Modeling and Analysis of Cloud Computing Network using M/G/k/PS Queues

Samer Salah Al_Yassin

Abstract: ***The cloud computing are gaining more and more popularity in business, enterprises and research areas. Flexible and transparent provision of services on-demand, height virtualization level, seamless interconnection between different network nodes are the reasons of good performance and scalability of cloud computing network. However cloud computing technology has not universal requirements and solutions for network deployment and implementation. The lack of standards leads to performance and functionality reduction. The analysis of functionality and performance of cloud computing solutions give ability find "bottlenecks" and improve services provision processes. The difference types of cloud computing architecture are analyzed in the paper. The different approaches to the performance and functionality analysis, such as real measurements, imitation and analytical modeling are observed. The analytical model of cloud network based on queuing theory is proposed. The selecting of queuing model parameters depended on the technology platform, the model of cloud services (SaaS, PaaS, IaaS), the type of applications. Proposed queuing model allows analyzing the main network characteristics affected on performance and services provision effectiveness.***

Keywords: cloud computing, queuing theory, performance, scheduling police, service on-demand

1. Introduction

Cloud computing is one of the emerging networking technology that has attracted attention from both industrial and research communities. In general cloud computing consist from a lot of parallel and distributed computing systems which compute, storage and provide recourses or services via the Ethernet, ATM, Fray Relay and other technologies [1-3]. In general case cloud computing offers both software and hardware components that can be virtualized and used "as a service" for small and large-scale data networks. Such world-wide known cloud computing network as Amazon[4], Google [5], App iCloud [6], Microsoft Azure [7] serve thousands of millions customers using different software decisions that located on thousands of servers and computing components. Cloud computing networks have structure that includes many heterogeneous distributed computation and commutation elements. The characteristics and parameters of each computing, communication and management node, communication channels and types of virtualization are influent to the end quality of provided services. An analysis of parameters that influent of cloud computing network effectiveness and quality of services specified in the SLA agreements is important part of future development.

Different type of methods allows to measure performance of cloud computing exist today. There are many methods that use test scenarios for real clouds i.e. benchmarking [8, 9], software and simulation modeling tools for analysis of cloud computing functionality [10, 11] and analytical methods based on mathematical modeling [12,13].

Benchmarking often uses comparison of different cloud platforms and virtualization platforms in order to choose the best and to check the validation. The measurements are carried out as follows: on the computer system is supplied pre-definiteness load, it may be what artificial or real, taken from logs journal.

While the real experiment and simulation modeling have some disadvantage, they are not be able to provide insights on scalability and performance in the case of large scale cloud networks, the analytical methods, in particular queueing theory, give ability modeling and analyze both small and large scale clouds. Moreover, queueing theory gives ability to model workloads in a variety of network traffic scenarios.

Queueing theory is a well established tool of analysis frequently used and well suited to communication networks. Use of such mathematical tools on existing cloud computing solutions could provide insight into how the performance changes as data volumes change [14]. In this way, the development a analytical model based on fundamental mathematical principles of queueing theory give ability to create a universal approach for analysis of cloud networks functionality with different architecture and service platforms.

2. Analysis of Service Models in Cloud Computing

The cloud technology can be represent as a complex of virtualization, distributed, parallel data computation that provide services on-demand. These technologies give ability to obtain the maximal transparent and flexible services and big data recourses to optimize the cost efficient of network infrastructure.

Infrastructure and architecture solutions of cloud computing networks provided by different companies have a number of common characteristics [3]. The common architecture of cloud computing network is depicted on the Figure 1.

