Guest Editors' Introduction: Special Issue on Science-driven Cloud Computing

[1]Ivona Brandic, [2]Ewa Deelman, [3,4]Ioan Raicu

[1] Information Systems Institute, Vienna University of Technology, Vienna, Austria
[2] Information Sciences Institute, University of Southern California, Marina Del Rey, CA, USA
[3] Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA
[4] Mathematics and Computer Science Div., Argonne National Laboratory, Argonne, IL, USA

ivona@infosys.tuwien.ac.at, deelman@isi.edu, iraicu@cs.iit.edu

1. INTRODUCTION

It is our honor to serve as guest editors of this special issue of the Scientific Programming Journal on Science-driven Cloud Computing. We are pleased to present 6 high-quality contributions that focus on scientific cloud computing. Scientific computing involves a broad range of technologies, from high-performance computing (HPC) which is heavily focused on compute-intensive applications, high-throughput computing (HTC) which focuses on using many computing resources over long periods of time to accomplish its computational tasks, many-task computing (MTC) which aims to bridge the gap between HPC and HTC by focusing on using many resources over short periods of time, to data-intensive computing which is heavily focused on data distribution and harnessing data locality by scheduling of computations close to the data.

This special issue on Science-driven Cloud Computing provides the scientific community a dedicated forum for discussing new research, development, and deployment efforts in running these kinds of scientific computing workloads on Cloud Computing infrastructures. This special issue will focus on the use of cloud-based technologies to meet new compute intensive and data intensive scientific challenges that are not well served by the current supercomputers, grids or commercial clouds. What architectural changes to the current cloud frameworks (hardware, operating systems, networking and/or programming models) are needed to support science? Dynamic information derived from remote instruments and coupled simulation and sensor ensembles are both important new science pathways and tremendous challenges for current HPC/HTC/MTC technologies. How can cloud technologies enable these new scientific approaches? How are scientists using clouds? Are there scientific HPC/HTC/MTC workloads that are suitable candidates to take advantage of emerging cloud computing resources with high efficiency? What benefits exist by adopting the cloud model, over clusters, grids, or supercomputers? What factors are limiting clouds use or would make them more usable/efficient?

Cloud computing first established in the business computing domain is now a topic of research in computer science and an interesting execution platform for science applications. Today there are a number of commercial and science cloud deployments, including those provided by Amazon, Google, IBM, Microsoft, and others. Campus and national labs are also deploying their own cloud solutions. The ability to control the resources and the pay-as-you go usage model enables new approaches to application development and resource provisioning. Science applications are looking towards the cloud to provide a stable and customizable execution environment. This special issue of the Scientific Programming Journal is dedicated to the computational challenges and opportunities of cloud computing. It is also an extension to several workshops the guest editors have been involved in, including the Scientific Cloud Computing (ScienceCloud) workshop series (see www.cs.iit.edu/~iraicu/ScienceCloud2011/).

This special issue reflects current research trends in developements of Cloud computing infrastructures for HPC and parallel computing area. The first trend is optimization of Cloud resources based on the available programming models for HPC (like workflows) where the control mechanisam for the application execution are external to the Cloud. The second trend are research issues in application of currently available public Clouds for HPC resulting in adaptation of data structures and programming models e.g., data format adaptation, performance and cost analysis for computations in

public Clouds and evaluation of available infrastructure for storage of large data sets.

2. In This Issue

It was a challenge to select just 6 of the 18 high-quality submissions for inclusion here. We thank the 56 reviewers for their thoughtful work. We briefly introduce the accepted articles in the following several paragraphs.

Manish Parashar et al.'s "**Autonomic Management of Applications Workflows on Hybrid Computing Infrastructure**" proposes a framework to manage autonomics, including scheduling the mix of hybrid resources, application management, monitoring resources and adaptation of resource provisioning based on objectives and metrics. The authors explored the autonomics using a real world scientific workflow, the Defiant reservoir simulator using Ensemble Kalman Filter with several different objectives (e.g. acceleration, conservation and resilience) and metrics (e.g. deadline and budget). They formed a hybrid infrastructure with TeraGrid and several instance types of Amazon EC2 and the results show that the proposed framework for autonomics works well efficiently for the application workflow achieving the objectives using the metrics.

Satish Srirama et al.'s "**Scalability of Parallel Scientific Applications on the Cloud**" studies the effects of moving parallel scientific applications onto the cloud through deploying several benchmark applications (e.g. matrix-vector operations, NAS parallel benchmarks, and DOUG – Domain decomposition On Unstructured Grids) on the cloud. The authors also observed the limitations of cloud and its comparison with cluster in terms of performance, and raises important issues around the necessity for better frameworks or optimizations for MapReduce style applications.

Keith Jackson et al.'s "**Performance and Cost Analysis of The Supernova Factory on The Amazon AWS Cloud**" studies the feasibility of porting the Nearby Supernova Factory pipeline (a complex pipeline of serial processes that execute various image processing algorithms in parallel on ~10TBs of data) to the Amazon Web Services cloud environment. Specifically the authors describe the tool set they developed to manage a virtual cluster on Amazon EC2, explore the various design options available for application data placement, and offer detailed performance results and lessons learned from each of the design options.

