SOME PRACTICAL ASPECTS IN MODELING SERVER CLUSTERS WITH BLOCKS OF RESERVE SERVERS

Lam Van Thanh Binh
Budapest University of Technology and Economics, Hungary

Abstract: Calculating the energy consumption plays an important role in planning as well as in managing the server clusters. It attracted many researchers in proposing the approaches to effectively reduce the energy consumption costs. However, some of these researchers have omitted several practical aspects with regard to the operation of servers when modeling their models. As a consequence, they may not have adequate facts for calculating the amount of electrical energy consumed by servers. This paper has investigated the impact of some practical aspects in the operation of servers on calculating the time that the servers consume energy. Three different aspects are considered: S1) the avoidance of turning off the server during the setup process, S2) the separate switching times required for each server, and S3) the necessity of allowing for the shutdown time of each server. The results show that, the assumption for turning off the server during its setup process should be carefully applied. Additionally, we might omit the shutdown time while still keeping the accuracy of models.

Keywords: Server cluster, practical aspects, setup time, shutdown time.

I. INTRODUCTION

Energy consumption accounts for a significant portion of a server cluster's budget. Therefore, the optimization of energy consumption has become a key focus point in the design and management of server clusters. This has resulted in much interest being generated over the years in how to most effectively reduce these energy consumption costs. Unfortunately, due to simplification and tractability of several models, many researchers have omitted some practical aspects with regard to the operation of servers. Theoretically, the energy consumption of a server is proportional to the time that it consumes electrical energy at each state. Thus, when researchers disregard these practical aspects, they may not have adequate facts for calculating the amount of electrical energy consumed by servers. This paper investigates the impact of some practical aspects in the operation of servers on calculating the time that servers consume energy. The considered aspects are S1) the avoidance of turning off the server during the setup process, S2) the separate switching times for each server, and S3) the necessity of allowing for the shutdown time of each server.

In [1-3], the authors applied the Dynamic Power Management (DPM) as well as the Dynamic Voltage/Frequency Scaling (DVFS) algorithms to optimize the energy consumption in their model. More precisely, the authors [4] applied DPM to activate/deactivate the servers, and utilized DVFS to choose the execution speed for the activated servers. They proposed a strategy to minimize the energy consumption and the mean response time. To do so, they assumed that the control period is sufficiently long enough to compensate for the energy consumption of the activation and the deactivation energy overhead of servers. However, the proposals based on DPM or DVFS only solve a part of the problem as the server still consumes around 60% of the peak power in its idle state [5]. Additionally, DPM should be applied carefully in order to reduce the energy consumption [6].

In [7, 8], the authors applied operational policies for turning on/off servers to control their energy consumption. Two thresholds which dynamically drive the number of idle servers were proposed [9]. Several analytical models [10-12] were introduced which control the number of idle servers to reduce the energy consumption. However, some practical aspects were omitted that would make these models mathematically tractable. More precisely, the authors disregarded the shutdown phase of servers. As a matter of fact, the servers still consume peak power during shutdown periods. Moreover, the number of operative servers is decreased during shutdown periods which directly limits the available servers to serve jobs. Another assumption is that servers can be turned off in the setup process.

The author [13] proposed having a subset of powered up/down servers in a block. They applied two thresholds to power up and power down a block of

Correspondence: Lam Van Thanh Binh,
Email: lvtbinh@hit.bme.hu
Manuscript communication: 3/2018, revised: 5/2018, accepted: 5/2018

servers called reserve servers. They further allowed turning off the servers during the setup process. They assumed that the block of reserve servers could be operative simultaneously after they finished the setup process. When turning off a server, if it is still busy, the serving job will be interrupted. This job will resume if there is any idle server available. Additionally, the server which is not serving any job will be idle. The shutdown time was disregarded in their model. Some of these assumptions are impractical in server operation as well.

Recently, the authors [14] first took into account some practical aspects in server operation under various workloads. However, they considered the accuracy of the models only rather than the operational policy as [13].

Motivated by [13] and [14], we investigate the impact of some practical aspects, i.e. S1, S2, and S3, in operating server clusters with a block of reserve servers. The operative servers are either idle or busy which consume power P_{idle} or P_{max}, respectively. It should be noted that the dynamics of our model is different from other model presented [13] in the sense that we do not turn off the server during setup process, we allow each server its own switching times, and we take into account the shutdown time required by the server to complete its shutdown process.