http://dx.doi.org/10.21275/v5i6.NOV164242

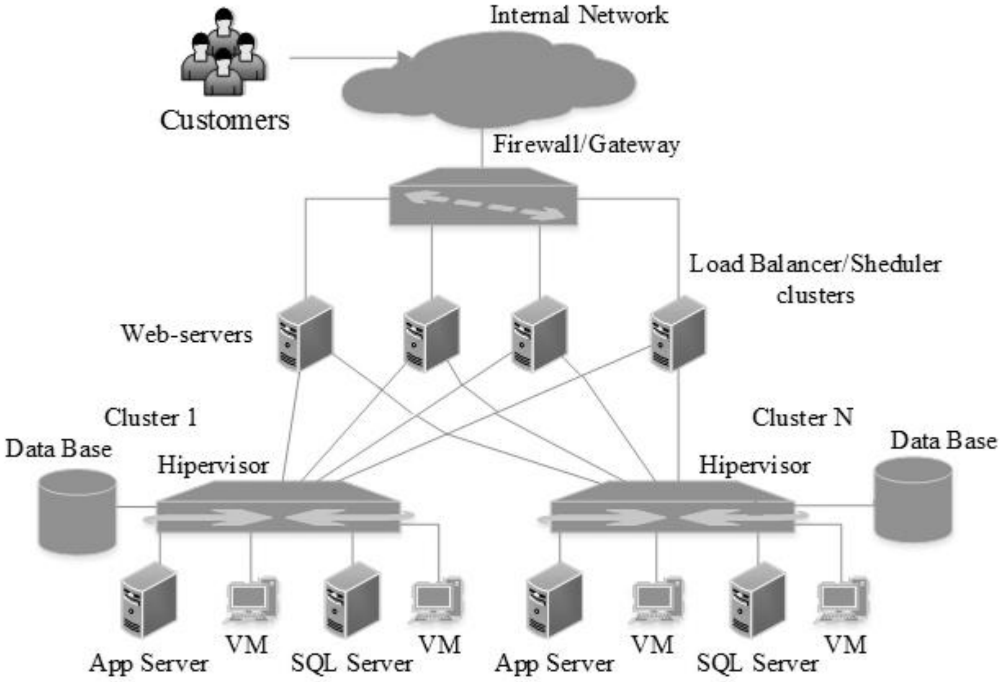


Figure 1: General model of cloud computing network

All solutions include standard servers (Web, SQL, application, mail servers), management nodes that have load balancing and task scheduling mechanisms, communication elements (switch, gateway, firewall etc.), data storage systems that are using horizontal scalability, commutation equipments and virtualization technologies.

In general cloud networks can be divided according to scale, mechanisms of services deliver, recourses storage and secure principles and services provided mechanisms. Three types of cloud infrastructures according to the level of virtualization and service provided mechanisms can be dedicated. This is Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) service delivery models [15].

SaaS is where application services are delivered over the network on a subscription and on-demand basis. The applications are hosted in “the cloud” and can be used for a wide range of tasks for both individuals and organizations.

PaaS consists of run-time environments and software development frameworks and components delivered over the network. PaaS offerings are typically presented as API to consumers.

IaaS is where compute, network, and storage are delivered over the network. As with all cloud computing services it provides access to computing resource in a virtualized environment, “the Cloud”, across a public connection, usually the internet. In the case of IaaS the computing resource provided is specifically that of virtualized hardware, in other words, computing infrastructure.

The division into the service model depends on the part of cloud-based information system is supported by service provider. The network parts that belong to service provider in different cloud infrastructure represent on Figure 2.

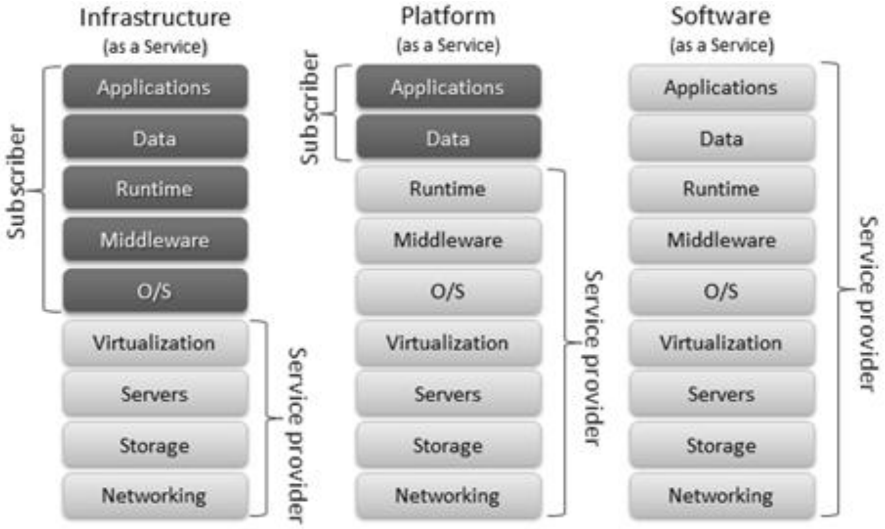


Figure 2: Cloud computing service delivery models

The performance of cloud computing network includes different parameters of system functionality as a high-level response: times, throughput, availability, and low-level: CPU performance, I/O devices and the network []. The choice of parameters depends on the particular research tasks, type of architecture, type and quality of services, etc. For example, provider responsible for all quality of service characteristics in case using SaaS service model. In this model performance evaluations should be done with height accuracy. In case using IaaS other characteristics should be taken into account.

