# Supplementary Material for Quantitative Modeling and Analytical Calculation of Elasticity in Cloud Computing

Keqin Li, *Fellow, IEEE*

## 1 Related Research

### 1.1 Modeling Cloud Platforms

A cloud platform essentially provides services to users, and is naturally modeled and treated as a queueing system. In [9], the authors investigated the problem of optimal multiserver configuration for profit maximization in a cloud computing environment by using an M/M/m queuing model. The study was further extended in [31]. In [10], the authors addressed the problem of optimal power allocation and load distribution for multiple heterogeneous multicore server processors across clouds and data centers, by modeling a multicore server processor as a queuing system with multiple servers. In [29], the authors focused on strategy configurations of multiple users to make cloud service reservation from a game theoretic perspective, and formulated the problem as a non-cooperative game among the multiple cloud users.

Another related analytical tool is the continuous-time Markov chain (CTMC) model, which has also been extensively used to study various properties of cloud computing systems. In [17], the authors quantified the power performance trade-offs by developing a scalable analytic model based on CTMC for joint analysis of performance and power consumption on a class of Infrastructure-as-a-Service (IaaS) clouds. In [25], the authors proposed an analytical performance model based on CTMC, which incorporates several important aspects of cloud centers to obtain not only detailed assessment of cloud center performance, but also clear insights into equilibrium arrangement and capacity planning that allow service delay, task rejection probability, and power consumption to be kept under control. In [33], the authors investigated the Markovian Arrival Processes (MAP) and the related MAP/MAP/1 queueing model to predict the performance of servers deployed in the cloud.

### 1.2 Assessing Elastic System Performance

Some efforts have been made for modeling the performance of clouds with elastic scaling strategies. In [21], the authors presented generic cloud performance models for evaluating Iaas, PaaS, SaaS, and mashup or hybrid clouds. Some real-life benchmark experiments were conducted mainly on IaaS cloud platforms over scale-out and scale-up workloads (see Section 3.2 for definitions). Cloud benchmarking results were analyzed with the efficiency, elasticity, QoS, productivity, and scalability (see Section 5 for definitions of these notions) of cloud performance. It was found that to satisfy production services, the choice of scale-up or scale-out solutions should be made primarily by the workload patterns and resources utilization rates required. Scaling-out machine instances have much lower overhead than those experienced in scale-up experiments.

• *K. Li is with the Department of Computer Science, State University of New York, New Paltz, New York 12561, USA.*
*E-mail: lik@newpaltz.edu*

*Manuscript received Month Day, 2016; revised Month Day, 2016.*

## References

[1] M. Aazam and E.-N. Huh, “Cloud broker service-oriented resource management model,” *Transactions on Emerging Telecommunications Technologies*, 2015.

[2] M. Aazam, E.-N. Huh, M. St-Hilaire, C.-H. Lung, and I. Lambadaris, “Cloud customer’s historical record based resource pricing,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 7, pp. 1929-1940, 2016.

[3] D. Ardagna, G. Casale, M. Ciavotta, J. F. Pérez, and W. Wang, “Quality-of-service in cloud computing: modeling techniques and their applications,” *Journal of Internet Services and Applications*, vol. 5, no. 11, 2014.

[4] J. R. Artalejo, D. S. Orlovsky, and A. N. Dudin, “Multi-server retrial model with variable number of active servers,” *Computers and Industrial Engineering*, vol. 48, no. 2, pp. 273-288, 2005.

[5] L. Badger, T. Grance, R. Patt-Corner, and J. Voas, “Cloud computing synopsis and recommendations,” Special Publication 800-146, National Institute of Standards and Technology, U.S. Department of Commerce, 5/29/2012.

[6] A. K. Bardsiri and S. M. Hashemi, “QoS metrics for cloud computing services evaluation,” *I.J. Intelligent Systems and Applications*, vol. 12, pp. 27-33, 2014.

[7] R. Buyya, J. Broberg, and A. Goscinski, eds., *Cloud Computing Principles and Paradigms*, John Wiley & Sons, Hoboken, New Jersey, 2011.

[8] R. N. Calheiros, C. Vecchiola, D. Karunamoorthy, R. Buyya, “The Aneka platform and QoS-driven resource provisioning for elastic applications on hybrid clouds,” *Future Generation Computer Systems*, vol. 28, pp. 861-870, 2012.

[9] J. Cao, K. Hwang, K. Li, and A. Zomaya, “Optimal multiserver configuration for profit maximization in cloud computing,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1087-1096, 2013.

[10] J. Cao, K. Li, and I. Stojmenovic, “Optimal power allocation and load distribution for multiple heterogeneous multicore server processors across clouds and data centers,” *IEEE Transactions on Computers*, vol. 63, no. 1, pp. 45-58, 2014.

[11] W. Dawoud, I. Takouna, and C. Meinel, "Elastic VM for cloud resources provisioning optimization," *Advances in Computing and Communications*, Communications in Computer and Information Science, vol. 190, pp. 431-445, 2011.

[12] S. Dustdar, Y. Guo, B. Satzger, and H.-L. Truong," "Principles of elastic processes," *IEEE Internet Computing*, vol. 15, no. 5, pp. 66-71, 2011.

