Uncertainty in Clouds: Challenges of Efficient Resource Provisioning*

Andrei Tchernykh [1], Uwe Schwiegelsohn [2], Vassil Alexandrov [3], El-ghazali Talbi [4]

[1] CICESE Research Center, Baja California, [2] Technische Universität Dortmund, [3] Barcelona Supercomputing Centre, ICREA-BSC, [4] LIFL, University of Lille 1

We discuss the role of uncertainty in the resource/service provisioning, investment, operational cost, programming models, etc. that have not yet been adequately addressed in the scientific literature.

Clouds differ from previous computing environments in the way that they introduce a continuous uncertainty into the computational process. The uncertainty becomes the main hassle of cloud computing bringing additional challenges to both end-users and resource providers. It requires to waive habitual computing paradigms, adapt current computing models, and design novel resource management strategies to handle uncertainty in an effective way [1].

In spite of extensive research of uncertainty issues in computational biology, decision making in economics, etc. a study of uncertainty for cloud computing is limited. Most of works examine uncertainty phenomena in users' perceptions of the qualities, intentions and actions of providers, privacy, security and availability [2].

We discuss several major sources of uncertainty in clouds: dynamic elasticity, dynamic performance changing, virtualization, loosely coupling application to the infrastructure, among many others. A workload in such an environment is not predictable and can be changed dramatically. It is impossible to get exact knowledge about the system. Parameters such as an effective processor speed, number of available processors, and actual bandwidth are changing over the time. Elastic escalation process has a higher repercussion on the QoS, but adds another factor of uncertainty.

Providers might not know the quantity of data and computation required by users. For example, every time when a user requires a status of his e-mail or bank account, it could generate different amount of data and take different time for delivering. A pool of virtualized, dynamically scalable computing resources, storages, software, and services add a new dimension to the problem. The manner in which the service provisioning can be done depends not only on the service property and needed resources, but also users that share resources at the same time, in contrast to dedicated resources governed by a queuing system [3, 4, 5, 6].

We also discuss a CA-DAG application model for cloud computing applications [7]. This communication-aware model allows making separate resource allocation decisions, assigning processors to handle computing jobs, and network resources for data transmissions. We discuss the benefits, weaknesses, and performance characteristics of such a model and resource allocation strategies in presence of uncertainty due to dynamic behavior of the execution context, job mix workloads, or uncertainty of the workflow properties.

References

1. Tchernykh A., Schwiegelsohn U., Alexandrov V., Talbi E., Towards Understanding Uncertainty in Cloud Computing Resource Provisioning. SPU'2015 - Solving Problems with Uncertainties (3rd Workshop). In conjunction with The 15th International Conference on Computational Science (ICCS 2015), Reykjavík, Iceland, June 1- 3, 2015. Procedia Computer Science, Elsevier, Volume 51, Pages 1772–1781, 2015, DOI: 10.1016/j.procs.2015.05.387

* This work is partially supported by CONACYT (Consejo Nacional de Ciencia y Tecnología, México), grant no. 178415.

.

2. Trenz M., Huntgeburth J.C., Veit D. The Role Of Uncertainty In Cloud Computing Continuance: Antecedents, Mitigators, And Consequences, ECIS, 2013, 147-147.
3. Tchernykh, A., Pecero, J., Barrondo, A., Schaeffer, E.: Adaptive Energy Efficient Scheduling in Peer-to-Peer Desktop Grids, Future Generation Computer Systems, 36:209–220 (2014).
4. Schwiegelshohn, U., Tchernykh, A.: Online Scheduling for Cloud Computing and Different Service Levels, 26th Int. Parallel and Distributed Processing Symposium Los Alamitos, CA, pp. 1067–1074 (2012)
5. Tchernykh, A., Lozano, L., Schwiegelshohn, U., Bouvry,P., Pecero, J., Nesmachnow, S., Drozdov, A. Online Bi-Objective Scheduling for IaaS Clouds with Ensuring Quality of Service. Journal of Grid Computing, Springer-Verlag, DOI 10.1007/s10723-015-9340-0 (2015)
6. Sequencing and Scheduling with Inaccurate Data. Editors: Yuri N. Sotskov and Frank Werner. Nova Science Pub, Applied Statistica Science, 2014, 442pp.
7. Kliazovich D., Pecero J. E., Tchernykh A., Bouvry P., Khan S. U., Zomaya A. Y. “CA-DAG: Modeling Communication-Aware Applications for Scheduling in Cloud Computing,” Journal of Grid Computing, Springer-Verlag, DOI 10.1007/s10723-015-9337-8 (2015)