Experiments with Complex Scientific Applications on Hybrid Cloud Infrastructures

Maciej Malawski[1,2], Piotr Nowakowski[1], Tomasz Gubała[1], Marek Kasztelnik[1], Marian Bubak[1,2],
Rafael Ferreira da Silva[3], Ewa Deelman[3], Jarek Nabrzyski[4]

AGH University of Science and Technology:
[1] ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Kraków, Poland
[2] Department of Computer Science, al. Mickiewicza 30, 30-095 Kraków, Poland
[3] ISI, University of Southern California, CA, USA
[4] Center for Research Computing, University of Notre Dame, IN, USA

emails: {bubak,malawski}@agh.edu.pl,{p.nowakowski,t.gubala,m.kasztelnik}@cyfronet.pl,
{rafsilva,deelman}@isi.edu, naber@nd.edu

1. Introduction

DICE Team at Department of Computer Science and Academic Computer Center CYFRONET of AGH collaborates with researchers at the University of Southern California and the Center for Research Computing at the University of Notre Dame. In the scope of this collaboration, we develop methods and tools supporting programming and execution of complex scientific applications on heterogeneous computing infrastructures. The developed methods and tools include:

- A cloud platform for scientific applications called Atmosphere [1], developed in the scope of the UrbanFlood and VPH-Share EU-funded research projects (http://www.urbanflood.eu and http://www.vph-share.eu). Atmosphere provides a development and execution environment based on state-of-the-art cloud technologies. It includes a resource allocation module, which is responsible for allocating resources from private clouds (e.g. OpenStack, hosted at CYFRONET) and public clouds such as Amazon EC2 basing on complex optimization conditions. Computational software is exposed as a set of services, which can be used either as standalone tools or as building blocks for larger workflows (http://dice.cyfronet.pl/projects/details/VPH-Share). Atmosphere is also the main frontend to cloud resources in the PL-Grid infrastructure for research in Poland (http://www.plgrid.pl/en).
- The HyperFlow workflow engine [3] that enables the execution of scientific workflows tasks on the available computing resources (e.g. Virtual Machines in a cloud). Development of HyperFlow also includes an innovative approach for deployment, execution and autoscaling of scientific workflows leveraging cloud resources, including the capability to deploy the entire workflow runtime environment as part of the workflow application. This approach allows us to avoid tight coupling to a particular cloud infrastructure and middleware.
- Methods for scheduling and cost optimization of scientific workflows on cloud infrastructures, based on mathematical programming and heuristic approaches [4,5,6]. The challenge in developing such optimization models lies in the necessity to find a good trade-off between the model complexity and its capability to find optimal solutions in reasonable time. So far,we have addressed the problem of bag-of-task applications and multi-level workflows under deadline constraints, executed on hybrid cloud infrastructures.
- Methodology for evaluation of cloud providers for scientific and industry applications [2,7,8]. The detailed benchmarks of multiple cloud providers and resource types are crucial not only for the selection of the best infrastructure for large-scale deployments, but also enables us to create the performance models of the applications. These, in turn, can be used as input data to the optimization and scheduling algorithms.

2. Experiments

The development of the methods and tools outlined above requires extensive experiments on real cloud infrastructures. Such experiments not only allow us to test the developed tools in a large scale before deployment in production, but also to evaluate how various configurations of deployments affect the performance of the tools, applications, and optimization algorithms.

The NFSCloud Chameleon and CloudLab facilities will provide the infrastructure to run the following experiments:

- Evaluation of advanced autoscaling techniques for Atmosphere cloud platform. The developed autoscaling solutions, based on complex event processing and time series databases, require repeated tests under varying workloads. NSFCloud facilities will help to conduct such experiments in an isolated environment, which is not possible e.g. with the production OpenStack installations that we work with on a daily basis.
- Scalability of scientific workflows in HyperFlow model: the heterogeneous infrastructure provided by NSFCloud facilities will enable to execute large-scale deployments on multi-site cloud, where the issues of data transfers and locality are important factors, not visible in private or single-cloud installations. In addition to the scalability tests, the experiments will allow us to calibrate the performance models of applications, taking into account the network latency and bandwidth limitations.
- Investigation of the influence of variability of cloud infrastructures on the quality of scheduling and optimization algorithms. Static workflow scheduling methods assume that the estimates of task runtimes are available prior to application execution, and this assumption holds to some extent when the performance model of the application and infrastructure is well known and measured. The problem arises, however, when the runtime variations and various uncertainties influence the actual execution. Having a large-scale experimental testbed will allow investigating the influence of the uncertainties on the performance of optimization algorithms, and help develop new models that mitigate their negative effects.
- Interoperation of cloud testbed of PL-Grid infrastructure with NSFCloud facilities, for transatlantic and global-scale experiments involving wide-area and high-latency networks.

To thoroughly analyze and understand the nature of hybrid cloud platforms, we consider combining the NSFCloud facilities with commercial cloud providers, such as Amazon EC2 or Google Compute Engine. Considering our to-date experience with cloud computing infrastructures, and the experiments we performed so far at AGH, ISI and at Notre Dame, we hope that this research will constitute an interesting contribution towards better understanding of using hybrid cloud computing resources for scientific applications.

3. References

1. P. Nowakowski, T. Bartyński, M. Bubak, T. Gubała, D. Harężlak, M. Kasztelnik, M. Malawski, J. Meizner, Development, Execution and Sharing of VPH Applications in the Cloud with the Atmosphere Platform, VPH 2014 Conference, Trondheim, Norway 2014.
2. M. Bubak, M. Kasztelnik, M. Malawski, J. Meizner, P. Nowakowski, and S. Varma: Evaluation of Cloud Providers for VPH Applications. CCGrid 2013, 13-16 May, 2013, Delft, the Netherlands (2013)
3. B. Balis, Increasing Scientific Workflow Programming Productivity with HyperFlow. In Proc. WORKS'14, 9th Workshop on Workflows in Support of Large-Scale Science, 2015.
4. M. Malawski, K. Figiela, J. Nabrzyski: Cost Minimization for Computational Applications on Hybrid Cloud Infrastructures. Future Generation Comp. Syst. 29(7): 1786-1794
5. Maciej Malawski, Kamil Figiela, Marian Bubak, Ewa Deelman, Jarek Nabrzyski: Cost Optimization of Execution of Multi-level Deadline-Constrained Scientific Workflows on Clouds. PPAM (1) 2013: 251-260
6. Maciej Malawski, Gideon Juve, Ewa Deelman, Jarek Nabrzyski: "Cost- and Deadline-Constrained Provisioning for Scientific Workflow Ensembles in IaaS Clouds", in 24th IEEE/ACM Conference on Supercomputing (SC12), 2012.
7. K. Zieliński, M. Malawski, M. Jarząb, S. Zieliński, K. Grzegorczyk, T. Szepieniec, M. Zyśk, Evaluation Methodology of Converged Cloud Environment, in: KU KDM 2014 : seventh ACC Cyfronet AGH users' conference : Zakopane, 12-14 Mar, 2014 : proceedings. — Kraków : ACK Cyfronet AGH, [2014], pp. 77-78
8. Sepideh Azarnoosh, Mats Rynge, Gideon Juve, Ewa Deelman Michal Nieć, Maciej Malawski and Rafael Ferreira da Silva, Introducing PRECIP: An API for Managing Repeatable Experiments in the Cloud, Workshop on Cloud Computing for Research Collaborations (CRC), 2013.