INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 6, ISSUE 4, APRIL-2015
ISSN 2229-5518

Survey on Enhancing Infrastructure Scalability by Predicting Future Requirements

Wegdan Altayeb, Ahmed Kayed

Abstract — Automatic resources scalability, performance and resources management are required properties for successful services in cloud infrastructure. There are some metrics that associated scalability and performance in cloud computing; that influenced by some software factors, hardware factors and workload factors. Infrastructure scalability and performance metrics which used to determine how to scale data centers resources and servers are critical issues for studies in cloud environments, since these metrics such as CPU load, memory load and throughput are changeable and have great impacts on scalability and performance. Predicting future resources changes in customer's requirement helps cloud providers to avoid wasted resources, cost and achieves high resources utilization. The paper will provide a survey for computing resources scalability, performance and management. Provides listed scalability servers and network metrics, scalability techniques, resources management and load balancing strategies, performance prediction.

Keywords— Horizontal scalability, Metrics, Performance, Prediction models, Resources managements, Scalability factors, Vertical scalability

1 Introduction

Scalability aims to determine when and how to scale resources. CSPs (Cloud Service Provider) need to determine customer's requirements dynamically and scale up or down as fast as they can to support required services with quality of services. They need to plan resources capacity and usage effectively to achieve enhanced resources utilization. Infrastructure as a service (Iaas) resources are automatically scaled, but the main issue is how to decide when and how to scale rely on some scalability metrics and factors that influence performance; metrics such as: servers and network metrics can be used as indications to forecast future scalability decisions.

The paper gives a survey on some scalability and performance related metrics and factors such as: response time, CPU usage, delay and server workload, methods and techniques that enhancing scalability.

- *Wegdan Altayeb is currently preparing for a PHD degree in cloud computing infrastructure in Sudan University of Science and Technology (SUST), Sudan. E-mail: wegdan_a_hamed@ hotmail.com*
- *Prof Ahmad Kayed awarded his PhD from University of Queensland, worked with Monash university –Australia. Prof. Kayed published in high impact journals (more than 300 citations). He held many hi administrative tasks: Dean of IT Faculty –MEU, University Research Board, Board of Trustees, and General Manager, etc. E-mail: drkayed@ymail.com*

2 Scalability

Scalability in cloud computing is the ability to increase or decrease computational resources numbers to process workload growing , while considering and maintaining performance. It requires automatic configuration, sizing of resources and enhancing throughput when extra resources are added [1, 2, 3].

2. 1 Scalability techniques

2.1.1 Vertical scalability

Vertical scalability is ability to maximize hardware or software efficiency by increasing resources [4]. The method provides ability to replicate servers or change the size of them. The main benefit of this method is the effective use for virtualization technique which provides resources sharing. The main goal of vertical machines scalability is to make work load threshold limit [5]. Cost and bandwidth requirements are the most points that use this scaling method. In hardware side this method concerns about adding processing power and memory, while for software concerns with optimizing algorithms and cost [1]. The drawback with vertical scalability appears when the server requests are huge which makes delay and also resources dead lock may occur [5].

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 6, ISSUE 4, APRIL-2015
ISSN 2229-5518

2.1.2 Horizontal scalability

Horizontal scalability focuses on handling demands increment by adding more virtual machines .In this method server's size are not change, but they can be replicated for more demands [6].The main advantage of this scaling technique is the high processing performance [7]. To reach target availability and speedy servers load balancing and clustering characteristics must applied. One of critical problems with cloud computing is availability and server failure which affect the hosted virtual machines and force them to restart at another server. Horizontal scalability solves that problem by adding more servers' hardware or software and makes them to work as one fail-safe unit [6].

Table.1 Vertical Scalability vs. Horizontal Scalability

Metric	Vertical Scalability	Horizontal Scalability
Implementation	- Done by regulating server's size. - Replace server by another new with more capabilities [8].	- Adding more servers and divide workload between them [6].
Availability	- Less resources availability, resources located only on same servers. - Server will be a single failure unit [9].	- Provide more resources availability since numbers of servers are increased. - More servers reduce failure probabilities [9].
Reliability	- Reliability is not ensured, server failure means to all system failure [10].	- Supports ensured reliability; redundant servers are available [10].
Simplicity	- Complex.	- Simple and easy to achieve.
Performance improvement	- Achieved by adding more resources.	- Achieved by adding more servers and distributing load between them.
Cloud architecture	- Not affected	- Architecture of the cloud will be Changed [11].
Resource allocation	- Provides a fine-grained allocation [12].	- Provides a coarse - grained allocation [12].
Quality of service(QoS)	- Achieved by identifying the optimum resources for virtual machines to satisfy user's requirements [12].	- Redundancy of physical servers and load balancing.

