Online Approach to Performance Fault Localization for Cloud and Datacenter Services

Jawwad Ahmed[1], Andreas Johnsson[1], Farnaz Moradi[1], Rafael Pasquini[3,4], Christofer Flinta[1], Rolf Stadler[2,4]
[1] Ericsson Research, Sweden, Email:{jawwad.ahmed, andreas.a.johnsson, farnaz.moradi, christofer.flinta}@ericsson.com
[2] ACCESS Linnaeus Center, KTH Royal Institute of Technology, Sweden, Email: stadler@kth.se
[3]Faculty of Computing (FACOM/UFU), Uberlândia – MG, Brazil, Email: rafael.pasquini@ufu.br
[4] Swedish Institute of Computer Science (SICS), Sweden

Abstract—Automated detection and diagnosis of the performance faults in cloud and datacenter environments is a crucial task to maintain smooth operation of different services and minimize downtime. We demonstrate an effective machine learning approach based on detecting metric correlation stability violations (CSV) for automated localization of performance faults for datacenter services running under dynamic load conditions.

I. INTRODUCTION

Next generation cloud services will demand greater flexibility and higher service quality from cloud systems while expecting more reliability. Real-time service assurance will become an integral part in transforming the general and flexible cloud into a robust and highly available cloud that can ensure low latency and agreed service quality for its customers.

A service assurance system must be able to detect problems that may violate the agreed service quality in a timely manner. This is a complex task already in legacy systems and will become even more challenging in the cloud, particularly for dynamic services that are distributed on multiple machines in the same datacenter or even across multiple geographically dispersed datacenters. One promising approach to service assurance is based on machine learning, where the service quality and behavior is learned from observations of the system. The ambition is to do real-time predictions of the service quality and in case of SLA (service-level-agreement) violations do the cause inference so that mitigation actions can be taken to remedy the detected faults. It is critical to restore the violated service as soon as possible to minimize the impact of potential penalties from SLA violations.

Machine learning has been used in the past for tasks like predicting user application quality-of-service (QoS) parameters, SLAs, and different type of anomalies for complex datacenter environments [1-3]. Predicting the SLA fulfillment level is an important tool for delivering high-quality services. But even more important is to have an effective real-time fault localization system so that the provider can take timely actions in contrast to traditional time consuming, costly and error-prone customer support services.

The core of our fault localization approach is based on a key observation that although workload changes may impact the performance of the application in a given time-period, the pairwise correlation values across most system metrics collected at the servers should remain stable. Correlation values change noticeably only if there is indeed a fault in the system (in that time-period) which we here refer to as correlation stability violations (CSV). This finding has already been reported in previous work [4]. However, the Fault Localizer (FL) we have implemented also takes into account the number of CSV occurrences and their magnitude to establish metric fault scores (MFS) to improve the accuracy of diagnosis.

We demonstrate the effectiveness of our approach for a video-on-demand (VoD) service use case. We show that the approach can be used to detect different performance faults (see Figure 1) under non-steady load conditions even when multiple faults occur concurrently in the system. Moreover, the unsupervised nature means that our approach is online and does not need any explicit training phase.

II. SYSTEM ARCHITECTURE AND TESTBED

Figure 2 shows the overall system architecture where the Cloud/DC monitoring system collects the data from the infrastructure both the system data as well as operational service data e.g., using system activity report (SAR). Data is then cleansed, normalized and smoothed at the 'Data Pre-processor' module also including the feature engineering. 'Fault Detector' (FD) sub-system uses predictive analytics techniques to detect any potential faults in the system [1-3]. In case a fault is detected then an alarm is generated towards the 'Fault Localizer' (FL) module which is the focus of this demonstration. Alarm filtering mechanism can optionally be turned-on to further reduce the false alarm rate. FL uses multi-variate statistical correlation methodology to pin-point faulty or suspicious metrics in the system. The approach is online and can adapt to system state changes over time. Decoupling of the detection and localization components provides a modular system with scalability and extensibility. The 'Fault Classifier' module is responsible for classifying the detected problem at a higher level, based on the details of the faulty metric(s) pointed-out by the FL. The last two modules are related to the system actuation to remedy the faults in the system by calculating parameters for a mitigation action.

Figure 3 shows the testbed system that we use for data collection [1]. It consists of a server connected to a client machine via a network. 'Load Generator' machine is used to execute different load profiles in the system. The client accesses a VoD service which resides on the server and a binary SLA is defined for the client-side service metrics. Figure 3 also shows the data processing pipeline. 'Big-data Path' uses ELK stack (i.e., Elasticsearch, Logstash, and Kibana) [5] to collect, process, transfer, and store the large amounts of data. Alternatively, low volumes of data is directly stored in the offline DB for analysis in 'Small-data Path'. For the implementation of the ML algorithms Spark, R or Python

978-3-901882-89-0 @2017 IFIP

libraries can be used. Finally, performance and diagnostic stats are rendered using a real-time dashboard application (i.e., Kibana).

III. Demonstration

The demo shows the fault diagnostic capabilities of our approach using the data collected in a testbed environment for a VoD service where service SLA violations are observed by the clients due to the infestation of system performance faults under dynamic load conditions. Dashboard displays the accuracy of our approach over-time on localizing concurrent occurrence of faults deduced from their degree of suspiciousness.

