Guest Editors' Introduction: Special Issue on Utility and Cloud Computing Science and Technology

Irena Bojanova and Ching-Hsien Hsu

1 Introduction

Computing is rapidly moving towards a model, where it is provided as services that are delivered in a manner similar to traditional utilities, such as water, electricity, gas, and telephony. In such a model, users access services according to their requirements, without regard to where the services are hosted or how they are delivered. Several computing architectures have evolved to realize this utility computing vision, including grid computing, service-oriented architecture (SOA), and cloud computing, which has recently shifted into the center of attention in the ICT industry. Increasing numbers of IT vendors are promising to offer applications, storage, and computation hosting services with conforming service-level agreements (SLA) to ensure quality of services (QoS) and performance [1]. Considering many of these services are hosted in traditional data centers, there is significant complexity involved in ensuring the scalability, availability, manageability and accessibility of applications, services, and data, as the scale of the systems as well as the users grows. As a result, it is becoming important to investigate the use of cloud computing techniques and its interoperability with utility computing. This special issue focuses on principles, paradigms, and applications of "utility computing" and its practical realization, especially in the context of cloud computing.

Despite a wide body of research development effort, ensuring the communications of cyberspaces and real spaces, how to realize cloud and utility services remains an open challenge. This special issue is in response to the increasing convergence between cloud and utility technologies and services. While different approaches exist, challenges and opportunities are numerous in this context. The research papers selected for this special issue represent recent progress in the field, including works on cloud architectures, mobile computing, security and energy issues, services computing and modelling, resources management, virtualization technologies and applications. All of these papers not only provide novel ideas and state-of-the-art techniques in the field, but also stimulate future research in merging cloud and utility services.

2 Virtualization and Communication Technologies

By abstracting from hardware addresses and lower-level communication, the publish/subscribe paradigm seems like an adequate abstraction for supporting communication across clouds, as it supports many-to-many communication between publishers and subscribers, of which one-to-one or one-to-many can be viewed as special cases. The paper by Chamikara Jayalath, Julian Stephen and Patrick Eugster entitled "Universal Cross-Cloud Communication" proposes a content-based publish/subscribe (CPS) architecture for cloud-of-clouds communication that can dynamically identify entourages of publishers and corresponding subscribers. The authors introduce a CPS system, named Atmosphere, which leverages this architecture to dynamically connect the publishers with their entourages through overlays. These overlays can transmit messages from a publisher to its corresponding subscribers with low latency. The experiments show that the proposed architecture makes the generic CPS abstraction viable and beneficial for many applications. In particular, the authors illustrate how Atmosphere allows to implement, with little effort, versions of the popular HDFS and ZooKeeper systems, which operate efficiently across data-centers.

Network I/O virtualization plays an important role in cloud computing. The paper by En-Hao Chang, Chen-Chieh Wang, Chien-Te Liu, Kuan-Chung Chen, and Chung-Ho Chen entitled "Virtualization Technology for TCP/IP Offload Engine" addresses the system-wide architecture issues of TCP/IP offload engine (TOE) virtualization and presents the architectural designs. The authors identify three critical factors that affect the performance of a TOE: I/O virtualization architectures, quality of service, and virtual machine monitor scheduler. In this proposed architecture, the VMM manages the socket connections in the TOE directly and thus can eliminate packet copy and de-multiplexing overheads as appeared in the virtualization of a layer-2 network card. To reduce hypervisor intervention, the direct I/O access architecture helps removing most of the VMM interventions. To continue serving the TOE commands for a VM, no matter the VM is idle or switched out by the VMM,

- *I. Bojanova is with the Information and Technology Systems Department, University of Maryland University College, Adelphi, MD. E-mail: irena.bojanova@umuc.edu.*
- *C.-H. Hsu is with the Department of Computer Science and Information Engineering, Chung Hua University, Taiwan. E-mail: chh@chu.edu.tw.*

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.

Digital Object Identifier no. 10.1109/TCC.2014.2321706

the TOE I/O command dispatcher is decoupled from the VMM scheduler. A VMM scheduler with preemptive I/O scheduling and a programmable I/O command dispatcher with deficit weighted round robin (DWRR) policy is able to ensure service fairness and at the same time maximize the TOE utilization.

Owing to the high interrupt frequency and heavy cost per interrupt in high-speed network virtualization, the performance of network virtualization is closely correlated to the computing resource allocation policy [2]. However, the I/O-intensive and CPU-intensive applications in virtual machines are treated in the same manner since application attributes are transparent to the scheduler in a hypervisor, and this unawareness of workload makes virtual systems unable to take full advantages of high performance networks. The paper by Haibing Guan, Jian Li and Ruhui Ma entitled "Workload-Aware Credit Scheduler for Improving Network I/O Performance in Virtualization Environment" discusses the networking solution and shows by experiment that the current credit scheduler in Xen does not utilize high performance networks efficiently. The authors propose a novel workload-aware scheduling model with two optimizations to eliminate the bottleneck caused by the scheduler. Guest domains are divided into I/O-intensive domains and CPU-intensive domains according to their monitored behaviour. I/O-intensive domains can obtain extra credits that CPU-intensive domains are willing to share. Experimental evaluations show that the new scheduling models improve bandwidth and reduce response time, by keeping the fairness between I/O-intensive and CPU-intensive domains. This enables the virtualized infrastructure to provide cloud computing services more efficiently and predictably.

