Recent Advances in Autonomic Provisioning of Big Data Applications on Clouds

Rajiv Ranjan, Lizhe Wang, *Senior Member, IEEE*, Albert Y. Zomaya, *Fellow, IEEE*, Dimitrios Georgakopoulos, Xian-He Sun, *Fellow, IEEE*, and Guojun Wang

CLOUD computing [1] assembles large networks of virtualised ICT services such as hardware resources (such as CPU, storage, and network), software resources (such as databases, application servers, and web servers) and applications. In industry these services are referred to as infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). Mainstream ICT powerhouses such as Amazon, HP, and IBM are heavily investing in the provision and support of public cloud infrastructure. Cloud computing is rapidly becoming a popular infrastructure of choice among all types of organisations. Despite some initial security concerns and technical issues, an increasing number of organisations have moved their applications and services in to "The Cloud". These applications range from generic word processing software to online healthcare. The cloud system taps into the processing power of virtualized computers on the back end, thus significantly speeding up the application for the user, which just pays for the used services.

Big Data [2], [3], [4], [5] applications has become a common phenomenon in domain of science, engineering, and commerce. Some of the representative applications include disaster management, high energy physics, genomics, connectomics, automobile simulations, medical imaging, and the like. The "BigData" problem, which is defined as the practice of collecting and analyzing complex data sets so large that it becomes difficult to analyse and interpret manually or using on-hand data management applications (e.g., Microsoft Excel). For example, in case of disaster management Big Data application there is a need to analyse "a deluge of online data from multiple sources (feeds from social media and mobile devices)" for understanding and managing real-life events such as flooding, earthquake, etc. Over 20 million tweets posted during Hurricane Sandy (2012) lead to an instance of the BigData problem. The statistics provided by the PearAnalytics study reveal that almost 44 percent of the Twitter posts are spam and pointless, about 6 percent are personal or product advertising, while 3.6 percent are news and 37.6 percent are conversational posts. During the 2010 Haiti earthquake, text messaging via mobile phones and Twitter made headlines as being crucial for disaster response, but only some 100,000 messages were actually processed by government agencies due to lack of automated and scalable ICT (cloud) infrastructure.

Large-scale, heterogeneous, and uncertain Big Data applications are becoming increasingly common, yet current cloud resource provisioning methods do not scale well and nor do they perform well under highly unpredictable conditions (data volume, data variety, data arrival rate, etc.). Much research effort have been paid in the fundamental understanding, technologies, and concepts related to autonomic provisioning of cloud resources for Big Data applications, to make cloud-hosted Big Data applications operate more efficiently, with reduced financial and environmental costs, reduced under-utilisation of resources, and better performance at times of unpredictable workload.

Targeting the aforementioned research challenges, this special issue compiles recent advances in Autonomic Provisioning [6], [7] of Big Data Applications on Clouds.

Following papers put their focus on infrastructure level Cloud management for optimizing big data flow processing:

- Virtualized clouds introduce performance variability in resources, thereby impacting the application's ability to meet its quality of service (QoS). This motivates the need for autonomic methods of provisioning elastic resources as well as dynamic task selection, for continuous dataflow applications on clouds. Kumbhare et al. extend continuous dataflows to the concept of "dynamic dataflows", which utilize alternate tasks definitions and offer additional control over the dataflow's cost and QoS. They formalize an optimization problem to automate both deployment time and runtime cloud resource provisioning of such dynamic dataflows that allows for trade-offs between the application's value and the resource cost. They propose two greedy heuristics, centralized and shared, based on the variable sized bin packing algorithm to solve this NP-hard problem. Further, they also present a genetic algorithm (GA)

- *R. Ranjan is with Digital Productivity Flagship CSIRO, Australia. E-mail: rajiv.ranjan@csiro.au.*
- *L. Wang is with the School of Computer Science, China University of Geosciences. E-mail: lizhe.wang@computer.org.*
- *A. Zomaya is with the School of Information Technologies, The University of Sydney, Australia. E-mail: albert.zomaya@sydney.edu.au.*
- *D. Georgakopoulos is with the RMIT University, Australia. E-mail: dimitrios.georgakopoulos@rmit.edu.au.*
- *X.-H. Sun is with Department of Computer Science, Illinois Institute of Technology. E-mail: sun@iit.edu.*
- *G. Wang is with the School of Information Science and Engineering, Central South University, Changsha, Hunan Province, P.R. China. E-mail: csgjwang@csu.edu.cn.*

For information on obtaining reprints of this article, please send e-mail to: reprints.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCC.2015.2437231

metaheuristic that gives a near optimal solution by exploring a wide range of possible configurations.