Specifically, our contributions in this paper include:

- Measuring the difference between the models with and without practical aspects of server operation in terms of the mean response time and the average energy consumption.
- Claiming that the omitted practical aspects in server operation in some models might not lead to a loss of accuracy in such models. However, the accuracy of calculating the amount of consumed energy used by the servers can be improved with knowledge of these aspects

The rest of the paper is organized as follows. Section II presents the abstract model which is considered throughout this paper. The performance measures and the energy consumption metrics are provided in Section III. The simulation results are demonstrated in Section IV. Finally, Section V concludes our work.

II. SYSTEM MODELING

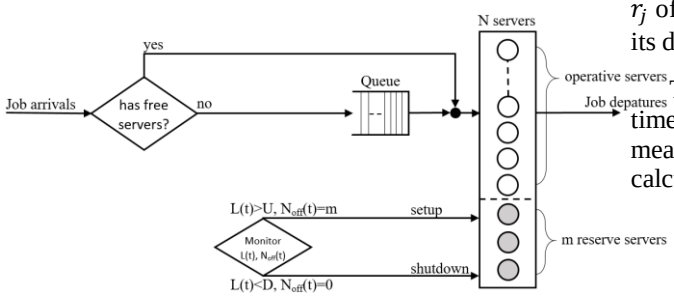

Fig. 1 illustrates the model of a server cluster with thresholds to power up/down the reserve servers (Table I explains the notations). The server cluster consists of N identical servers, of which m are reserved. Initially, the $(N-m)$ operative servers are idle and are ready to serve jobs, the remaining m reserve servers are off. An operative server can serve one job at a time following the FCFS policy. The service time of each server follows an exponential distribution with mean $1/\mu$.

There are two thresholds to power up and power down the reserve servers.

- Threshold U: when the number of jobs in the system at time t, i.e. $L(t)$ exceeds the upper threshold and the reserve servers are off, they will be simultaneously powered up; however, each server has its own setup period which is distributed exponentially with mean $1/\vartheta$.
- Threshold D: when the number of jobs in the system at time t, i.e. $L(t)$ drops below the lower threshold and there are no off servers, there are m servers will be powered down. Each server takes an exponentially distributed shutdown time independently of the others with mean $1/\omega$.

Furthermore, when a server is in setup process, it can not be turned off. The operator must wait until that such server has finished the setup process before it can be powered down if needed.

Jobs arrive at the servers according to a Poisson process with rate $\lambda > 0$. Jobs will be enqueued if there are no available servers to serve them. When the operator chooses m servers to power down, if there are not enough idle servers in the system, he must power down other servers that are occupied. This action results in interrupted jobs moved back to the queue with a priority. When there is an idle server available, the interrupted job will be resumed. Note that the number of servers powered up/down does not exceed m.

III. PERFORMANCE MEASURES AND ENERGY CONSUMPTION METRICS

Following [6, 14], let w_j be the waiting time in the queue of a job j before service and s_j be the service time needed to process a job j. The interrupted time i_j of a job j is the time since that job was interrupted to the time it resumes. Obviously, $i_j = 0$ if a job j was not interrupted, otherwise, $i_j > 0$. The response time r_j of a job j is the time period between its arrival and its departure. Therefore, $r_j = w_j + s_j + i_j$.

The mean waiting time $WT(n)$, the mean service time $ST(n)$, the mean interrupted time $IT(n)$, and the mean response time $RT(n)$ of n completed jobs are calculated as follows:

$$WT(n) = \frac{1}{n}\sum_{j=1}^{n} w_j \quad (1)$$

$$ST(n) = \frac{1}{n}\sum_{j=1}^{n} s_j \quad (2)$$

$$IT(n) = \frac{1}{n}\sum_{j=1}^{n} i_j \quad (3)$$

$$RT(n) = \frac{1}{n}\sum_{j=1}^{n} r_j \quad (4)$$

The long term average waiting time, the long term average service time, the long term average interrupted time, and the long term average response time are defined as

$$WT = \lim_{n\to\infty} WT(n) \quad (5)$$

$$ST = \lim_{n\to\infty} ST(n) \quad (6)$$

$$IT = \lim_{n\to\infty} IT(n) \quad (7)$$

$$RT = \lim_{n\to\infty} RT(n) \quad (8)$$

Servers consume peak energy in serving job, in setup process, and in shutdown process. When the servers are busy, the consumed energy is used to process jobs. In the setup and the shutdown periods, consumed energy is needed because of the natural dynamics of the servers, yet it can not be used for processing jobs, therefore, it should be minimized.