2.2 Scalability Metrics

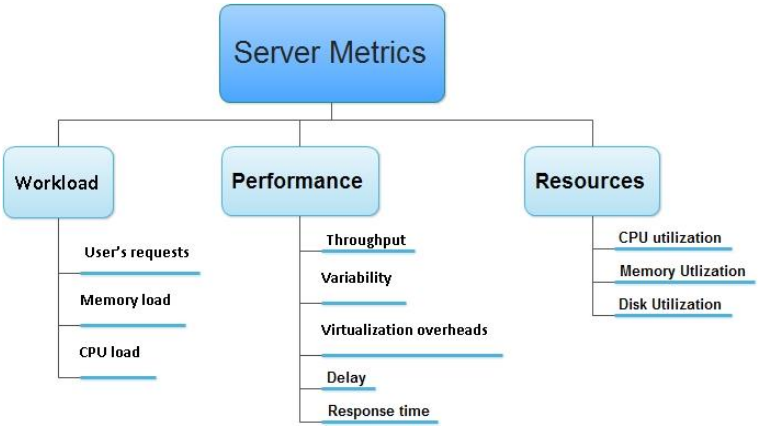

Fig. 1 Server scalability metrics

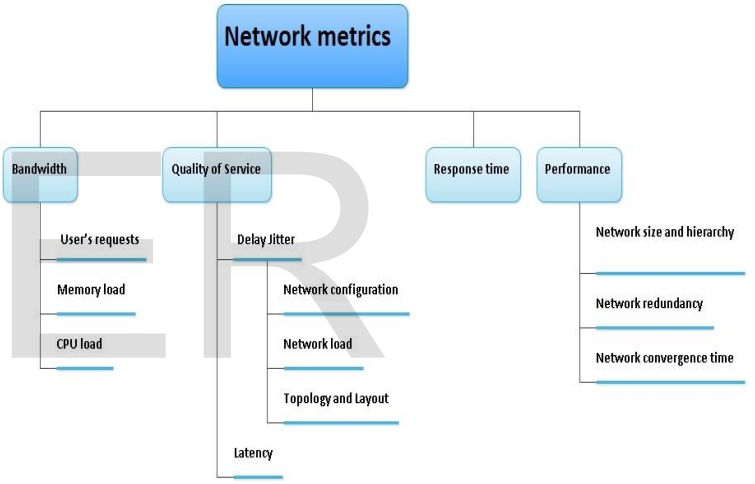

Fig. 2 Network scalability metrics

2.3 Scalability Factors

Scalability is qualified property for cloud environments which targets to enhance cloud performance and quality of service (QoS). Scalability described as ability to adjust resources as the workload increased; there some measurement factors must be consider and their associated metrics must specified to achieve desired scalability. Scalability factors are measurements of actual resource usage [13]. System factors that impact scalability metrics are classified as: hardware factors, software factors and load factors [14]. Table 2 provides some hardware scalability factors and how to measure them.

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 6, ISSUE 4, APRIL-2015
ISSN 2229-5518

Table.2 Scalability hardware factors

HW factors	CPU computing power and usage	Measurement (s) Throughput (The received data in a second by the destination) [15]. 2. Speedup (ratio between throughput and CPU core). 3. CPU utilization (percentage of CPU core) [14]. 4. CPU load (Cycles perform for the current processing task) [16].
	Memory usage and consumption	1. The number of accessed pages included the cached pages. 2. Maximum throughput. 3. Memory used by Applications/services [17].
	Disk usage and storage system	1. Disk time (I/O per second). 2. Execution time (E-time relative to storage system) 3.Load (Disk operations performed per second) [16].
	Virtualization overhead	1. Workload and CPU overheads. 2. Average response time. 3. Throughput. 4. Memory access overheads (when cache misses increase). 5.Network virtualization overheads [15].
	Throughput(Th)	1.Latency(hops and hosts numbers from user to reach another node on the network) - Latency= processing delay+ queuing delay+ transmission delay + propagation delay [18] 2. Response time. 3. Bandwidth (Th = total B – B required for reply) -Workload (Requests and responses sizes) [14].