- Elasticity has now become the elemental feature of cloud computing as it enables the ability to dynamically add or remove virtual machine instances when workload changes. However, effective virtualized resource management is still one of the most challenging tasks. When the workload of a service increases rapidly, existing approaches cannot respond to the growing performance requirement efficiently because of either inaccuracy of adaptation decisions or the slow process of adjustments, both of which may result in insufficient resource provisioning. As a consequence, the QoS of the hosted applications may degrade and the service level objective (SLO) will be thus violated. Liu et al. introduce SPRNT, a novel resource management framework, to ensure high-level QoS in the cloud computing system. SPRNT utilizes an aggressive resource provisioning strategy which encourages SPRNT to substantially increase the resource allocation in each adaptation cycle when workload increases.
- Public clouds allow virtually any institution in the world to deploy a set of virtual machines (VMs) to do their analytics. It becomes increasingly complex and time consuming in the setting of hundreds or thousands of VMs crunching tens or hundreds of TB in public Clouds. Vaquero et al. present techniques for on-demand, smarter data distribution on top of public clouds.
- To achieve energy-efficiency control and simultaneously satisfying QoS guarantees have become critical issues for large-scale cloud providers. Chiang et al. propose three green operating systems implemented in cloud to mitigate server idle power. The challenge of controlling service rate and applying N-policy to optimize cost and simultaneously meet performance guarantee in different green systems is studied first. A cost function is developed by taking power consumption cost, system congestion cost and mode-switching cost into consideration. The effects of energy-efficiency controls and operating configurations on system performances and operational cost are demonstrated and compared. The first proposed EGC algorithm allows cloud providers to make a costs/performances tradeoff analysis and solve the nonlinear constrained optimization problem.

Some work in the special issue put their effort on developing techniques for cloud job management:

- The scheduling of multitask jobs on clouds is an NP-hard problem. The problem becomes even worse when complex workflows are executed on large elastic clouds, such as Amazon EC2 or IBM RC2. The main difficulty lies in the large search space and high overhead for generation of optimal schedules, especially for real-time applications with dynamic workloads. A new iterative ordinal optimization (IOO) method is proposed by Zhang et al. The ordinal optimization method is applied in each iteration to achieve suboptimal schedules. IOO aims at generating more efficient schedules from a global perspective over a long period. They prove through overhead analysis the advantages in time and space efficiency in using the IOO method. The IOO method is designed to adapt to system dynamism to yield suboptimal performance.
- As cloud computing makes computing a utility, scientists across different application domains are facing the same challenge of reducing financial cost in addition to meet the traditional goal of performance optimization. Wu et al. develop a prototype generic workflow system by leveraging existing technologies for a quick evaluation of scientific workflow optimization strategies. They construct analytical models to quantify the network performance of scientific workflows using cloud-based computing resources, and formulate a task scheduling problem to minimize the workflow end-to-end delay under a user-specified financial constraint.
- Despite recent efforts toward designing resource-efficient MapReduce schedulers for large-scale cloud applications, existing solutions that focus on scheduling at the task-level still offer sub-optimal job performance. This is because tasks can have highly varying resource requirements during their lifetime, which makes it difficult for task-level schedulers to effectively utilize available resources to reduce job execution time. To address this limitation, Zhang et al. introduce PRISM, a fine-grained resource-aware MapReduce scheduler that divides tasks into phases, where each phase has a constant resource usage profile, and performs scheduling at the phase level.