Let $\tau_k(t)$, $\sigma_k(t)$, $\zeta_k(t)$ and $\kappa_k(t)$ denote the time a server k spent in the busy, idle, setup and shutdown periods within the time interval t, respectively. Let the departure time of the job n^{th}, which is completed, be t_n. Define $AE_u(t_n)$ as the mean useful energy consumption per job which a server consumes to process a job up to the time t_n, $AE_d(t_n)$ as the mean idle energy consumption per job which a server consumes at idle state up to the time t_n, and $AE_w(t_n)$ as the mean switching energy consumption per job which a server consumes during either setup or shutdown periods up to the time t_n. We have,

$$AE_u(t_n) = \frac{P_{max}\sum_{k=1}^{N}\tau_k(t_n)}{n} \quad (9)$$

$$AE_d(t_n) = \frac{P_{idle}\sum_{k=1}^{N}\sigma_k(t_n)}{n} \quad (10)$$

$$AE_w(t_n) = \frac{P_{max}\sum_{k=1}^{N}(\zeta_k(t_n)+\kappa_k(t_n))}{n} \quad (11)$$

where P_{max} and P_{idle} are the power consumption of each server at busy/setup/shutdown and idle states, respectively. Therefore, the mean energy consumption per job up to t_n is calculated as

$$AE(t_n) = AE_u(t_n) + AE_d(t_n) + AE_w(t_n) \quad (12)$$

The long term average energy consumption per job are defined as

$$AE_u = \lim_{n\to\infty} AE_u(t_n) \quad (13)$$

$$AE_d = \lim_{n\to\infty} AE_d(t_n) \quad (14)$$

$$AE_w = \lim_{n\to\infty} AE_w(t_n) \quad (15)$$

$$AE = \lim_{n\to\infty} AE(t_n) \quad (16)$$

Table I shows the parameters, the performance measures, and the energy metrics which were used in the simulation.

Table I. Notations

N	number of servers in the system
m	number of reserve servers
$N_{off}(t)$	number of servers which is off at time t
$L(t)$	number of jobs in the system at time t
U	upper threshold of the number of jobs in the system to power up servers
D	lower threshold of the number of jobs in the system to power down servers
P_{max}	average active power of server
P_{idle}	average idle power of server
WT	average waiting time
ST	average service time
IT	average interrupted time
RT	average responese time
AE_u	average useful energy consumption per job
AE_d	average idle energy consumption per job
AE_w	average switching energy consumption per job
AE	average energy consumption per job

IV. SIMULATION EXPERIMENTS

We built a software to simulate the abstract model. Simulation runs were performed with the confidence level of 99% and the accuracy (i.e. the ratio of the half-width of the confidence interval and the mean of collected data) is less than 10^{-4}.

Throughout the simulation, the number of servers is $N = 20$ the mean service time is $1/\mu = 1s$, the mean setup time is ten times longer than the mean service time, i.e. $1/\vartheta = 10s$, the mean shutdown time (if any) is $1/\omega = 2s$ which is a fifth of the mean setup time. We chose the reference server [15] which consumes average active power and average idle power of $P_{max} = 56.1W$ and $P_{idle} = 13.1W$, respectively.

We compared 4 following models:

Table II. Models in simulation

Model	Disallow turning	Disallow finishing	Server has

	off a server in setup process (S1)	switching periods of servers simultaneously (S2)	shutdown time (S3)
M003			x
M000			
B123	x	x	x
B120	x	x	

Model **M000** has the same dynamics as [13]. The shutdown time (S3) is added into model **M003**. We applied all three practical aspects in model **B123**. The model **B120** does not take the shutdown time into consideration.

Fig. 2 shows the mean response time for all models versus threshold U. Apparently, there is asymptotic behavior of the mean response time between the pair of models **M00x** as well as that of models **B12x** against threshold U. It suggests that we might omit the aspect S3 in modeling the server clusters while still maintaining the accuracy of models in terms of the mean response time. However, these pairs of models have the reversed trends. At the small values of threshold U, the models with aspects S1 and S2 (**B12x**) give a higher mean response time than the models without them (**M00x**). Yet for the larger values of threshold U, this order is reversed. Interestingly, at thresholds $U = 30$ and $U = 35$, four models show almost the same mean response time. The reason for this phenomenon is that at these values of U, the difference between U and D is around 20 which is equal to the number of servers in the system ($N = 20$). Therefore, the rate of switching reserve servers is medium. This implies that after the system powers up the reserve servers, it runs with the maximum number of operative servers for a long duration before powers down them. In other words, there are more available servers to serve jobs. Hence, the jobs incur shorter waiting times.