3 RESOURCES MANAGEMENT STRATEGIES

While cloud data centers provide customers with required resources, those resources should be sufficient. Some problems arise behind those computational resources such as bandwidth, response time and delay; resources can be managed among applying load balancing techniques considering some parameters such as performance and scalability [19]. Table 3 provides a brief comparison between some load balancing methods for managing cloud resources.

Table.3 Resources management and load balancing strategies

Strategy	Metrics	Target	Methodology
Linear scheduling	-Response time - Waiting time. -Process activities[2o]	-Resource distribution to increased QoS, resources utilization [21].	-Forecasting first resources response for specific period of time [21].
Process Migration Strategies	Bandwidth. -Migration Overheads [22]. -Migration delay [23].	- Process live migration [23].	-Page level protection hardware [22].
Match Making	-Response time, Job features and hardware [24].	- Identifying jobs order execution and mapping to resources [25].	- Match user task to resources [25].
Just-in-Time Resource Allocation	-Cost of: SLA, reconfigured Application, and machine leased and released [25].	-Workload provisioning and optimization [21]. - Minimize idle resources [25]	-Adjust interval time for resources and continuously monitoring workload [25].

4 RESOURCES AND PERFORMANCE PREDICTION

4.1 Performance prediction

Services, applications and resources performance in cloud environments are effected by some factors such as virtualization factors, network factors and servers workloads. Performance forecasting is big challenge and need to measure, manage, execute some workloads, and gather resources usage information, monitor resources and any other factors. Gathering some historical data helps grasping and clarifying reasonable performance and scalability of resources in cloud environment. Understanding cloud performance needs to gather reasonable amount of data among some experiments and measure that data. There are some complicated issues which go beside performance measurement such as configuration, data gathering and processing, data analysis, data heterogeneity, parameters configuration, resources and availability; all these influence performance and scalability which makes prediction more complex process. Figure 3 provides cloud computing performance prediction methodologies [28, 29, 30, 3].

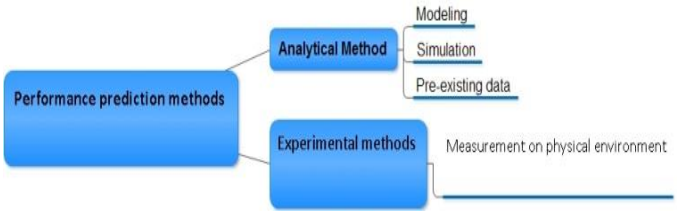

Fig. 3 Performance prediction methods

4.2 Resources scalability prediction models

Accurate scalability prediction can be achieved among black box or white box methods. Table 4 compares the three prediction and testing approaches: white box, black box and grey box.

Table. 4 Resources prediction models comparison

Metric	White box	Black box	Grey box
Complexity and Prediction	- Complex, high cost. - Not appropriate for performance prediction [32].	-Less complexity, less cost, simple [32]. -Suitable for performance and scalability prediction [32].	-Not suitable for scalability and performance prediction [33].
Testing	-Test source code, not look at the general system. - Useful for testing stages [34].	- Tests the system as general [34].	- Enables programmer to coding for expecting results [35].
Evaluation	-Provides maximum level of information For both evaluator and end user [36].	-Provides minimum level of information. - Suitable for vendors [36].	-Offer sufficient level of information for evaluation [36].
Framework	- End user should aware about how the white box framework work and all details [33].	-End user not aware about framework details and its works [33].	- Used by programmer who should know grey box internal details [35].
Program Code Analysis	-Deals with program source code -Provides static code analysis -may provide non-actual reports [33].	-Look for the general knowledge and deals with applications not source code [33].	- Uses static analysis of white box [33].

4.3 Scalability prediction and knowledge representation approaches

Recently, most complex systems and problems are represented mathematically using model base approaches such as control theory. Model base approaches and mathematical formulation give comprehensible and reasonable results. While model base defined mathematically, rule based approaches such as fuzzy logic and neural network are used in soft computing [37]. Table 5 provides comparison between rule base and model based approaches.