Two papers makes advances in large-scale cloud data management:

- In order to tackle the challenge is that the computational burden is too huge for the users to compute the public authentication tags of file blocks in large-scale cloud storage, Li et al. propose a new cloud storage architecture with two independent cloud servers, that is, the cloud storage server and the cloud audit server, where the latter is assumed to be semi-honest. In particular, they consider the task of allowing the cloud audit server, on behalf of the cloud users, to pre-process the data before uploading to the cloud storage server and later verifying the data integrity.
- In large cloud computing environments, existing range-aggregate queries are insufficient to quickly provide accurate results in big data environments. Yun et al. propose FastRAQ—a fast approach to range-aggregate queries in big data environments. FastRAQ first divides big data into independent partitions with a balanced partitioning algorithm, and then generates a local estimation sketch for each partition. When a range-aggregate query request arrives, FastRAQ obtains the result directly by summarizing local estimates from all partitions.

Concerning the cloud network aspect, bandwidth efficient execution of online big data analytics in telecommunication networks demands for tailored solutions. Existing streaming analytics systems are designed to operate in large

data centers, assuming unlimited bandwidth between data center nodes. Applying these solutions without modification to distributed telecommunication clouds, overlooks the fact that available bandwidth is a scarce and costly resource making the telecommunication network valuable to end-users. Theeten et al. present Continuous Hive (CHive), a streaming analytics platform tailored for distributed telecommunication clouds. The fundamental contribution of CHive is that it optimizes query plans to minimize their overall bandwidth consumption when deployed in a distributed telecommunication cloud.

In large-scale cloud computing application domains, Baek et al. propose a secure cloud computing based framework for big data information management in smart grids, which is called "Smart-Frame." The main idea of our framework is to build a hierarchical structure of cloud computing centers to provide different types of computing services for information management and big data analysis. In addition to this structural framework, they present a security solution based on identity-based encryption, signature and proxy re-encryption to address critical security issues of the proposed framework.

Rajiv Ranjan
Lizhe Wang
Albert Zomaya
Dimitrios Georgakopoulos
Xian-He Sun
Guojun Wang
Guest Editors

ACKNOWLEDGMENTS

Lizhe Wang is the corresponding author.

REFERENCES

[1] *Cloud Computing: Methodology, Systems, and Applications*, L. Wang, R. Ranjan, J. Chen, and B. Benatallah Eds. Boca Raton, FL: CRC Press, Taylor and Francis Group, p. 844, Oct. 2011.
[2] R. Pepper and J. Garrity, "The internet of everything: How the network unleashes the benefits of big data," *CISCO* [Online]. Available: http://blogs.cisco.com/wp-content/uploads/GITR-2014-Cisco-Chapter.pdf. Accessed on May 27th, 2015.
[3] Bringing Big Data to the Enterprise, IBM [Online]. Available: http://www-01.ibm.com/software/in/data/bigdata/. Accessed on May 27th, 2015.
[4] J. Gantz et al., "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," *IDC IVIEW*, Sponsored by EMC Corporation. Available at: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf. Accessed on May 27th, 2015.
[5] Tomorrow's Internet Today, CISCO [Online]. Available: http://www.cisco.com/c/dam/en/us/products/collateral/routers/asr-9000-series-aggregation-services-routers/brochure_tomorrows_internet_today.pdf. Accessed on May 27th, 2015.
[6] R. Ranjan, "Streaming Big Data Processing in Datacentre Clouds, " *IEEE Cloud Computing, BlueSkies Column*, vol. 1, no. 1, pp. 78–83, May 2014.
[7] L. Wang and R. Ranjan, "Processing distributed internet of things data in clouds, " *IEEE Cloud Computing, BlueSkies Column*, vol. 2, no. 1, pp. 76–80, Apr. 2015.

Rajiv Ranjan is a Julius fellow (2013-2016), senior research scientist and project leader in the Digital Productivity and Services Flagship of Commonwealth Scientific and Industrial Research Organization (CSIRO – Australian Government's Premier Research Agency). His main research interests include autonomic management of applications in distributed systems (clouds, datacentres, big data). He has published more than 120 scientific papers in premier venues such as the *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Transactions on Computers*, *IEEE Transactions on Cloud Computing, ACM/IEEE World Wide Web Conference, and IEEE Transactions on Cloud Computing*. According to Google Scholar, his papers have received more than 4,000 citation with an h-index of 25. More information can be found at: http://rajivranjan.net.