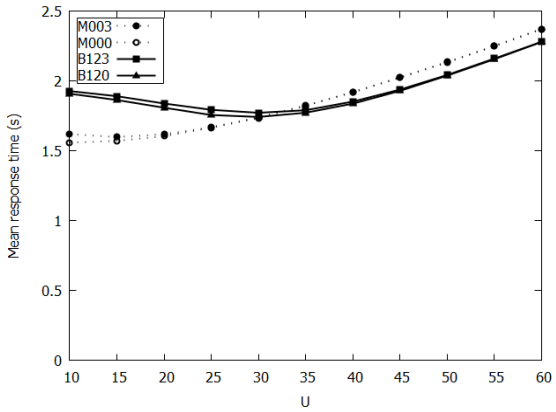

Figure 2. Mean response time versus up thresholds. ($N = 20$; $\rho = 0.6$; $1/\mu = 1s$; $1/\vartheta = 10s$; $1/\omega = 2s$; $m = 8$; $D = 11$)

Fig. 3 shows the mean response time of the job at different workloads. As expected, the figure reveals that the mean response time is asymptotic to the mean service time at $\rho < 0.5$, whereas it becomes larger as ρ approaches to 1. For the workload greater than 0.5, the significant differences between the models **M00x** and models **B12x** increase gradually. The explanation for these differences is as follows. When the arrival rate is high enough, the number of jobs in the waiting queue is high which causes accumulative long waiting time for the jobs in all models. Additionally, in the models with aspects S1 and S2, if the number of jobs in the system drops below threshold D during the setup processes, the reserve servers do not allow being turned off. In other words, the reserve servers have to wait until finishing their setup processes. During these periods, due to the high arrival rate of jobs, there might have many jobs accumulated in the waiting queue which results in long waiting times. Whereas, this phenomenon does not occur in the models without aspects S1 and S2 due to the reserve servers are turned off immediately during setup process. Hence, the mean response time of jobs in the models with aspects S1 and S2 is longer than the mean response time of jobs in the models without them.

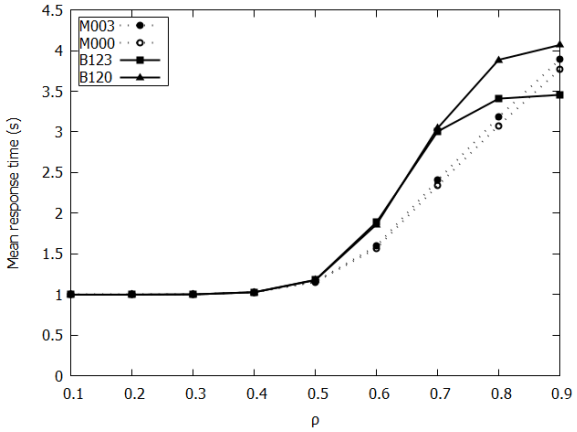

Figure 3. Mean response time versus workload. ($N = 20$; $1/\mu = 1s$; $1/\vartheta = 10s$; $1/\omega = 2s$; $m = 8$; $D = 11$; $U = 15$)

Fig. 4 illustrates the mean response time of job in the case of the mean service time is large ($1/\mu = 50s; 300s$) compared to the mean setup time ($1/\vartheta = 10s$) and the mean shutdown time ($1/\omega = 2s$). It is interesting to observe that, when the service time is large enough, the impact of these practical aspects is insignificant.

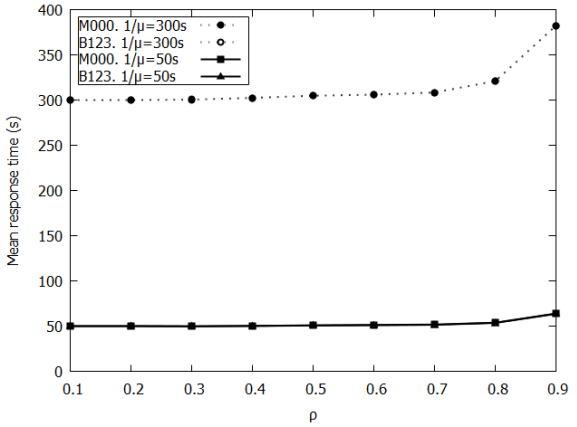

Figure 4. Mean response time versus workload. ($N = 20; {}^{1}/_{\mu} = 50s, 300s;\ {}^{1}/_{\vartheta} = 10s;\ {}^{1}/_{\omega} = 2s; m = 8; D = 11; U = 15$)

