercomputing Front. Innov, 2014.

[3] A. Gainaru, et al., "Failure prediction for HPC systems and applications: Current situation and open issues," Int. J. High Perform. Comput. Appl. 2013.

[4] S. Ghiasvand et al., "Lessons learned from spatial and temporal correlation of node failures in high performance computers", in International Conference on Parallel, Distributed and Network-Based Processing, 2016.

[5] S. Ghiasvand et al., "Analysis of Node Failures in High Performance Computers Based on System Logs ", Poster presented at Supercomputing, 2015.

TECHNISCHE UNIVERSITÄT DRESDEN

ZIH Center for Information Services & High Performance Computing

cfaed

UNI BASEL