Lizhe Wang is a professor at the Institute of Remote Sensing & Digital Earth, Chinese Academy of Sciences (CAS) and a ChuTian chair professor at the School of Computer Science, China University of Geosciences (CUG). received the BE and ME degrees from Tsinghua University and Doctor of Engineering degree from the University of Karlsruhe (Magna Cum Laude), Germany. He serves as an associate editor in the *IEEE Transactions on Computers* and *IEEE Transactions on Cloud Computing*. His main research interests include high-performance computing, e-Science, and spatial data processing. He is a fellow of the IET, fellow of the British Computer Society.

Albert Y. Zomaya is the Chair Professor of High Performance Computing & Networking in the School of Information Technologies, The University of Sydney. He is also the Director of the Centre for Distributed and High Performance Computing which was established in late 2009. Professor Zomaya published more than 500 scientific papers and articles and is author, co-author or editor of more than 20 books. He served as the Editor in Chief of the *IEEE Transactions on Computers* (2011-2014). Also, Professor Zomaya serves as an associate editor for 22 leading journals. Professor Zomaya is the recipient of the *IEEE Technical Committee on Parallel Processing Outstanding Service Award* (2011), the *IEEE Technical Committee on Scalable Computing Medal for Excellence in Scalable Computing* (2011), and the *IEEE Computer Society Technical Achievement Award* (2014). He is a Chartered Engineer, a Fellow of AAAS, IEEE, IET (UK). Professor Zomaya's research interests are in the areas of parallel and distributed computing and complex systems.

Dimitrios Georgakopoulos has recently joined RMIT as a professor at the School of Computer Science and Information Technology in Melbourne, Australia. Until June 2014, he was the director in the Information Engineering Laboratory, CSIRO's ICT Centre. Before coming to CSIRO in 2008, he held research and management positions in several industrial laboratories in the US. From 2000 to 2008, he was a senior scientist with Telcordia, where he helped found Telcordia's Research Centers in Austin, Texas, and Poznan, Poland. From 1997 to 2000, he was a technical manager at the Microelectronics and Computer Corporation (MCC), and the project manager and chief architect in MCC's Collaboration Management Infrastructure (CMI) consortial project. From 1990 to 1997, he was a principal scientist at GTE (currently Verizon) Laboratories Inc. He is an adjunct professor at the Australian National University and a CSIRO adjunct science fellow. He has published more than 130 scientific papers.

Xian-He Sun is a distinguished professor of computer science in the Department of Computer Science, Illinois Institute of Technology (IIT). He is the director in the Scalable Computing Software Laboratory, IIT and a guest faculty in the Mathematics and Computer Science Division, Argonne National Laboratory. Before joining IIT, he was at DoE Ames National Laboratory, at ICASE, NASA Langley Research Center, at Louisiana State University, Baton Rouge, and was an ASEE fellow at Navy Research Laboratories. He is known for his memory-bounded speedup model, also called Sun-Ni's Law, for scalable computing. His research interests include parallel and distributed processing, memory and I/O systems, software systems for big data applications, and performance evaluation. He has more than 200 publications and five patents in these areas. He is a former IEEE CS distinguished speaker, a former vice chair in the IEEE Technical Committee on Scalable Computing, the past chair in the Computer Science Department at IIT (September 2009-August 2014), and is serving and served on the editorial board of most of the leading professional journals in the field of parallel processing. He is a fellow of the IEEE. More information about him can be found at his web site www.cs.iit.edu/∼sun/.

Guojun Wang received the BSc degree in geophysics, and the MSc and PhD degrees in computer science from Central South University, China. He is the head and a professor in the Department of Computer Science and Technology, Central South University. He is also the director in the Trusted Computing Institute at Central South University. He has been an adjunct professor at Temple University, US; a visiting scholar at Florida Atlantic University, US; a visiting researcher at the University of Aizu, Japan; and a research fellow at Hong Kong Polytechnic University. His research interests include network and information security, trusted computing, transparent computing, and cloud computing. He is a distinguished member of the CCF, and a member of the the IEEE, ACM, and IEICE.