Zach Hill et al.'s "**Early Observations on the Performance of Windows Azure**" presents the results of comprehensive performance experiments conducted on Windows Azure from October 2009 to February 2010. The authors present performance and reliability observations and analysis from their deployment of a large-scale scientific application hosted on Azure, called ModisAzure, that show unusual and sporadic VM execution slowdown of over 4X in some cases and affected up to 16% of task executions at times. In addition to a detailed performance evaluation of Windows Azure, the authors provide recommendations for potential users of Windows Azure based on these early observations. Although the discussion and analysis is tailored to scientific applications, the results are broadly applicable to the range of existing and future applications running in Windows Azure.

Gabriela Turcu et al.'s "**Reshaping Text Data for Efficient Processing on Amazon EC2**" investigates provisioning on the Amazon EC2 cloud from the user perspective, attempting to provide a scheduling strategy that is both timely and cost effective. The authors rely on the empirical performance of the application of interest on smaller subsets of data, to construct an execution plan. Using predictions of the performance of the application based on measurements on small data sets, the authors devise an execution plan that meets a user specified deadline while minimizing cost.

Ani Thakar et al.'s "**Large Science Databases - Are Cloud Services Ready For them?**" reports on attempts to put an astronomical database – the Sloan Digital Sky Survey science archive – in the cloud. The authors find that it is very frustrating to impossible at this time to migrate a complex SQL Server database into current cloud service offerings such as Amazon (EC2) and Microsoft (SQL Azure). Certainly it is impossible to migrate a large database in excess of a TB, but even with (much) smaller databases, the limitations of cloud services make it very difficult to migrate the data to the cloud without making changes to the schema and settings that would degrade performance and/or make the data unusable. Preliminary performance comparisons show a large performance discrepancy with the Amazon cloud version of the

SDSS database. The authors describe a powerful new computational instrument that they are developing in the interim – the Data-Scope – that will enable fast and efficient analysis of the largest (petabyte scale) scientific datasets. Data-Scope uses a two-tier architecture and commodity components to provide a data-intensive analysis platform that maximises sequential disk throughput while minimizing power consumption and overall cost.

3. GUEST EDITORS

Ivona Brandic is Assistant Professor at the Distributed Systems Group, Information Systems Institute, Vienna University of Technology (TU Wien). Prior to that, she was Assistant Professor at the Department of Scientific Computing, Vienna University. She received her PhD degree from Vienna University of Technology in 2007. From 2003 to 2007 she participated in the special research project AURORA (Advanced Models, Applications and Software Systems for High Performance Computing) and the European Union's GEMSS (Grid-Enabled Medical Simulation Services) project. She is involved in the European Union's SCube project and she is leading the Austrian national FoSII (Foundations of Self-governing ICT Infrastructures) project funded by the Vienna Science and Technology Fund (WWTF). She is Management Committee member of the European Commission's COST Action on Energy Efficient Large Scale Distributed Systems. From June-August 2008 she was visiting researcher at the University of Melbourne. Her interests comprise SLA and QoS management, Service-oriented architectures, autonomic computing, workflow management, and large scale distributed systems (Cloud, Grid, Cluster, etc.).

Ewa Deelman is a Project Leader in the Advanced Systems Division at the University of Southern California (USC) Information Sciences Institute (ISI). She is also a Research Associate Professor in the Computer Science Department at USC. Her main area of research is scientific workflow management in Grids. As part of this work, she is leading the design and development of the Pegasus software that maps complex application workflows onto distributed resources. Pegasus is being used in a variety of scientific applications. She is also interested in large-scale data management issues, especially concerning metadata management. In particular, she is leading the design and development of the Metadata Catalog Service (MCS), a catalog that allows for the storing and querying of descriptive attributes associated with data objects such as files. Before joining ISI, she was a Sr. Development Engineer in the Parallel Computing Laboratory at UCLA. She received her PhD in Computer Science from Rensselaer Polytechnic Institute in 1997, focusing on "Optimizing Parallel Discrete Event Simulation for Spatially Explicit Problems".

Ioan Raicu is an assistant professor in the Department of Computer Science at Illinois Institute of Technology (IIT), as well as a guest research faculty in the Math and Computer Science Division at Argonne National Laboratory. He is also the founder and director of the Data-Intensive Distributed Systems Laboratory at IIT. He has received the prestigious NSF CAREER award in 2011 for his innovative work on distributed file systems for exascale computing. He was a NSF/CRA Computation Innovation Fellow at Northwestern University in 2009-2010, and obtained his Ph.D. in Computer Science from University of Chicago under the guidance of Dr. Ian Foster in 2009. He is a 3-year award winner of the GSRP Fellowship from NASA Ames Research Center. His research work and interests are in the general area of distributed systems. His work focuses on a relatively new paradigm of Many-Task Computing (MTC), which aims to bridge the gap between two predominant paradigms from distributed systems, High-Throughput Computing and High-Performance Computing. His work has focused on defining and exploring both the theory and practical aspects of realizing MTC across a wide range of large-scale distributed systems. He is particularly interested in resource management in large scale distributed systems with a focus on many-task computing, data intensive computing, cloud computing, grid computing, and many-core computing.