Performance Monitoring of Clouds for Network Centric Applications

YOUNGMIN KIM†1 HEEJU JOO†1 JAECHUL UM†1
HONGUK WOO†2 EUNHO HEO†2 CHAN-GUN LEE†1

Abstract: Recently much interest in various applications on cloud computing is getting increased rapidly. We argue that there is short of enough studies regarding effective tools and methods for evaluating performance of cloud services. This is even worse for the network centric applications on the clouds where the network is known to be a typical performance bottleneck and has significant impacts on the quality of service. In this paper, we propose a monitoring framework for gathering and analyzing the performance of the cloud for network centric applications. The framework measures various aspects of cloud system performance including network metrics regarding both of intra-cloud and WAN as well as the metrics for CPU, storage, and other resources. We present our initial design and key features of the framework.

Keywords: cloud computing, performance, cloud service, network centric, monitoring framework

1. Introduction

Recently, many companies, such as Amazon, Google and Microsoft, have launched their cloud infrastructures and services. The cloud services allow users to construct their own flexible applications suited for the business needs and to deploy the applications on-demand. However, due to the lack of objective performance analysis and experiences of cloud computing, it is difficult for the users to determine which cloud provider is the best to implement their services. In addition, it is hard for them to predict the performance of their applications due to highly dynamic natures of the cloud services. For this reason, the framework that can present a comprehensive performance analysis of cloud system and help them to diagnose their application behaviors is strongly required. In addition, monitoring service at the service-level is needed to improve and optimize the user applications.

According to development of smart devices, cloud-based multipoint communication services are gaining tremendous popularities. They enable the users to communicate anytime and anywhere via the service. Since the quality of service (QoS) and application performance are critical to the application users, we need effective metrics for performance analysis and monitoring of the application to guarantee the QoS and application performance. Unfortunately, the measurements of cloud system have been performed in a limited range of metrics such as CPU scheduling pattern among the users and bandwidth between the computing resources and the storages until now. The results from these studies are not enough for our purpose described in the above.

In this paper, we focus on multicast network based services and propose a framework that provides comprehensive performance analysis and service-level monitoring of key elements of cloud computing. Our framework is designed to work with public cloud infrastructure as a service (IaaS). Currently we are using Amazon Web Services (AWS), hence various AWS components such as EC2 and S3 will be mentioned for illustrating main features of our system throughout the paper. However, we expect that other IaaS can be also integrated with minor extensions.

Our framework collects various metrics regarding network performance such as intra-cloud network and Wide Area Network (WAN) as well as other resources like CPU, storage, and etc. We expect that this framework is useful to those developing the cloud applications in improving the quality and performance of the services and also helpful for those planning to make business decisions regarding cloud applications.

In order to provide various geographical measurement points for WAN performance, we incorporate with DipZoom which has more than 400 world-wide proxies.

The rest of this paper is organized as follows. Section 2 gives related studies about evaluation of cloud system and multicast services. Section 3 defines the framework proposed in our approach. In this section, we introduce our approach to evaluate various resources including intra-cloud network, WAN, and CPU and the mechanism to recommend the best server in multicast network based services. Section 4 proposes the advantages of our monitoring framework and concludes the paper.

2. Related Work

2.1 Performance Evaluation of Clouds

Recently a few of efforts on the evaluation of cloud system were performed from various viewpoints. Simson et al. [1] evaluated the performance of Amazon S3 with different sizes of objects. They transmitted probes continuously in order to measure the end-to-end performance between EC2 and S3. Jörg et al. [2] tried to analyze various aspects of EC2 performances such as instance startup, CPU, memory speed, disk, and network bandwidth for S3. They presented analysis results for different instance types and regions of EC2. Guohui et al. [3] investigated the patterns of process sharing among the EC2 virtualized machines, packet delays, and throughputs and packet loss rates of TCP/UDP. They measured the progress of packet delays among the EC2 instances and packet loss rates at data center. Cloudcmp[4], a tool for measuring the Amazon cloud system, has been developed. Its main features include benchmark performance, network measurement, and EC2 evaluation.

In spite of numerous previous efforts toward the cloud performance evaluation, we argue that more intensive studies for a systematic monitoring framework for gathering and

†1 Dept. of Computer Science and Engineering, Chung-Ang University, Seoul, Korea. remnant1120, soulmateof88, steveum0105@gmail.com; cglee@cau.ac.kr
†2 Samsung Electronics. eunho.heo, honguk.woo@samsung.com

analyzing the performance of the cloud are needed to support network centric applications with critical constraints on various resources and service quality.

2.2 Multipoint Communication Services on Clouds

Among the various cloud applications, we are mostly interested in a multi-user video conferencing service. There exist several studies on development of a model for workloads used in enterprise IP telephony applications [5] and audio/video conferencing service based on commercial distributed system [6]. In addition, a few of research groups focused on distributed network based real-time streaming service [7][8][9]. In this paper, we consider a multi-user video conferencing application as a typical example of multipoint communication services. We designed our monitoring framework for the service and illustrated its features and architecture in Section 3.

