# Online monitoring of HPC resources using decentralized clustering

Alejandro Pelaez
Rutgers University
alejo.pelaez@rutgers.edu

Manish Parashar
Rutgers University
parashar@rutgers.edu

James C. Browne
University of Texas at Austin
browne@cs.utexas.edu

Andres Quiroz
Xerox
Andres.Quiroz@xerox.com

Edward Chuah
University of Texas at Austin
edwardchuah@acm.org

## ABSTRACT

Large-scale clusters often fail due to either the failure of an individual component or a failure arising from a fault in the complex interactions among components. As a result, monitoring these complex systems and interactions, and detecting and predicting failures and other conditions based on the large amount of monitoring data have become critical for ensuring reliability. System event logs have become one of the primary sources of information for this purpose, and faults recorded in the logs have been shown to provide good indicators of failures, often minutes or even hours before the system failure actually occurs. Consequently, it is essential that such monitoring and analyses processes be done continually and in soft realtime if the faults that will lead to failures are to be identified or perhaps even predicted and averted. However, as system scales grow, centralized solutions can become cost-prohibitive due to the sheer amount of data movement as well as the computation (and energy) required to transport and process the entirety of the data.

In this poster, we address the problem of soft realtime monitoring and analysis of large-scale clusters by exploring a decentralized online clustering solution, or DOC, which has been developed for distributed system monitoring and resource provisioning. By partitioning the logs of monitored data along multiple semantic dimensions of interest, and using a locality preserving distribution of the data along with several caching strategies, we can achieve low data movement and local coherence and can decrease the overall data clustering time costs. This allows us to quickly identify local clusters of log entries (indicating patterns or âĂIJnormalâĂİ behavior) and aggregate them with only a few extra messages, and to perform in situ analysis on each node to identify outliers in the clustered data, which are hypothesized to lead to failure conditions.

Our initial results using logs from the Ranger cluster at TACC have shown our approach to perform quite well, scaling to hundreds of nodes, maintaining low total bandwidth and memory usage, and being able to run every 5 minutes with low performance overhead on each node and an overall impact of less than 1% on the execution time of popular benchmarks such as LINPACK or NPB. However, enabling effective soft realtime identification (or prediction) of system failures will require further effort in decreasing the still significant number of false positives and a very small number of false negatives that arise using our current algorithms to predict failures. We are currently investigating several hybrid approaches to lower these false positives, where we use centralized and more powerful techniques, including pattern identification and correlation analysis, but only applied to the subset of anomalies previously found in order to keep the performance impact low.