Main parameters that impact on performance of cloud computing network

The high performance is one advantages of the cloud computing what should be satisfactory for each delivered services [3, 8, 10]. Hence, performance evaluation for cloud providers and users is important.

The analysis of research [16, 17] and industrial [18] publication shows that the main criteria are:

- Average response time (R_T). The average response time can be calculated by formula: $R_T = S_T + W_T$, where S_T - service time (data computation process), W_T - waiting time;
- Network capacity of individual connection ($c_{ij}(t)$) can be calculated by formula: $c_{ij}(t) = \frac{C_{ij}}{\sum d_{ij}(t)}$, where C_{ij} is total channel capacity from node i to node j, $d_{ij}(t)$ is current amount of data transfer through the cannel;
- The number of I/O commands per second. This parameter depends of processor core characteristics. I/O commands can represent as: $p_i(t) = \frac{CPU}{CPU_{background} + CPU_{\sum S(t-1)}}$, where $CPU_{background}$ is the CPU that spent on background applications, $CPU_{\sum S(t-1)}$ is the CPU that spent on services of-demand.
- Average waiting time per unit time;
- The number of requests executed per unit time;
- The number of requests per unit time buffer;
- The number of rejected requests per unit time;
- The proposed criteria can be separate on network characteristics and characteristics of computation node.

In this case, the cloud infrastructure model is appropriate to be represented by an analytical model that takes into account the relationship between the computational component (node) and the communication channel. Such representation in its turn, will allow to detect “bottlenecks” in cloud infrastructure.

The queuing model for analysis of cloud computing network

For today a few queuing theory models are proposed in literature as a solution for analysis and evaluation of different cloud computing characteristics. In literature the cloud computing system is modeled as open, closed, and probabilistic queuing models and various amount of nodes with different distribution types *M/M/1/k* [19], *M/H/m/k* [11], *M/G/m* [20]. However, in the proposed solutions analysis and performance evaluation were made within one tier (fragment) of cloud models that significantly reduce evaluation ability. In the way the dynamic distribution of workload can be analyze only on one fragment and don’t allow predict distribution on another layer. In obtained during evaluation results shows only the particular case.

The predominant architecture of cloud computing is a multi-tier architecture, which implies the use of multi-layer interconnection between Web-servers, application server to implement application logic and database servers, computing recourses (VMs). The model of cloud computing with multi-tier applications is shown in Figure 1.

Controller clouds, as the «main entrance» to the cloud, set the queue for each application deployed in it. In addition, the cloud controller reserves the necessary amount of storage and creates the requirement number of virtual machines cloud nodes. Under the node cloud meant a physical server that is running the virtual machine monitor. The initial number of virtual machines specified in the service level agreement (SLA) or determined empirically.

The number of running virtual machines can vary and depends of current load on each node. The virtual machines that depend to one tier coordinated work on one or more nodes of the cloud, each machine has the same amount of physical computing resources. Virtual machines that are running the same application instance form a virtual cluster. Formation of the cluster is independent of the physical location of the virtual machines.

The provision services on-demand can be represent in next steps:

1) Customer generate request.
2) The cloud controller receive request, analyze and prepare scheduling and, according the scheduling rules, forward to the virtual cluster.
3) The request on the cluster can be served both application server and computation node.
4) In the case when customer’s request served on compute node the virtual machine will be deploying.
5) The reply for customer will be send after all computing recourses are successfully created.

The sequences of customers’ requests create their individual session. The processing time of various requests in the session may vary significantly, so they are not homogeneous. The inputs flow of customer’s request can be represent by exponential distribution M with variation coefficient v=1. But the request execution on communication and computing nodes has different nature. The outcomes flow proposed to represent by general G distribution low. The each single tier of cloud computing model is represented as *M/G/1* queering model without any specific scheduling policy, the elastic flows are assumed to divide the bottleneck link bandwidth evenly. In this way the virtual cluster can be modeled by *M/G/k* with specific scheduling policy - *M/G/k/PS*.

