Name: Prashanth Mallyampatti
Unity ID: pmallya
Student ID: 200250501

CSC724-Advanced Distributed Systems
Paper Review

Capturing, Indexing, Clustering, and Retrieving System History

Summary:

This paper addresses the problems involved in identifying system problems, identification and retrieval of similar problems using the system's failure history, and how to use, effectively apply the captured data & state of the machine for detecting performance anomalies.

This paper provides a method which extracts the raw system metrics, and machine state and clusters them into indexable groups based on the instances previously recorded. These indexable groups are termed as signatures which accounts for diagnosing system problems, reference for future categorization of problems. Here SLO violation, compliance and low-level system metrics are measured in time intervals and Tree-Augmented Naïve Bayes models (TAN) are used to classify them into various signatures. These signatures are clustered whose centroids are syndrome of that problem. The entropy of the clusters is calculated to know the purity of the clusters. Previous instances which are similar and closest to the current signature are retrieved to know the diagnoses and repairs applied. Trace collection was collected from a wide variety of sources – controlled laboratory, and globally distributed environment.

The proposed method was tested on both experimental testbed, global environments, and synthetic workloads. The quality of signature clustering and retrieval wherein only 10% of them weren't able to separate SLO violations and compliance. A precision of 92% of signature retrieval was measured. Second identifying the recurrent problems had far fewer errors in annotations and recurring problems were easily identified the second time they occur. Successful cross site usage of signature has also been showed. All these results indicate that the proposed method addresses system problems in a more sophisticated way and can be applied to wide variety of environments. Since system state and metrics are recorded continuously the model adapts and improves its clustering efficiency.

Strong Points:

1) Considering the state of the system rather than only metrics is good approach as system metrics can be similar for various states and thus address different violation (or compliance).
2) The paper presents related work which gives an insight of other approaches which can be compared with the proposed method.
3) The proposed method being able to identify similar or recurring problems, even when no problem annotations are available and also giving an insight for diagnoses of future problems which are similar to the recorded ones is a plus.

Weak Points:

1) The overheads of gathering and clustering data have not been discussed in detail. These clustering might use large amounts of system resources for large distributed systems.
2) Many organizations require immediate root cause for new problems arising, which cannot be obtained using this method.
3) The assumption of continuous measure of state and data imposes problems such as a state responsible for a system crash may not be measured and thus cannot be foreseen if occurred again.