Evaluation of Latency in Delay Sensitive Cloud Services

*Rahul Kumar Sharma[1], Mayank Deep Khare[2], Amrendra Singh Yadav[3], Ayushi Singhal[4]
[1234]*Assistant Professor, Department of Computer Science & Engineering, NIET, Greater Noida, India*
rahulsharma9045@gmail.com, mayankdeepkhare20@gmail.com,
yadavamrendrasingh@gmail.com, zealous.ayushi@gmail.com

Abstract- **Cloud computing makes the dream of computing real as a tool and in the form of service. This internet - based ongoing technology which has brought flexibility, capacity and power of processing has realized service- oriented idea and has created a new ecosystem in the computing world with its great power and benefits. Cloud capabilities have been able to move IT industry one step forward. Nowadays, large and famous enterprise has resorted to cloud computing and have transferred their processing and storage to it. Due to popularity and progress of cloud in different organizations, cloud performance evaluation is of special importance and this evaluation can help users make right decisions. In this paper we define the cloud performance issues which are cause for quality of any cloud or database. In the cloud we can improve the performance by the reducing these issues. We also provide a large view of latency which is a hot topic in the research field for improving the performance of cloud.**

Keywords- **Cloud Computing, Latency, Virtualization, Jitter, Round trip time.**

I. INTRODUCTION

cloud computing enables a new business model that supports on- demand, pay-for-use, and economies-of-scale IT services over the Internet. The Internet cloud works as a service factory built around virtualized data centers.1 Cloud platforms are dynamically built through virtualization with provisioned hardware, software, networks, and datasets. The idea is to migrate desktop computing to a service- oriented platform using virtual server clusters at data centers. However, a lack of trust between cloud users and providers has hindered the universal acceptance of clouds as outsourced computing services. To promote multitenancy, we must design the cloud ecosystem to be secure, trustworthy, and dependable.2 In reality, trust is a social problem, not a purely technical issue. However, we believe that technology can enhance trust, justice, reputation, credibility, and assurance in Internet applications. To increase the adoption of Web and cloud services, cloud service providers (CSPs) must first establish trust and security to alleviate the worries of a large number of users. A healthy cloud ecosystem should be free from abuses, violence, cheating, hacking, viruses, rumors, pornography, spam, and privacy and copyright violations. Both public and private clouds demand "trusted zones" for data, virtual machines (VMs), and user identity, as VMware and EMC3 originally introduced. Data integrity issues in the cloud. differ from those in traditional database systems. Cloud users are most concerned about whether data-center owners will abuse the system by randomly using private datasets or releasing sensitive data to a third party without authorization. Cloud security hinges on how to establish trust between these service providers and data owners. To address these issues, we propose a reputation-based trust-management scheme augmented with data coloring and soft-ware watermarking. Information about related trust models is available elsewhere.2,4

II. PERFORMANCE ISSUES

A. *Virtual Machine Migration* -Virtual Machine migrations provide major benefit in the cloud computing through load balance across the data centre. It provides robust and high response in the data centre. Virtual Machine migration was extended from the process migration. During data transfer virtual machine migration maintains consistency for application by considering resources and physical servers.

B. *Server consolidation*-Server consolidation is an efficient approach is to minimize energy consumption for make best use of resource utilization. Server consolidation is not depending on the application performance. Sever consolidation is known as the resource usage means individual VMs may vary time to time. When resource congestions are occur, then system react quickly.

C. *Performance Unpredictability*-Sharing I/O is complex in the cloud computing while multiple VMs can share CPU and main memory easily.

*Author for correspondence

Virtualization of I/O interrupts and channel is one solution to improve efficiency of the operating system and improve the architecture. Another unpredictability problem concerns to the scheduling of virtual machine for various classes of batch processing programs, exclusively for high performance computing. The problem is that various HPC applications require ensuring that all threads of the program are running simultaneously.

D. *Bugs in Large-Scale Distributed System*-Another challenge issue in the Cloud Computing is removing errors in these large scale distributed systems. The debugging of these bugs has to be done at the large scale in production data centers as these bugs cannot be reproduced in the smaller configurations.