Table. 5 Scalability prediction approaches comparison

Approach	Algorithms	Advantages	Disadvantages
Rule-based	-Reinforcement learning. - Statistical machine learning. - Fuzzy control. -Support vector machine [38]. - Neural networks (NN). -Genetic algorithms (GA) [37].	-Can infer required resources using historical data, without a knowledgeable model. -Suitable for complex systems. -Not need huge knowledge [38]. -Provide appropriate level of data classification; produce more rules and accuracy [39].	-QoS and performance requirements are not ensured. - Very complex in computation. - Require long time for learning from the data, which affect the performance [38]. -Hard to provide system generalization, because of small set of rules [39].
Model based	-Control theory (such as using black box). -Queuing theory [38].	- Ensured QoS. -Provide strict approach to model, analyze and understand the system [38]. - Proper to use in knowledge-based system that help to show, interpret and maintain the knowledge. - Have obvious features and simple in design [39].	-This approach needs much knowledge [38]. -Difficult to formulate as it's represented mathematically [37].

5 Related Works

Automatic scaling is an important issue for both cloud providers and consumers to get adequate level of cloud resources and cost. Predicting future changes in resources requirements helps to choose the right decisions to scale cloud resources (computing, storage, and network).

Some related works are proposed for automatic scaling in cloud environments. In [40] a prediction model for performing automatic scaling resources was proposed. The work focuses on studying system behavior and resources using the past performance information and provide a framework for inferring demands, make decisions and prediction. The author uses machine learning techniques for decision making and prediction processes.

Chandan Banerjee1, Anirban Kundu2, et al. In [41] proposed scalable resources selection framework in cloud computing and load balancing. They show in their framework that resources are accessed by the customer as flow from cloud service layer at the top till the resources location at the bottom. So they use top-down engineering technique. The model consists of four layers which are: Cloud Service Abstraction, Resource Manager, Load Sharing and Load Balancing and Resource Allocation. Customers access cluster resources by cluster index which hold information such as: number of resources in a cluster, load balancing selection scalability factors. They use CPU, memory measurement as scalability factors.

In [42] dynamically cloud resources architecture is provided with SLA violation avoidance. The problem is divided into three units: host dynamic configuration and SLA, monitoring components and quality of services requirements, and finally dispatching and load balancing. The architecture uses actual response time as load balancing metric. Fuzzy logic and expert knowledge are used to model information in non-numeric form, but linguistic variables. The system scaling is modeled by using IF-Then rules. The architecture of proposed system has two scaling modules: blue box, data collector and fuzzy control. Data collector gathers data such as CPU usage and response time, controls scaling action data, load data and categorized the data to be ready for using by fuzzy control module. The result sends to inference engine and the de-fuzzified values are the VM[s] number need to start or stop to reach the target scaling. The scaling data send to scale control model which produces cloud management system.

Petter Sv¨ard in [8] shows enhanced scalability approach (server disaggregation approach) that aggregates resources to create virtual machine larger than physical machine that host it. The approach uses vertical scaling on-demand and not takes into account physical machines boundaries; physical machine permits the guests to use resources from more than one physical machine. The main benefit provided by the approach is the high improvement of physical machines utilization in cloud data center.

Sadeka Islam , Jacky Keunga, et al. In [9] proposed a performance prediction model for dynamic application scalability by estimating future resources demands and predicting them; targeted to improve performance and availability. The model concerns with resources prediction from cloud service provider point of view. The model applies standard benchmark to generate historical data and uses that data for prediction among several learning algorithms (Error Correction Neural Network (ECNN) and linear Regression. Precision of the prediction framework is evaluated using some metrics (Mean Absolute Percentage Error (MAPE, Root Mean Squared Error (RMSE) and R2 Prediction Accuracy.

6 Conclusions

The paper provides a survey on Infrastructure as a Service (IAAS) scalability enhancement. Its focuses on scalability metrics and factors that impact performance; gives some methods and approach which used for IaaS resources prediction.

The paper shows many methods, models, approaches existed for enhancing performance and scalability in cloud data centers. Scalability prediction entails intensive measurements on workload, server and network.

Automated scalability and performance are required features to build successful cloud service provider (CPS) network and provide qualified, reasonable cost, services and resources to customers.

- Future Recommended Issues

Some solutions for enhancing cloud scalability are existed, but most of them use:

- Current states of scalability parameters.
- Use one technique for each solution process such as prediction, learning and evaluation.
- Use one scalability technique (vertical, horizontal).

Three points mentioned above clarify that there are wide needs to cover these research areas:

- Consider intensive past and present scalability measurements.
- Apply more than one technique or tool and compare their results for scalability parameters, specifying which tool is more perfect for each parameter.
- Mix vertical and horizontal scalability approaches to get better performance, management, cost benefit and better quality of service (QoS).