Fig. 5 presents the average energy consumption for each job. The general trends are that the average energy consumption per job decreases with the values of threshold U . When U is large, the aspects S1 and S2 might be omitted as well as S3 while maintaining the accuracy of the model. Conversely, when U is small, there is a trade-off between the mean response time of the job (Fig. 2) and the average energy consumption per job (Fig. 5). Thus, we should omit the practical aspects with care in the case of the threshold U is small compared to the threshold D as well as the number of servers in the system N.

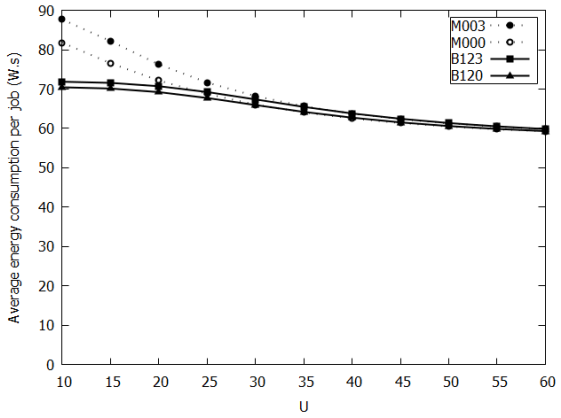

Figure 5. Average energy consumption per job versus up thresholds. ($N = 20;\ \rho = 0.6;\ {}^{1}/_{\mu} = 1s;\ {}^{1}/_{\vartheta} = 10s;\ {}^{1}/_{\omega} = 2s; m = 8; D = 11$)

Fig. 6 shows the same trends with Fig. 5. It is worth noting that the average idle energy consumption in this scenario is very small, so the average switching energy consumption contributes a significant portion to the average energy consumption. At small values of U, the models with aspects S1 and S2 consumes less energy than those without that of aspects. It is trivial to explain the decreasing average energy consumption at the high values of U while the value of D remains constant. Take threshold $U = 30$ as an example, at first state, the model runs with ($N - m = 20 - 8 = 12$) operative servers. The condition to power up $m = 8$ reserve servers is that the number of jobs in the system exceeds 30. While at threshold $U = 60$, to power up the reserve servers, the number of jobs in the system must be at least 61 which is two times larger compared to the former case. Hence, the rate of switching reverse servers decreases with threshold U.

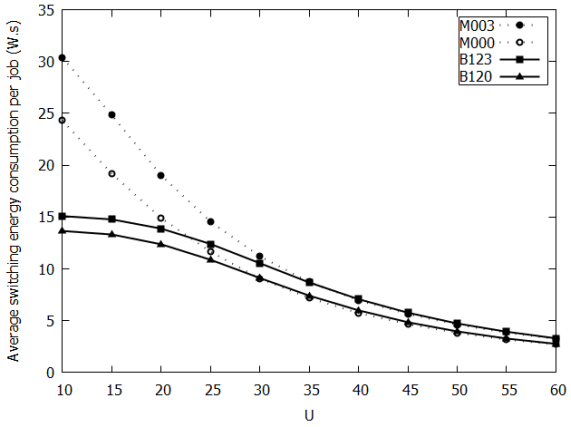

Figure 6. Average switching energy consumption per job versus up thresholds. ($N = 20;\ \rho = 0.6;\ {}^{1}/_{\mu} = 1s;\ {}^{1}/_{\vartheta} = 10s;\ {}^{1}/_{\omega} = 2s; m = 8; D = 11$)

V. CONCLUSION

This paper has investigated the performance measures in terms of the mean response time and the average energy consumption per job in models with and without practical aspects, i.e. not turning off the servers during setup, allowing each server its own switching times, and requiring shutdown time when powering down the server. The results show that:

- The assumption for turning off the server during its setup process should be carefully applied in modeling the server cluster. This aspect shows the considerable impact of energy consumption on the operative servers and affects the length of time that energy is consumed.
- The shutdown time might be omitted while still keeping the accuracy of models.
- The models with/without these practical aspects do not show significant differences when the mean service time is large enough compared to the mean switching times.