The scheduling policy is chosen according to the type of response and type of delivered services (voice, video, data). We assume that the request flows and replies will divide the bottleneck link bandwidth evenly. The policy serves discipline on each single tier modeled by type FCFS (first come -first serve). In MLPS, jobs are served with a discipline that will depend on their attained amount of service. Processor-Sharing (PS) or Foreground Background (FB) contain served processes with smallest attained service is served first. Average response time is performers metric for *M/G/k/FCFS (MPLS, PS, FB)* model [21].

The interaction between tiers is organized by feedback link. The feedback flows have a different intensity μ [22]. The queuing model that takes into account multi-tier architecture of cloud computing network is depicted on Figure 3.

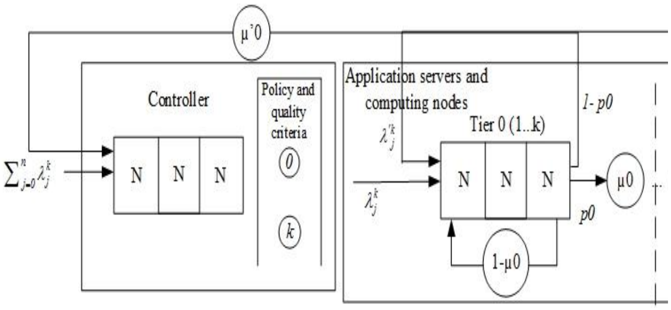

μ'0
Controller
Policy and quality criteria
0
k
$\sum_{j-0}^{n}\lambda_j^k$
N
N
N
Application servers and computing nodes
Tier 0 (1..k)
$\lambda_j^{'k}$
λ_j^k
1- p0
p0
μ0
1-μ0

Figure 3: *M/G/k/PS* queuing model for cloud computing networks

Initial arrival intensity of requests from each customers to controller is matched as λ_j^k the total intensity of request can be characterized as $\sum_{j=0}^{n}\lambda_j^k$. Service rates of each node is matched as μ_j^d , where the d – id the node number. This intensity depends of nodes characteristic. The μ_j^d can be divided into two parts: $\mu^i{}_d = \mu^{ip}{}_d + \mu^{i(1-p)}{}_d$, where the p is probability that request (traffic) passing from node *i* to node *j*, *(1*-p) is the probability of feedback, q - arbiter criterion indicated the total cluster load, *k* – variable determined the type of response or service on-demand, N – total number of customer' requests.

We identified three time metrics to evaluate cloud computing performance:

- Conditional mean delay;
- Average waiting time;
- Average response time.

Chosen parameters depend of scheduling policy and amount of tiers. Conditional mean delay for proposed model can be calculated as:

$$D[T^{PS(\lambda_i)}(i)] = \begin{cases} N\dfrac{\lambda^k{}_i}{\mu_i{}^k(1-p)} \\ E[T^{PS}(\lambda_{i-1})] + N\dfrac{\lambda^k{}_i}{\mu_i{}^k p} \end{cases} \qquad (1)$$

The average waiting time can be calculated as:

$$W(T^{FCFS(\lambda)}(\mu))_{ij} = \begin{cases} \dfrac{\dfrac{(\lambda_i{}^k)^2}{\mu_{ij}}}{\eta - \dfrac{\lambda_i{}^k}{\mu_{ij}}(1-p)}; \\ \dfrac{(1-\eta)^2\mu_{ij}^2\dfrac{(\lambda^k)^2}{\mu_{ij}}}{\eta\mu_0^2(\eta\mu_0-(1-\eta)\mu_{ij}\dfrac{\lambda^k}{\mu})} \end{cases}, \qquad (2)$$

where η is variation of performance between clusters for different service type (*k*), $\eta = \sum_{i=1}^{k} S^k{}_i - \dfrac{\sum_{i=1}^{k} S^k{}_{i-1}}{2c}$.