E. *Quickly Scaling*-Pay-as-you-go certainly applies to the network bandwidth and storage on basis of used bytes count. Depending on the virtualization level, computation is slightly different. Google AppEngine automatically scales in response to the load up and down, and users are charged by usage of the cycles.

F. *Latency*-Latency is research issue in the Internet. Any performance in the cloud computing is going the same meaning of performance of the result on the client. The latency in a cloud introduces is not to be tedious. Latency is compressed back for understand where and how they are running with both smartly-written application and intelligently planned infrastructure. In the future, cloud computing capacity and cloud based applications are rapidly increases and the latency is also increases.

III. CLOUD LATENCY

Latency is research issue in the Internet. Any performance in the cloud computing is going the same meaning of performance of the result on the client. The latency in a cloud introduces is not to be tedious. Latency is compressed back for understand where and how they are running with both smartly-written application and intelligently planned infrastructure. In the future, cloud computing capacity and cloud based applications are rapidly increases and the latency is also increases. In cloud, Latency can occur due to delay in the network cause of congestion. Latency can be measure by applied some formula or algorithm

Even fiber optics are limited by more than just the speed of light, as the refractive index of the cable and all repeaters or amplifiers along their length introduce delays.

A. *Intra Cloud Latency*

In the cloud, the Latency could be occur when two virtual Machine co-located at the same server communicate with each other. Nahanni is introducing this problem, a port of well known memcached that uses inter-VM shared memory instead of virtual network of for cache read. Facebook is an example for employs memcached as one of the several caching layer

B. *Network Latency*

Network Latency cause application to spend amount of time waiting for response from distant data centre, then the bandwidth may not be fully utilized and then performance will be suffer. Good network design can minimize node delay and congestion delay but not propagation delay.

- *Propagation Delay* - It is time duration which is taken to packet travel between one place to another place at the speed of light.
- *Transmission Delay*- The medium itself introduce some delay, which vary from one medium to another medium. The size of the packet data introduce delay in a round sense a large packet take longer to received and return than the short one.
- *Congestion Delay*- Network congestion occur when a node is carrying so much data that its quality of services deteriorates.
- *Processing Delay*- The time which take to process the data is called processing delay. It is important component of network delay. Let us consider the distance between to LAN to be 400 miles and assume each router adds 2ms. Current network utilization without storage application is 16%. So amount of bandwidth available for new storage network application is to be 85%.The distance between two end points of network link is 400 miles. Therefore round Trip Time propagation delay is 400*2=800miles or equal to 8ms. If there are two routers in the path taken by data. So estimated RTT transmission delay is 2 nodes*2ms equal to 4ms. Now with the congestion processing delay is increased to 4ms/0.85 equal to 5ms. Hence total network latency [Propagation Delay+ Transmission Delay + Congestion Delay] is 17 ms {i.e. 8ms+4ms+5ms}.

IV. RELATED WORK

Simulation in wide environments such as cloud computing environments may be done in different categories, different environments or different criteria. The references in this study are in the following categories and have been conducted since 2013 onwards: Performance evaluation based on different criteria, evaluation and simulation is usually performed based on criteria in accordance with goals and the results are also studied for proving goal. In studies related to cloud computing performance evaluation , some criteria such as average waiting time, load balancing and number of requests , cost and throughput, workload , the rate of transactions [1], response time [1], time of allocation and release of resources [2], different scheduling algorithms [2, 3], effectiveness, delays in service and productivity [3], the number of input and output operations in the network , etc. are studied.

Evaluation based on different methods, tools and simulations for cloud environments and other web-based environments, there are various methods and simulation tools, such as parto traffic methods, fuzzy systems, a discrete event simulation [3], a tool like CloudAnalyst and etc.

Performance evaluation based on specific applications or services, Variety of services and applications are offered in cloud environments each of which can be topic for the simulation scenarios, such as scientific computing, e-learning software, high-performance computing, inbound and outbound applications on the network, services with assuring quality of service, Multi-Tier Cloud Applications and etc.