[13] J. O. Fitó, I. Goiri, and J. Guitart, "SLA-driven elastic cloud hosting provider," *16th Euromicro Conference on Parallel, Distributed and Network-Based Processing*, pp. 111-118, 2010.

[14] G. Galante and L. C. E. de Bona, "A survey on cloud computing elasticity," *IEEE/ACM Fifth International Conference on Utility and Cloud Computing*, pp. 263-270, 2012.

[15] R. Ghosh, D. Kim, K. S. Trivedi, "System resiliency quantification using non-state-space and state-space analytic models," *Reliability Engineering and System Safety*, vol. 116, pp. 109-125, 2013.

[16] R. Ghosh, F. Longoy, V. K. Naikz, and K. S. Trivedi, "Quantifying resiliency of IaaS cloud," *29th IEEE International Symposium on Reliable Distributed Systems*, pp. 343-347, 2010.

[17] R. Ghosh, V. K. Naik, and K. S. Trivedi, "Power-performance trade-offs in IaaS cloud: A scalable analytic approach," *IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops*, pp. 152-157 , 2011.

[18] Z. Gong, X. Gu, and J. Wilkes, "PRESS: predictive elastic resource scaling for cloud systems," *International Conference on Network and Service Management*, pp. 9-16, 2010.

[19] N. R. Herbst, *Quantifying the Impact of Platform Configuration Space for Elasticity Benchmarking*, Study Thesis, Karlsruhe Institute of Technology, 2011.

[20] N. R. Herbst, S. Kounev, and R. Reussner, "Elasticity in cloud computing: what it is, and what it is not," *10th International Conference on Autonomic Computing*, pp. 23-27, 2013.

[21] K. Hwang, X. Bai, Y. Shi, M. Li, W.-G. Chen, and Y. Wu, "Cloud performance modeling with benchmark evaluation of elastic scaling strategies," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 1, pp. 130-143, 2016.

[22] S. Islam, K. Lee, A. Fekete, and A. Liu, "How a consumer can measure elasticity for cloud platforms," Technical Report 680, School of Information Technologies, University of Sydney, 2011.

[23] A. I. Ivaneshkin, "Optimizing a multiserver queuing system with a variable number of servers," *Cybernetics and Systems Analysis*, vol. 43, no. 4, pp. 542-548, 2007.

[24] H. Khazaei, J. Mišić, and V. B. Mišić, "A fine-grained performance model of cloud computing centers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 11, pp. 2138-2147, 2013.

[25] H. Khazaei, J. Mišić, V. B. Mišić, and S. Rashwand, "Analysis of a pool management scheme for cloud computing centers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 5, pp. 849-861, 2013.

[26] L. Kleinrock, *Queueing Systems, Volume 1: Theory*, John Wiley and Sons, New York, 1975.

[27] M. Kuperberg, N. Herbst, J. von Kistowski, and R. Reussner, "Defining and quantifying elasticity of resources in cloud computing and scalable platforms," Karlsruhe Reports in Informatics 2011, 16, Karlsruhe Institute of Technology.

[28] K. Li, "Improving multicore server performance and reducing energy consumption by workload dependent dynamic power management," *IEEE Transactions on Cloud Computing*, vol. 4, no. 2, pp. 122-137, 2016.

[29] C. Liu, K. Li, C.-Z. Xu, and K. Li, "Strategy configurations of multiple users competition for cloud service reservation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 2, pp. 508-520, 2016.

[30] M. Mao and M. Humphrey, "A performance study on the VM startup time in the cloud," *IEEE Fifth International Conference on Cloud Computing*, pp. 423-430, 2012.

[31] J. Mei, K. Li, A. Ouyang, and K. Li, "A profit maximization scheme with guaranteed quality of service in cloud computing," *IEEE Transactions on Computers*, vol. 64, no. 11, pp. 3064-3078, 2015.

[32] P. Mell and T. Grance, "The NIST definition of cloud computing," Special Publication 800-145, National Institute of Standards and Technology, U.S. Department of Commerce, September 2011.

[33] S. Pacheco-Sanchez, G. Casale, B. Scotney, S. McClean, G. Parr, and S. Dawson, "Markovian workload characterization for QoS prediction in the cloud," *IEEE Fourth International Conference on Cloud Computing*, pp. 147-154, 2011.

[34] N. Roy, A. Dubey, and A. Gokhale, "Efficient autoscaling in the cloud using predictive models for workload forecasting," *IEEE Fourth International Conference on Cloud Computing*, pp. 500-507, 2011.

[35] U. Sharma, P. Shenoy, S. Sahu, and A. Shaikh, "A cost-aware elasticity provisioning system for the cloud," *31st International Conference on Distributed Computing Systems*, pp. 559-570, 2011.

[36] Z. Shen, S. Subbiah, X. Gu, and J. Wilkes, "CloudScale: elastic resource scaling for multi-tenant cloud systems," *Proceedings of the 2nd ACM Symposium on Cloud Computing*, Article No. 5, 2011.

[37] P. Sobeslavsky, *Elasticity in Cloud Computing*, Master Thesis, Joseph Fourier University, 2011.