The average value of the time interval between successive requests can be calculated by the formula:

$$\bar{\tau} = \frac{1}{n}\sum_{i=0}^{n}(t_{i+1}-t_i), \qquad (3)$$

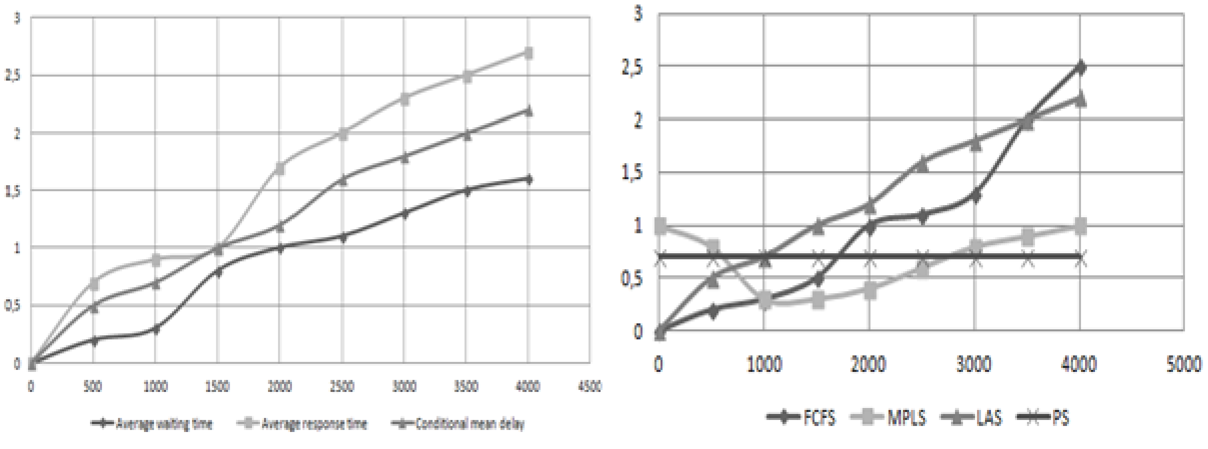
where t_i is timestamp of *i*-th request arriving; *n* is number of time intervals between the analyzed request $\overline{1,i}$.

The average response time for proposed model:

$$R[T^i] = \sum T_{ij} + \sum_{i=1}^{c}\frac{\lambda_i^k}{N\sum\lambda}\frac{W(T)^k{}_i}{S_i - q_i}, \qquad (4)$$

where c - amount of virtual cluster in network, S_i - amount of virtual machines in the cluster, q_i- nominal load of cluster c, $W(T)^k$ - average waiting time for different services (*k*), λ_i^k - responses intensively calculated for different services (*k*).

Matlab package was used for obtaining experimental of result for queuing model types: *M/G/k/FCFS, M/G/k/MPLS, M/G/k/ PS, M/G/k/FB.* The obtained results is show on Figure 4.



a) b)

Figure 4: Experimental results. Average waiting time, response time and delay dependences on amount of customers (a). Characteristics of different scheduling policies *M/G/k/PS* (b)

Then, select a suitable queueing model from the list provided, use the current parameter values, e.g. arrival rate and service rate µ, and show the network performance as computed by the queueing model.

The obtained results shows that a response time increases abruptly when the number of queries greater than 2000. The delay time in waiting time (for discipline) increases smoothly (Figure 5a). average waiting time are depend of scheduling discipline, as shown on Figure 5b, for non-critical time application FCFS is the betters solution (in this variant the biggest response time), for critical time application – MPLS (Multilevel Processor Sharing).

3. Conclusion

Analysis of functionality and performance of cloud computing network give ability to find limitations of scale and bottlenecks in existing solutions. While such methods as real-time analysis and imitation modeling have disadvantages (bad scalability, restricted amount of tested scenarios, deployment complexity) analytical methods allow determining wider specter of problems in network functionality.

Queueing theory has a well established record for successfully modelling networks and computing systems. The proposed queueing model bases on M/G/k/PS queue, it give ability to analyze interaction between tiers and take into account different scheduling policies.

The equations that represent time metrics for cloud computing performance evaluation (conditional mean delay, average waiting time, average response time) proposed in the paper. The average response time equation can be used for evaluating performance of cloud infrastructure with SaaS and IaaS service model, taking into account the multi-tier applications and different class of requests and scheduling policy *M/G/k/FCFS*, *M/G/k/MPLS*, *M/G/k/ PS*, *M/G/k/FB.*

References

[1] Buyya I. R., Yeo C. and other. “Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the fifth Utility,” Future Generation Computer Systems, vol. 25, no. 6, pp. 599–616, 2009.

[2] “Twenty Experts Define Cloud Computing”, SYS-CON Media Inc, http://cloudcomputing.sys-con.com/read/612375_p.htm, 2008.