References

[1] R. Yanggratoke, et al., "Predicting Real-Time Service-Level Metrics from Device Statistics", In IFIP/IEEE IM, Canada 2015.

[2] J. Ahmed, et al. "Predicting SLA conformance for cluster-based services using distributed analytics." NOMS 2016. IEEE, 2016.

[3] R. Yanggratoke, et al., "Predicting service metrics for cluster-based services using real-time analytics", In CNSM, 2015.

[4] S. Bikash, et al., "Cloudpd: Problem determination and diagnosis in shared dynamic clouds." In IEEE DSN 2013.

[5] ELK stack, URL: https://www.elastic.co/products.

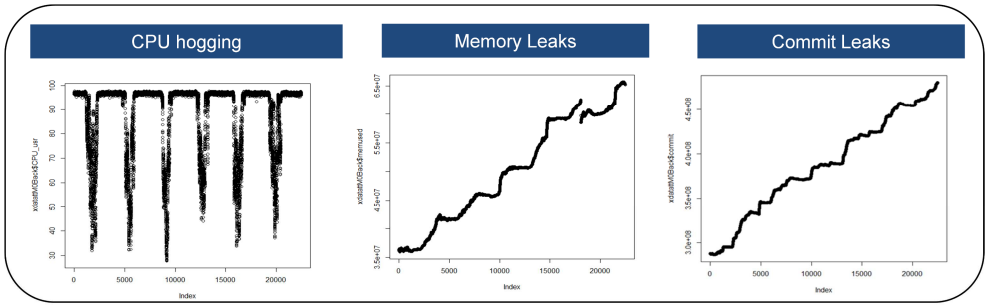

Fig. 1. Different type of performance faults manifesting themselves in the physical host machines(s)

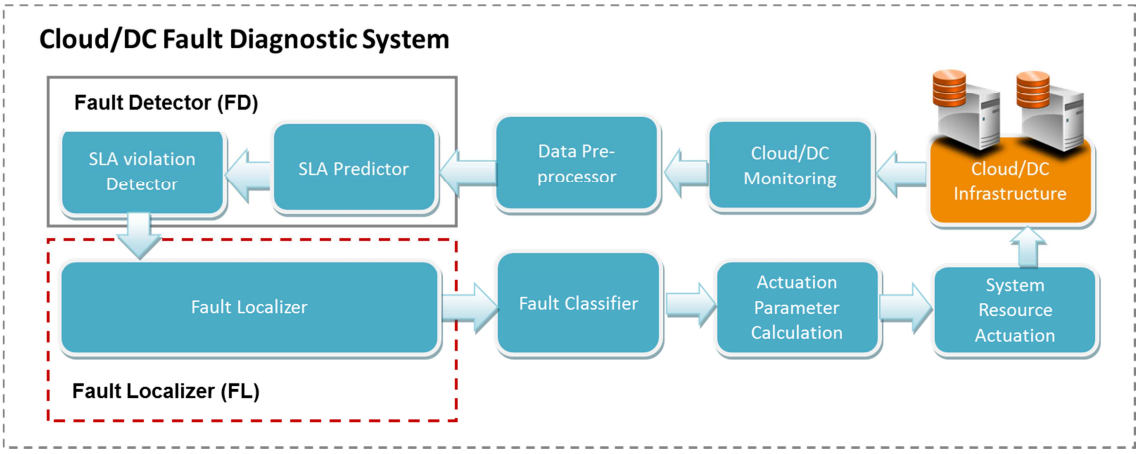

Fig. 2. A high-level architecture of Cloud/DC Fault Diagnostic System

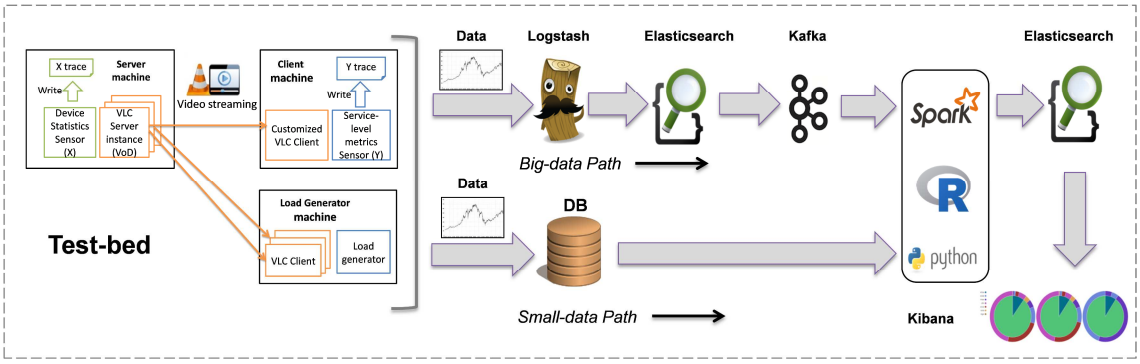

Fig. 3. Testbed and Data Processing Pipeline for the Cloud/DC Fault Diagnostic System