

## **Challenges addressed: Resilience through failure avoidance: New detectors of failure precursors and improved prediction workflow**

Franck Cappello\*<sup>o</sup>, Ana Gainaru<sup>o</sup>,  
\*INRIA, <sup>o</sup>UIUC  
cappello@illinois.edu

Failure avoidance relies on runtimes, OS, and system management environments to perform failure prediction and proactive actions. Making failure avoidance a credible solution for Exascale resilience supposes (i) to improve failure prediction drastically and (ii) to adapt the relevant software layers for failure avoidance. This white paper proposes an approach to address the most critical problem: improving failure prediction drastically.

The main foundations of failure avoidance are (i) precise failure prediction, (ii) large failure prediction coverage and (iii) enough time lags between correctly predicted failures and the real failures to trigger and perform proactive actions like migration, replication or checkpointing.

In the past few years, several key results have demonstrated that new anomaly/symptom detection and correlation analysis algorithms can provide precise on-line failure prediction in HPC systems. The time lag observed for the most efficient prediction approaches is consistent with the time taken by fast proactive actions, like checkpointing on local SSD or migration. However the proportion of failures that could be predicted over all observed failures is still low and stays below 50% even for the most advanced prediction approaches. The objective is then to improve the failure prediction coverage from 50% toward 80% or 90%, in HPC systems.

**Context:** A large fraction of the literature in failure prediction for HPC systems focuses on event analysis. HPC systems are producing events related to the state of their software and hardware components. Events of same types can be clustered into groups. Event correlation analysis allows establishing correlation graphs between events of a same group or/and of different groups. Correlation graphs essentially contain two categories of events: precursors and critical events. When a critical event is in a correlation graph, all previous events in the graph are called precursors (precursors potentially also include critical events). Recent advances in event clustering [6], anomaly detection [3], event correlation [2], correlation graph construction [4] and online detection of correlation graphs [5] lead to the conclusion that the reasons of the low failure prediction coverage are 1) the lack of precursor events (some failures have no identified precursors) and 2) the precision losses at each step of the failure prediction workflow.

**Objective:** We consider that failure prediction in HPC systems could be improved drastically by (i) developing and improving failure precursor detectors, especially for failures having no precursor with current analysis methods and (ii) quantifying and reducing the precision losses in every stage of the prediction workflow.

An example of existing precursors to a failure that is often seen on HPC systems is the case of hard memory errors that are announced with minutes beforehand by an unusual number of informational events related to correctable errors that are detected.

An example of event without known precursor is a NFS error that indicates unavailability of the network file system for a machine. With current analysis methods this failure is unpredictable. So the goal will be to analyze the reasons that lead to the NSF error and develop precursor detectors based on them.

**Maturity:** The overall failure prediction workflow [1], its limitations and needed improvements are reasonably well identified. There are many known approaches to develop new precursors detectors: function approximation, classifiers, system models and time series analysis. Quantifying precision loss in the prediction workflow is an open problem. Our methodology here is to start with running the prediction workflow from synthetic event flows that theoretically offer perfect prediction potential

according to the prediction workflow capabilities and to analyze the source of prediction precision losses in every stage of the prediction workflow independently. The next step will be to develop improvements for every stage.

**Uniqueness to Exascale:** Optimizations of the classic checkpoint/restart are already known for Petascale (application level checkpointing, multilevel-checkpoint restart, precise calculation of the optimal checkpoint interval, etc.). While failure prediction can also be used at Petascale, it becomes a key mechanism to reduce the time to solution and the energy spent at Exascale. By predicting accurately failures with enough time lag, proactive actions can reduce the execution slowdown due to failures drastically (the overhead of proactive migration or checkpointing is orders of magnitude lower than the overhead associated with the current checkpointing approaches, in presence of failures).

**Novelty:** Previous researches on on-line failure prediction have focused either (i) on developing symptom detectors for a single component (disk, memory, piece of software) and had to deal with a relatively small amount of events or (ii) on proposing algorithms dealing with large flows of existing events for specific stages of the prediction workflow [1]. Our approach is different since we need both to develop new precursor detectors and to deal with very large flow of events. We focus on (i) understanding and characterizing the types of failures in HPC system that show no precursors and developing new failure precursor detectors for them and (ii) improving the accuracy of the prediction workflow as a whole by quantifying the precision losses in every stage of the failure prediction workflow and developing improvements.

**Applicability:** The prediction methodology consists in using precursor detectors to predict the occurrence of critical events. Precursor detectors detect deviations of system parameters from normality. This requires (i) to establish relevant metrics, (ii) to qualify normality from these metrics, (iii) to quantify deviations and (iv) to define thresholds to qualify precursors events (a deviation exceeding a threshold would be considered as precursor events). Precursors and critical events are currently envisioned for failures, but with the relevant metrics, they might be used for power and performance.

**Effort:** Use of sensor to predict circuit failure is common in new scaled CMOS circuits [8]. We will focus our effort on developing new software precursor detectors. We anticipate the need of 1 man-year to discover, for a given system, failures that have no precursors. We consider that several man-years are needed to develop new failure precursor detectors for the most frequent failures, for that system. We also consider 1 man-year to quantify the losses of precision in every stage of the failure prediction workflow. Several man-years will be needed to reduce the precision losses in the failure prediction workflow.

## References:

- [1] Felix Salfner, Maren Lenk, and Miroslaw Malek. 2010. A survey of online failure prediction methods. *ACM Comput. Surv.* 42, 3, Article 10 (March 2010), 42 pages.
- [2] Ana Gainaru, Franck Cappello, Marc Snir, William Kramer, Fault prediction under the microscope: A closer look into HPC systems, *Proceedings of IEEE/ACM SC12*.
- [3] Ana Gainaru, Franck Cappello, William Kramer, Taming of the Shrew: Modeling the Normal and Faulty Behavior of Large-scale HPC Systems, *Proceedings of IEEE IPDPS 2012*.
- [4] Eric Heien, Derrick Kondo, Ana Gainaru, Dan LaPine, Bill Kramer, Franck Cappello, Modeling and Tolerating Heterogeneous Failures in Large Parallel Systems, *Procs. of IEEE/ACM SC11*.
- [5] Ana Gainaru, Franck Cappello, Joshi Fullop, Stefan Trausan-Matu, Bill Kramer, Adaptive Event Prediction Strategy with Dynamic Time Window for Large-Scale HPC Systems, *Proceedings of Managing Large-Scale Systems via the Analysis of System Logs and the Application of Machine Learning Techniques (SLAML) 2011*.
- [6] Ana Gainaru, Franck Cappello, Stefan Trausan-Matu, Bill Kramer, Event log mining tool for large scale HPC systems, *Proceedings of EuroPar conference 2011*.
- [7] Mridul Agarwal, Varsha Balakrishnan, Anshuman Bhuyan, Kyunglok Kim, Bipul C. Paul, Wenping Wang, Bo Yang, Yu Cao, Subhasish Mitra, Optimized Circuit Failure Prediction for Aging: Practicality and Promise. *IEEE International Test Conference, ITC 2008*.