Raihana Abdullah, M.Faizal Abdullah, Zul Azri Muhamad, M.Zakri Mas, Siti Rahayu Selamat and Roboah Yusuf in [4] had address current trend of Botnet detection techniques and identifies significant criteria in each technique. Several existing are analyzed from various researcher and capability criteria of Botnet detection techniques are analyzed. These techniques have been shown on the selected detected criteria.

Predeep Sharma, Sumeet Kaur and sandeep sood [5] had proposed the benefits of cloud computing along with its flip side. This paper is also introduces various issues in cloud computing and suggested the possible measure to overcome them and data proposed algorithm is used to calculate and net revenue by using the cloud and data centre.

V. IMPACT OF LATENCTY

In past, the latency has three different measures: jitter, endpoint computational speed and roundtrip time (RTT). Now adding traceroutes as a tool, each of these is important to the understanding true effect of latency, and only after understanding each of these metrics can you get full picture.

- Round trip time measures time it takes one packet to transit to network from source to destination and back to source, or time it takes for the initial server connection. This is useful in the interactive applications and it also in examining app-to-app situations, like as measuring way a Web server and database server interact and exchange data also.
- Jitter: - Jitter is a variation in the packet transit delay caused by the queuing, contention and serialization effects on path through the network. So this can have a large impact on interactive applications such as video or voice in the network.
- The speed of computers at core of the application: their configuration will determine how quickly they can process data. While this seems so simple, it can be difficult to calculate once we start using the cloud-based computer servers.

Trace route is a popular command which examines individual hops or the network routers that a packet takes to go from one place to another. Each hop can also introduce less or more latency. For example: the lowest latency and the fastest path between a computer in Sydney Australia and one in Singapore. Let's keep each of these in the mind as we look at how the cloud plays into the each calculation.

Many businesses are extremely demanding and will continue to require lowest latencies possible from their Internet connections. Applications such as algorithmic frequency trading, video streaming, more complex web/3-D engineering modeling and database services are in this category. But some applications such as email, analytics and some kinds of document management are not as demanding. As a way down the path towards understanding the latency, perhaps we need to start with some kind of triage and separating those applications that will really benefit from the lowest latencies.

We preparing a requirement based chart such as the one shown opposite that classifies our applications according to their various properties of computing intensity, the network bandwidth and latency requirements [8].

Advantage and disadvantage: According to Equation Research3 sanctioned by Gomez, a poor web experience results in lost revenue opportunity, a decreased customer perception of your company and can drive customers to your competitor.

78% of site visitors have gone to a competitor's site due to poor performance during peak times. 88% are less likely to return to a site after a poor user experience. 47% are left with a less positive perception of the company. Aberdeen Group4 provides a similar snapshot of demanding user requirements for website performance.

A one second delay reduces customer conversions by 7%. A one second delay decreases page views by 11%.

Shopzilla: A 5-second increase in website performance resulted in 25% more page views, a 12% increase in revenue and a 50% reduction in hardware.

Amazon.com: Every 0.1 seconds in latency reduces sales by 1%.

Google: Every 0.5 seconds in latency reduces traffic by 20%.

VI. SOLUTIONS FOR IMPROVING CLOUD LATENCIES

Providing more consistent network As you can see, it isn't just poor latency in the cloud but the unpredictable nature of the various network connections between your on-premises applications and your cloud provider that can cause problems. What is needed is some way to reduce these daily or even minutely- minute variations so you can have a better handle on what to expect.

For the connection from the customer's premises to the Direct Connect locations these metrics will be subject to strict quality of service guarantees, i.e. a bandwidth of X with a defined maximum latency and an SLA for the connection. For the connection from the Direct Connect locations to Amazon Web Services cluster in your region, you can expect improved network characteristics but there is not an SLA that defines guaranteed bandwidth.

Direct connect options like Amazon, but also Windows Azure, offer to build a hybrid solution that uses both on-premise and cloud-based resources. Application code and data can be stored in an appropriate on- premise location according to regulations, privacy concerns, and a measurement of acceptable risk, while solution components requiring the features and pricing model of cloud computing can be migrated to the cloud.

