1 ARTIFACT DESCRIPTION APPENDIX: [HOLISTIC ROOT CAUSE ANALYSIS OF NODE FAILURES IN PRODUCTION HPC]

1.1 Abstract

A brief description of the measurement-driven approach is provided here to illustrate the major steps performed in the root cause diagnosis of node failures.

1.2 Description

The artifact description enumerates the basic requirements for conducting the empirical study. We mention which components of the logs where correlated to decipher node failure causes and derive lead times.

1.2.1 Check-list (artifact meta information).

- **Algorithm: Time Correlation, Failure Pattern Matching**
- **Program: Bash scripts, Python**
- **Compilation: None**
- **Transformations: Forming correlations of Cabinet, Blade, Node Identifiers with various system events**
- **Binary: None**
- **Data set: Real production logs**
- **Run-time environment: Linux 4.10.13-1.el7.elrepo.x86 64**
- **Hardware: Intel processors**
- **Run-time state: Unused compute nodes**
- **Execution: Bash Shell**
- **Output: Cause and effect between various components and layers of HPC subsystems**
- **Experiment workflow: Consideration of spatial and temporal locality, Analysis performed from cabinet to blade to node**
- **Experiment customization: Certain time frames are more unhealthy, Distribution of faults and failures vary accordingly**
- **Publicly available?: No**

1.2.2 How software can be obtained (if available). Subject to eventual publication.

1.2.3 Hardware dependencies. Any Intel platform

1.2.4 Software dependencies. None

1.2.5 Datasets. Log archives from well used contemporary HPC clusters.

1.3 Installation

Not Applicable. The required logs need to be cleansed and segregated into different components (environment, console, controller etc.) for root cause diagnosis.

1.4 Experiment workflow

The major steps are as follows:

(1) We analyze the controller logs consisting of events pertaining to the blade and cabinet controllers. These entail incidents pertaining to sensor measurement and heartbeat faults of nodes, blades and cabinet controllers. Some prior indications are available through alerts caused by environmental conditions or hardware faults.
(2) The event router logs along with SMW (System Management Workstation, a centralized administrator's console for Cray systems) messages provide additional information of system faults including location and timing. These help us formulate inter- and intra-node dependencies.
(3) The internal logs of compute nodes consisting of console logs are then correlated with the analysis of steps 1 and 2 to derive lead times to failures. The lead times obtained from compute node specific events alone are the base lead times. Subsystem correlations considering environmental logs in steps 1 and 2 increase the lead times of certain failures.

1.5 Evaluation and expected result

During the unhealthy time frames, few timestamps will overlap between controller, event and console logs since logs are routed to the SMW from different locations. The automated calculation of lead times are obtained by deducing the earliest indication of system malfunctioning pattern to the event after which the node became unresponsive. The chain of events considering node-specific events in isolation provides a certain lead time, say 2 minutes. For certain failures, lead time increases to 10 minutes when steps 1 and 2 (from Section 1.4) are taken into account. Such cases can be cross-validated from the log data.

1.6 Experiment customization

Based on the target systems and format of logs, time frames must be chosen carefully to capture the events when unintended failures have happened. Based on the location granularity (chassis/cabinet) and their capacity, frequency and distribution of faults vary.

1.7 Notes

Several anomalous faults do not always lead to failures. There are additional characteristics of environment and hardware problems which eventually lead to failures. Hence, holistic analysis strengthens lead time confidence in proactive fault tolerance.