The results in this paper might be useful to operators who desire to evaluate the performance and to calculate the energy consumption of their server clusters. In the future, we would consider the practical aspects in terms of migration costs when transferring the interrupted jobs to another server which were not considered in [13].

REFERENCES

[1] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, and N. Gautam, "Managing server energy and operational costs in hosting centers," in *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '05. New York, NY, USA: ACM, 2005, pp. 303–314. [Online]. Available: http://doi.acm.org/10.1145/1064212.1064253.

[2] Y. Wang, X. Wang, M. Chen, and X. Zhu, "Power-efficient response time guarantees for virtualized enterprise servers," in *2008 Real-Time Systems Symposium*, Nov 2008, pp. 303–312.

[3] V. Petrucci, E. V. Carrera, O. Loques, J. C. B. Leite, and D. Moss, "Optimized management of power and

performance *for virtualized heterogeneous server clusters," in 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, May 2011, pp. 23–32* .

[4] *S. Wang, W. Munawar, J. Liu, J. J. Chen, and X. Liu, "Powersaving design for server farms with response time percentile guarantees," in 2012 IEEE 18th Real Time and Embedded Technology and Applications Symposium, April 2012, pp. 273–284.*

[5] *L. A. Barroso and U. Hlzle, "The case for energy-proportional* computing," *Computer*, vol. 40, no. 12, pp. 33–37, Dec 2007.

[6] T. V. Do, B. T. Vu, X. T. Tran, and A. P. Nguyen, "A generalized model for investigating scheduling schemes in computational clusters," *Simulation Modelling Practice and Theory*, vol. 37, pp. 30 – 42, 2013.

[7] A. Gandhi, V. Gupta, M. Harchol-Balter, and M. A. Kozuch, "Optimality analysis of energy-performance trade-off for server farm management," *Perform. Eval.*, vol. 67, no. 11, pp. 1155–1171, Nov. 2010. [Online]: http://dx.doi.org/10.1016/j.peva.2010.08.009.

[8] B. Guenter, N. Jain, and C. Williams, "Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning," in *2011 Proceedings IEEE INFOCOM*, April 2011, pp. 1332–1340.

[9] A. Gandhi, M. Harchol-Balter, and I. Adan, "Server farms with setup costs," *Performance Evaluation*, vol. 67, no. 11, pp. 1123 – 1138, 2010, performance 2010.

[10] J. R. Artalejo, A. Economou, and M. J. Lopez-Herrero, "Analysis of a multiserver queue with setup times," *Queueing Systems*, vol. 51, no. 1, pp. 53–76, Oct 2005. [Online]. Available: https://doi.org/10.1007/s11134-005-1740-6.

[11] T. Do, "Comparison of allocation schemes for virtual machines in energy-aware server farms," vol. 54, pp. 1790–1797, 02 2011.

[12] T. V. Do and C. Rotter, "Comparison of scheduling schemes for on-demand iaas requests," *Journal of Systems and Software*, vol. 85, no. 6, pp. 1400 – 1408, 2012, special Issue: Agile Development.

[13] I. Mitrani, "Managing performance and power consumption in a server farm," *Annals of Operations Research*, vol. 202, no. 1, pp. 121–134, Jan 2013. [Online]. Available: https://doi.org/10.1007/s10479-011-0932-1.

[14] N. H. Do and T.-B. V. Lam, "Some practical aspects on modelling server clusters," in *Advanced Computational Methods for Knowledge Engineering*, H. A. Le Thi, N. T. Nguyen, and T. V. Do, Eds. Cham: Springer International Publishing, 2015, pp. 403–414.

[15] Standard Performance Evaluation Corporation (SPEC), "Standard performance evaluation corporation (spec), fujitsu server primergy rx1330 m3," May 2017. [Online] https://www.spec.org/power_ssj2008/results/res2017q2/power_ssj2008-20170315-00744.html.

Lam Van Thanh Binh, received the B.S degree in Telecommunication from the Nha Trang Telecommunication University, Vietnam in 2006. He graduated M.A degree in Electronics Engineering from the Posts and Telecommunications Institute of Technology, Vietnam in 2011. He is currently pursuing PhD degree in Electrical Engineering at the Budapest University of Technology and Economics, Hungary. His interested fields are: telecommunication, queueing theory, cloud computing, performance evaluations.