Traceroute is a popular command which examines individual hops or the network routers that a packet takes to go from one place to another. Each hop can also introduce less or more latency. For example: the lowest latency and the fastest path between a computer in Sydney Australia and one in

TABLE I: Requirement infrastructure in cloud[on the basis of high to low level network Latency

Applications	Computer Intensity	Network BW	Network Latency
Testing & Development	High	Low	High
Web Browsing	Low	Low	High
Backup & Recovery	Low	High	High
Email & Calendar	Low	Low	High
HRM	Medium	Low	High
Document Management	Low	High	Medium
CRM	Medium	Low	Medium
Finance & Accounting	High	Medium	Medium
ERP	High	Medium	Medium
Payment & Transactions	Medium	High	Medium
Virtual Desktops	Medium	Medium	Low
Network Storage	High	Medium	Low
Unified Communication	High	Medium	Low
Online Gaming	High	Medium	Low
HD Video Streaming	High	High	Low
Black Box (M2M) Trading	High	High	Proximity

Singapore. Let's keep each of these in the mind as we look at how the cloud plays into the each calculation.

Many businesses are extremely demanding and will continue to require lowest latencies possible from their Internet connections. Applications such as algorithmic frequency trading, video streaming, more complex web/3-D engineering modeling and database services are in this category. But some applications such as email, analytics and some kinds of document management are not as demanding. As a way down the path towards understanding the latency, perhaps we need to start with some kind of triage and separating those applications that will really benefit from the lowest latencies.

VII. CONCLUSIONS AND FUTURE WORK

This paper highlighted evaluation on effect of latency in cloud computing and also present a view of latency and how we measure it network and we try to define the problem which increase the latency. We need to develop algorithm which search the minimum delay path between two nodes. As the future work, we can direct to our

research towards reducing the latency in the cloud access data. In addition we plan to apply methodology to more platform to understanding their strengths and weakness.

REFERENCES

[1] Vladimir Stantchev, (2013) "Performance Evaluation of Cloud Computing Offerings"; Third International Conference on Advanced Engineering Computing and Applications in Sciences IEEE

[2] Nezih Yigitbasi, (2013) "C-Meter: A Framework for Performance Analysis of Computing Clouds", IEEE/ACM International Symposium on Cluster Computing and the Grid

[3] Ioannis A. Moschakis & Helen D. Karatza , (2011) "Performance and Cost evaluation of Gang Scheduling in a Cloud Computing System with Job Migrations and Starvation Handling", IEEE

[4] Raihana Abdullah, M.Faizal Abdullah, Zul Azri Muhamad, M.Zakri Mas, Siti Rahayu Selamat and Roboah Yusuf," Revealing the Criterion on Botnet detection technique " IJCSI, Vol.10 issue 2.

[5] Pardeep Sharma, Sendeep Sood and Sumeet Kaur, "Cloud computing Issues and what do compute on cloud " in ICACCN.

[6] Chang F, Dean J et al (2011) Bigtable: a distributed storage system for structured data. In: Proc of OSDI

[7] CAR R, N.Rough Type [online]. 2013. Available from: http://www.roughtype.com.

Rahul Kumar Sharma completed M.Tech. (Computer Science) from Madan Mohan Malaviya university of Technology, Gorakhpur in 2015.He is working at NIET, Greater Noida for last one year as a dynamic, successful and responsible faculty

Mayank Deep Khare completed M.Tech. (Information Technology) from Madan Mohan Mohan Malaviya University of Technology, Gorakhpur in 2015. He is working at NIET, Greater Noida for last one year as Assistant Professor in department.

Amrendra Singh Yadav is Assistant Professor of Computer Science & Engineering at Noida Institute of Engineering & Technology, Greater Noida. He recieved his Masters degree in Infomation Technology from Madan Mohan Malaviyan University of Technology, Gorakhpur (U.P) His Research interests include Mobile ad-hoc network, wireless sensor network. He has five research papers in national and international journals and conference.

Ayushi Singhal is Assistant Professor of Computer Science & Engineering at Noida Institute of Engineering & Technology, Greater Noida. She recieved his Masters degree in Infomation Technology from Madan Mohan Malaviyan University of Technology, Gorakhpur (U.P) Her Research interests include Mobile ad-hoc network, wireless sensor network. She has three research papers in national and international journals and conference.