[3] Mohammed, J. Altmann and J. Hwang, “Cloud Computing Value Chains: Understanding Business and Value Creation in the Cloud,” In: D. Neumann, M. Baker, J. Altmann and O. Rana, Eds., Economic Models and Algorithms for Distributed Systems, Birkhäuser, Basel, 2010, pp. 187-208.

[4] Amazon Elastic Compute Cloud (Amazon EC2)[electronic resourse]. Available at http://aws.amazon.com/ec2 , 30 may 2016.

[5] Newson P. Google Cloud Storage Nearline [electronic resource] available at: https://cloud.google.com/files/GoogleCloudStorageNearline.pdf, 30 may 2015

[6] iCloud – iCloud Drive [electronic resource] available at: http://www.apple.com/ru/icloud/icloud-drive/

[7] P. L. Aidan Finn, Hans Vredevoort and D. Flynn, Microsoft Private Cloud Computing. Wiley Publishing, Inc, 2012, ch.3, 5, pp. 89–116.

[8] N. Handigol, B. Heller, V. Jeyakumar, B. Lantz, and N. McKeown, “Reproducible network experiments using container-based emulation,” in Proc of the 8th International Conference on Emerging Networking Experiments and Technologies (CoNEXT), Nice, France, 10-13 Dec 2012, pp. 253–264.

[9] Q. Xu, S. Mehrotra, Z. Mao, and J. Li, “PROTEUS: Network Performance Forecast for Real-time, Interactive Mobile Applications,” in Proc. of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys), Taipei, Taiwan, 25-28 Jun 2013, pp. 347–360.

[10] Rajkumar Buyya, Rajiv Ranjan and Rodrigo N. Calheiros1,"Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities", Proc. of International Conference on High Performance Computing & Simulation, 2009, June 2009. pp. 1-11.

[11] Rodrigo N. Calheiros, Rajiv Ranjan, and Rajkumar Buyya," Virtual Machine Provisioning Based on Analytical Performance and QoS in Cloud Computing

Environments", Proc. International Conference on Parallel Processing (ICPP), Sept 2011, pp. 295_304.

[12] Hamzeh Khazaei,. Jelena Misic and Vojislav B Misic, "Modelling of Cloud Computing Centers Using M/G/m Queues", Proc. 2011 31st International Conference on Distributed Computing Workshops, pp.87-92.

[13] Roy, K. Yocum, and A. C. Snoeren, “Challenges in the Emulation of Large Scale Software Defined Networks,” in Proc. of the 4th Asia-Pacific Workshop on Systems (APSys), Singapore, 29-30 July 2013.

[14] Jordan Ansell, Winston K.G. Seah, Bryan Ng and Stuart Marshall. Making Queueing Theory More Palatable to SDN/OpenFlow-based Network Practitioners, Wirgin, 2016 – 6 p.

[15] Cloud computing services models: IaaS, SaaS, PaaS [electronic resource] available at: http://www.interoute.com/, 31 may, 2016.

[16] R.Kanniga Devi , S.Sujan" A Survey on Application of Cloudsim Toolkit in Cloud Computing" in International Journal of Innovative Research in Science, Engineering and Technology Vol. 3, Issue 6, June 2014

[17] M. Kesavan, et al., "Practical Compute Capacity Management for Virtualized Datacenters," IEEE Transactions on Cloud Computing, vol.1, no.1, pp. 88-100, 2013.

[18] B. Sharma, et al., "Modeling and synthesizing task placement constraints in Google compute clusters," in Proc. ACM Symp. on Cloud Computing, pp. 1-14, 2011.

[19] Updating data centers. electronic resource] available at: intel.com/content/www/us/en/data-center-efficiency/intel-it-data-center-efficiency-upgrading-datacenter-network-architecture-to-10-gigabit-ethernet-practices.html 31 may, 2016.

[20] Wendy Ellens, Miroslav Zivkovi’, Jacob Akkerboom, Remco Litjens, Hans van den Berg. Performance of Cloud Computing Centers with Multiple Priority Classes. In proc. of 2012 IEEE Fifth International Conference on Cloud Computing, pp. 245_252.

[21] L. Kleinrock, Queueing Systems, vol. 2, John Wiley and Sons, 1976.

[22] L. Kleinrock, R.R. Muntz, and E. Rodemich, “The processor sharing queueing model for time-shared systems with bulk arrivals,,” Networks Journal, , no. 1, pp. 1–13, 1971.