3. Proposal of Network Monitoring Framework

3.1 Overview of Framework

This section presents main features of our monitoring framework and its design in detail. In the following, we describe the main features. Firstly, it measures different aspects of network performance of intra-cloud network and WAN. The users are informed of various metrics such as latency, throughput, jitter, and packet loss rates. Secondly, it evaluates elastic compute cluster. For this we measure CPU usage of each cloud node. This feature is particularly important for the cloud nodes shared by other users. For example, the Amazon's Elastic Compute (EC) nodes with small instances are known to be shared by other cloud users [1]. Lastly, it recommends the best server based on the analysis results from the monitoring so that the cloud application users can be served by the most appropriate server dynamically. The overview of our framework is shown in Figure 1.

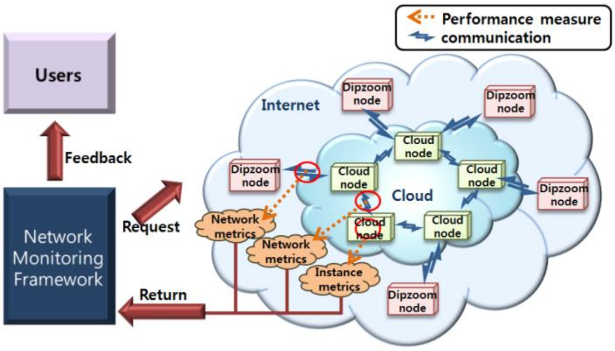

Figure 1. Overview of monitoring framework

3.2 Measuring Performance of Intra-Cloud Network

The intra-cloud network enables the communication between the nodes in cloud as illustrated in Figure 2. The measurement of intra-cloud network is an important factor to evaluate the performance of cloud network.

The inner nodes typically work as servers providing the primary services of the applications. In addition, those servers are assigned to customers using the shared services offered by the cloud. Therefore, the intra-cloud network determines how fast the server nodes comprising of the application cores can interact with each other. In addition, the inner nodes in the cloud may be assigned to a particular job and operate in a distributed fashion. In such a case, a group of nodes have to closely work together and communicate with each other. In the end, the intra-cloud network is significantly related to the whole performance of distributed applications.

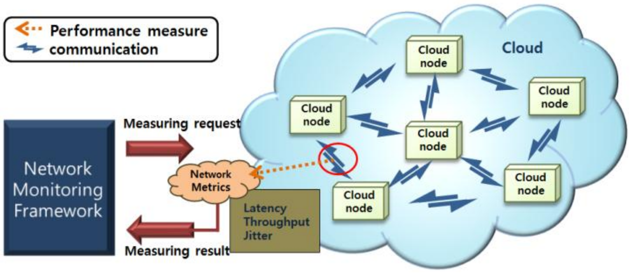

Figure 2. Measurement of intra-cloud network

We present the measurement metrics for the intra-cloud network in the following. Firstly, we measure the TCP/UDP throughput between the inner nodes of the cloud. It has impacts on data transfer performance and the way to handle congestion events. Secondly, we measure the network latency between the inner nodes. It determines the end-to-end response time, which is important factor for many applications. Lastly, we use packet loss rates between the inner nodes. It impacts the efficiency of data transfer and error control.

3.3 Measuring Performance of Wide Area Network

The WAN provides network services between arbitrary client nodes and cloud computing nodes e.g. Amazon ECs in the cloud. The metrics for the measurement of WAN are identical to those of intra-cloud network as shown in Figure 3. We measure network performance between an arbitrary node on the Internet and one of cloud servers. One of the most difficult problems for measuring the network performance of WAN is that it requires the installation of measuring tools in advance. This limitation can significantly hamper the applicability of monitoring in practice because we cannot simply install the needed tools on an arbitrary node without proper permissions. Even in the situations where we are granted the necessary permissions, it may require complicated and manual installations of the tools and the solution would not scale. In order to overcome these technical difficulties we have decided to utilize DipZoom [10] which is an Internet measurement platform. It features about 400 measurement points around the world and provides Java APIs with the users. Hence, custom benchmark programs using the APIs can be developed and deployed. In addition, DipZoom is free for the academic research purpose. Recent research utilizing DipZoom for network measurement of peer-to-peer systems introduced in [11]. The work showed the reliability and applicability of DipZoom as an WAN performance measurement tool. In addition, the comparison with King[12] and GNP[13] from the perspective of measuring accurate latencies between arbitrary network nodes was performed.

