Contribution ID: **40** Type: **Oral Presentation**

A Blueprint of Log Based Monitoring and Diagnosing Framework in Large Distributed Environments

Wednesday, 3 April 2019 14:20 (20 minutes)

Distributed systems have grown larger and larger since this concept appears, and they soon evolve to environments that contain heterogeneous components playing different roles, e.g. data centers and computing units. From security point of view, it is a difficult task to get an idea of how such large environment works or if any undesired matters happened. Logs, produced by devices, sub-systems and running processes, are a very important source to help system maintainers to get relative security knowledge. But there are too many logs and too many kinds of logs to deal with, which makes manual checking impossible.

CNGrid is a good example of a large distributed environment. It is composed of 19 HPC clusters contributed by many research institutes and universities throughout China. Computer Network Information Center of Chinese Academy of Sciences is the operation and management center of CNGrid, and is responsible for keeping the environment running smoothly and efficiently. The maintenance team has been working for years on using logs generated in CNGrid to help us analyzing system behaviors and addressing sources of failures.

In this work we will share some of our experiences by representing a general framework that monitors events, analyzes hidden information and diagnoses the healthy state for large distributed computing environments bases on logs. There are 4 major steps in this framework: 1) identifying necessary logs that are produced by key modules in the environment; 2) classifying these logs either using variables or non-variable contents to obtain a set of log types; 3) performing analyses from various angels such as user behaviors, type associations, etc., using data mining and machine learning techniques with the help of the obtained set of log types; 4) gathering results from previous analyses and find some metrics that can numerically represents the vulnerability level of the environment, and produce a diagnosis report.

We will use CNGrid as an example to demonstrate these 4 steps. First, we pick up OS system logs and SCE middleware logs as the target to be analyzed, according to our demands of responding to system failures, defending on malicious attacks and providing user services. Second, log pattern extraction algorithms are used to classify types for system logs, and SCE logs are grouped by usernames. We organize these log types as the log library and implement interfaces to access it. Third, by applying the log library, a number of log analyzing applications are developed to perform analyses, including log flow detection, fault prediction bases on log type associations and user behavior analysis. Finally, results of these analyses would be turned into some measurable factors, so that they can be combined to reach an overall conclusion. We plan to weight these factors so that the produced diagnosis report is more adequate to describe the vulnerability level of the environment.

Although the framework we initially designed was for the maintenance for CNGrid, it is believed that the process is adaptable to other distributed computing environments.

Primary author: Dr ZHAO, Yining (Computer Network Information Center, Chinese Academy of Sciences)

Co-authors: Mr XIAO, Haili (Supercomputing Center, Chinese Academy of Sciences); Mr WANG, Xiaodong (Computer Network Information Center, Chinese Academy of Sciences); Prof. CHI, Xuebin (Computer Network Information Center, Chinese Academy of Sciences)

Presenter: Dr ZHAO, Yining (Computer Network Information Center, Chinese Academy of Sciences)

Session Classification: Networking, Security, Infrastructure & Operations

Track Classification: Network, Security, Infrastructure & Operations