References

[1] Nezih Yigitbasi, Alexandru Iosup, Dick Epema, "C-Meter: A Framework for Performance Analysis of Computing Clouds".

[2] Eileen Marie Hanna, Nader Mohamed, Jameela Al-Jaroodi, "The Cloud: Requirements for a Better Service", 2012, 12th IEEE/ACM International Symposium on Cluster, Cloud and

Grid Computing.

[3] Daniel .F. GARCIA, Rodrigo GARCIA, and Joaquín ENTRIALGO, "Experimental Evaluation of Horizontal and Vertical Scalability of Cluster-Based Application Servers for Transactional Workloads", 2008.

[4] Lloyd G. Williams, Connie U. Smith, "Web Application Scalability: A Model-Based Approach", Software Engineering Research and Performance Engineering Services 2004.

[5] Rahul Sharma, Mohit Mathur, "Achieving Vertical Scalability: A hindrance to Cloud Computing", National conference In=NDIACom-2010.

[6] Herminder Singh & Babul Bansal," analysis of security issues and performance enhancement In cloud computing".

[7] S. Bibi, D. Katsaros, and P. Bozanis, "Application Development: Fly to the clouds or stay in-house?" in Proceedings of the 2010 19th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, ser. WETICE '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 60–65.

[8] Petter Sv¨ard,"Dynamic Cloud Resource Management", PHD thesis, April 2014 , department of computing science Ume°a university Sweden.

[9] Sadeka Islam, Jacky Keung, [etal], "Empirical prediction models for adaptive resource provisioning in the cloud", Elsevier journal 2012, Future Generation Computer Systems 28 (2012) 155–162.

[10] Andre Romeo Botes, "An artefact analysis unstructured document data stores", school of information technology , North-West university, master degree research, 2014.

[11] Peter Mell ,Timothy Grance , "The NIST Definition of Cloud Computing ",Reports on Computer Systems Technology , National Institute of Standards and Technology, 2011.

[12] D Huang, B He, C Miao, "A Survey of Resource Management in Multi-Tier Web Applications", IEEE 2014.

[13] Will Venters, Edgar A Whitley, "A critical review of cloud computing: researching desires and realities", Journal of Information Technology (2012).

[14] Nuno Filipe, "State Machine Replication: from Analytical Evaluation to High-Performance Paxos", PHD thesis, 2012.

[15] Pradeep Padala1, Xiaoyun Zhu2, ZhikuiWang2, "Performance Evaluation of Virtualization Technologies for Server Consolidation", University of Michigan, Enterprise Systems and Software Laboratory, April , 2007.

[16] K. N. Honwadkar, T. R. Sontakke," Configuring Hard Disk Drives On Nodes Of A Lan Toimprove Capacity And Efficiency Of Network Storage", International Journal of Information Technology and Knowledge Management, Volume 4, No. 2, pp. 659-667, 2011.

[17] Arupratan Santra, I V Murali Krishna, Anindita Das, "Measurement of memory usage in J2EE applications", Journal of Scientific and Industrial Research, Vol 68, pages: 786, 788, 2009.

[18] Minseok Kwon," A Tutorial on Network Latency and its Measurements", Dept. of Computer Science, Rochester Institute of Technology, 2014.

[19] Nidhi Jain Kansal and Inderveer Chana , "Cloud Load Balancing Techniques: A Step Towards Green Computing", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, 2012 .

[20] Veronica Haring, "Linear Scheduling: Special Purpose Simulation Template Developed for Simphony.Net", Master of Science, University of Alberta, 2014.

[21] Abirami S.P., Shalini Ramanathan , "Linear Scheduling Strategy for Resource allocation in Cloud Environment", International Journal on Cloud Computing and Architecture, vol.2, No.1, February 2012.

[22] Rakhi k Raj and Getzi Jeba Leelipushpam., "Live Virtual Machine Migration Techniques – A Survey", International Journal of Engineering Research and Technology, Volume 1 Issue 7, September 2012.

[23] Amirreza Zarrabi, "A Generic Process Migration Algorithm", "International Journal of Distributed and Parallel Systems (IJDPS)", Vol.3, No.5, September 2012.

[24] Malarvizhi Nandagopal et. Al, "fault tolerant scheduling Strategy for computational grid environment", International Journal of Engineering Science and Technology", Vol. 2(9), 2010.