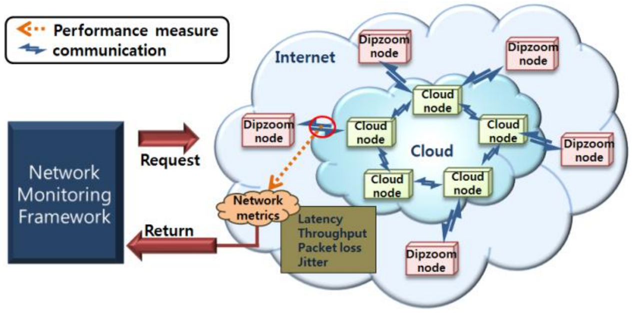

Figure 3. Measurement of WAN

3.4 Measuring Performance of CPU, Storage, and Other Resources

The elastic compute cluster is an on-demand computational service of the cloud. As discussed in the previous section, sometimes an Amazon's EC can be shared by multiple cloud customers. Hence it is needed to monitor the performance of EC in run-time to guarantee the overall performance of the service. To evaluate this, we measure benchmark finishing time, cost, and scaling latency of the EC. The benchmark finishing time is duration time to finish a specific benchmark task by a cloud node. The cost is monetary cost until benchmark tasks are completed. This cost varies on different commercial cloud services. The scaling latency is the delay time to launch a new instance of EC on the cloud. For storage performance, we measure latency and throughput for accessing the stored data. The access can be made either in intra or outside of the cloud. The performance metrics of other resources are collected from the default monitoring service provided by the cloud infrastructure and various logs of the system.

3.5 Recommending Optimal Server

The most appropriate server is recommended based on the analysis of the performance measurement. Table 1 presents various metrics utilized for the multicast network based services on previous research. In general packet loss rate, latency, jitter, and throughput are the main factors of network centric applications. In addition, we can find that cloud based E-learning service considered CPU and memory as well.

Table 1. Metrics for multicast network based services.

Previous studies	Metrics
Cloud based virtual server network for enterprise IP telephone service[5]	Packet loss rate Latency
A distributed system based on audio/video conference service[6]	Latency, Jitter Throughput
Cable network based on IPTV streaming service [14]	Throughput, Jitter
Hybrid P2P and streaming service based on Cloud [7]	Packet loss rate Throughput, Jitter
Cloud based E-leaning service[9]	CPU, I/O, memory

Our monitoring framework suggests a combination of metrics for server recommendation. Firstly, the network latency determines the response time on near real-time cloud services. In case the network latency exceeds the time limit of the application then it will lead to deterioration of the quality service (QoS) especially for the multipoint communication services. Hence, the network latency is a primary metric for our targeting applications. Secondly, the packet loss rate also impacts on QoS significantly. We note that moderate degree of packet loss may be tolerable in our application; however, exceeding the limit may render the multipoint communication services unusable at all. Other parameters affecting the recommendation include server availability, throughput, and

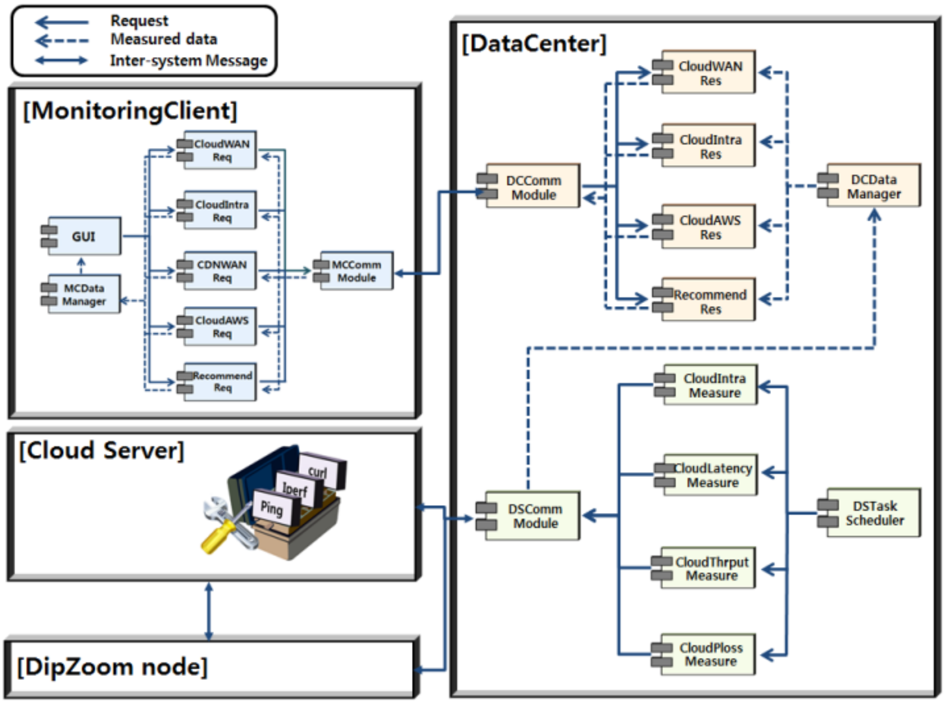

Figure 4. Module diagram of monitoring framework

jitter, variation of latency. Hence, our framework recommend the best server based on the combinations of the above various factors. We are looking forward to deriving the proper weight for each metric and the composition of them by the performance modeling and experiments.

3.6 Module Organization

The module organization of the proposed monitoring framework is shown in Figure 4. Our framework is composed of the subsystems MonitoringClient, DataCenter and CloudServer. The features of each subsystem are shown in the following in detail.

MonitoringClient provides the functions of GUI environment, selection of optimum servers suited for a service, and request for outcome of diverse network measurements to users. These results are shown to users through the visualization module and the notification service. The modules requesting performance measurements include CloudIntraReq, CloudWANReq, and CloudAWSReq. DataCenter requests the measurement for intra-cloud, WAN, and cloud servers periodically and stores the returned outcomes on database. A cloud server indicates the target of measurement and each server is equipped with performance measurement toolsets. DataCenter is composed of a component delivering data from database to MonitoringClient and a group of components measuring performance of intra-cloud network, WAN, and cloud servers. Measurement modules are called upon by daemon DSTaskScheduler module periodically. The results of measurement are saved in database by DCDataManager module.

4. Conclusion

In this paper, we proposed a monitoring framework for performance analysis of network centric applications on cloud services. . The performance of the monitored components is reported in the form of various metrics regarding elastic compute clusters, Intra-cloud network, and WAN. We are looking forward to completing the implementation of the proposed framework and applying it to a practical environment soon. As indicated in the previous sections, the primary candidate case would be a multi-user video conferencing service on the cloud. We will perform the case study for the above scenario and report the results. For the future work, we hope to expand our framework to provide much comprehensive performance metrics than the current version and enable it to handle more versatile applications on the cloud.

Acknowledgements

This work was supported by the National Research Foundation of Korea Grant funded by the Korean government (No. 2011-0013924) and a grant from Samsung Electronics.

Reference

1) Simson L. Garfinkel,"An Evaluation of Amazon’s Grid Computing Services: EC2, S3 and SQS", Harvard University Technical Report TR-08-07, 2007.

2) Jörg Schad, Jens Dittrich, and Jorge-Arnulfo Quiané-Ruiz, "Runtime measurements in the cloud: observing, analyzing, and reducing variance", Proc. of VLDB Endow., 2010.

3) Guohui Wang and T. S. Eugene Ng, "The Impact of Virtualization on Network Performance of Amazon EC2 Data Center", Proc. of IEEE INFOCOM, San Diego, CA, March 2010.

4) Ang Li, Xiaowei Yang, Srikanth Kandula, and Ming Zhang, "CloudCmp: comparing public cloud providers", Proc. of annual Conference on Internet Measurement (IMC), 2010.

5) D. Patnaik, A. S. Krishnakumar and P. Krishnan, "Performance Implications of Hosting Enterprise Telephony Application on Virtualized Multi-Core Platforms," Proc. of International Conference on Principles, Systems and Applications of IP Telecommunications, 2009.

6) A. Uyar, S. Pallickara and G. Fox, "Towards an Architecture for Audio/Video Conferencing in Distributed Brokering Systems," Proc. of International Conference on Communications in Computing, 2003.

7) J. Cervino, P. Rodriguez, I. Trajkovska, A. Mozo and J. Salvachua, "Testing a Cloud Provider Network for Hybrid P2P and Cloud Streaming Architectures," Proc. of IEEE International Conference on Cloud Computing(CLOUD), 2011.

8) M. Chesire, A. Wolman, G. M. Voelker and H. M. Levy, "Measurement and Analysis of a Streaming-Media Workload," Proc. of conference on USENIX Symposium on Internet Technologies and Systems, 2001.

9) B. Dong, Q. Zheng, J. Yang, H. Li and M. Qiao, "An E-learning Ecosystem Based on Cloud Computing Infrastructure," Proc. of IEEE International Conference on Advanced Learning Technologies, 2009.

10) Dipzoom, http://dipzoom.case.edu.

11) Z. Wen, S. Triukose, and M. Rabinovich, “Facilitating focused internet measurements,” Proc. of ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, 2007.

12) K. Gummadi, S. Saroiu, and S. Gribble, “King: Estimating latency between arbitrary internet end hosts,” Proc. of the Second SIGCOMM Internet Measurement Workshop, 2002.

13) T. S. E. Ng and H. Zhang, “Predicting internet network distance with coordinates-based approaches,” Proc. of IEEE INFOCOM, 2002.

14) Y. J. Won and J. Hong, "Measurement of Download and Play and Streaming IPTV Traffic," IEEE Communications Magazine, 2008.