3 Secure, Green, and Utility Clouds

The adoption of cloud computing as the fifth utility requires the solution of several information confidentiality problems [3], especially when considering the database as a service (DaaS) paradigm that can support the most important Internet-based applications. The paper by Luca Ferretti, Fabio Pierazzi, Michele Colajanni, and Mirco Marchetti entitled "Performance and cost evaluation of adaptive encryption for cloud database services" proposes an adaptive encryption architecture that represents the best solution to the trade-off between data confidentiality and usability in public cloud databases. The authors demonstrate the feasibility and the performance of the proposed solution through a software prototype. This study also investigated the cost increases due to the integration of adaptive encryption schemes. The paper opens a possible way to combine data confidentiality and cloud database services.

Many researchers in market-based literatures have highlighted user satisfaction as a significant antecedent to user loyalty. SLA violation as an important factor can decrease users' satisfaction level. The amount of that decrease depends on user' characteristics. Some of these characteristics are related to QoS requirements and announced to the service provider through SLAs. However, some user's characteristics are hidden for the service provider and selfish users are not interested to reveal them truly. The paper by H. Morshedlou and M.R. Meybodi entitled "Decreasing Impact of SLA Violations: A Proactive Resource Allocation Approach for Cloud Computing Environments" uses two user's hidden characteristics, named willingness to pay for service and willingness to pay for certainty, to present a new proactive resource allocation approach with aim of decreasing impact of SLA violations. New methods based on learning automaton for estimation of these characteristics are provided as well. To validate the proposed approach, the authors conducted numerical simulations in critical situations. The results illustrate that their approach has ability to preserve users' satisfaction in a good level while there is not any reduction in profitability.

Energy conservation is a major concern in cloud computing systems, because it can bring several important benefits such as reducing operating costs, increasing system reliability, and prompting environmental protection. Power-aware scheduling approach is a promising way to achieve that goal. The paper by Xiaomin Zhu, Laurence T. Yang, Huangke Chen, Ji Wang, Shu Yin Member and Xiaocheng Liu entitled "Real-Time Tasks Oriented Energy-Aware Scheduling in Virtualized Clouds" proposes a novel rolling-horizon scheduling architecture for real-time task scheduling in virtualized clouds. A task-oriented energy consumption model is presented and analyzed. Based on the proposed scheduling architecture, the authors develop a novel energy-aware scheduling algorithm for real-time, aperiodic, independent tasks. The proposed architecture employs a rolling-horizon optimization policy and can also be extended to integrate other energy-aware scheduling algorithms. They propose two strategies in terms of resource scaling up and resource scaling down to make a good trade-off between task's schedulability and energy conservation. Simulation experiments injecting random synthetic tasks, as well as tasks following the last version of the Google cloud tracelogs are conducted to validate the superiority of the proposed architecture by comparing it with some baselines. The experimental results show that the proposed architecture significantly improves the scheduling quality and is suitable for real-time task scheduling in virtualized clouds.

4 Mobile Cloud Computing Architectures

Mobile cloud computing improves the computational capabilities of resource-constrained mobile devices. However, mobile users demand also certain level of QoS provisioning while they use services from the cloud, even if the interfacing gateway changes due to their mobility. The paper by Sudip Misra, Snigdha Das, Manas Khatua and Mohammad S. Obaidat entitled "QoS-Guaranteed Bandwidth Shifting and Redistribution in Mobile Cloud Environment" identifies, formulates, and addresses the problem of QoS-guaranteed bandwidth shifting and redistribution among the interfacing gateways for maximizing their utility. Due to node mobility, bandwidth shifting is required for providing QoS-guarantee to the mobile nodes. However, shifting alone is not always sufficient for maintaining QoS-guarantee because of varying spectral efficiency across the associated channels, coupled with the corresponding protocol overhead involved with the computation of utility. The authors formulate bandwidth redistribution as a utility maximization problem and solve it using a modified descending bid auction. In the proposed scheme, each gateway aggregates the demands of all the connecting mobile nodes and makes a bid for the required amount of bandwidth. The cloud service provider terminates the auction process provided that the bandwidth distribution is optimal. The authors investigate the existence of Nash equilibrium in the proposed solution. Theoretically, they deduce the maximum and minimum selling prices of bandwidth, and prove the convergence of AQUM. Simulation results establish the correctness of the proposed algorithm.

5 Cloud Monitoring and Management

Compared to traditional distributed computing, it is nontrivial to optimize cloud tasks execution performance due to more constraints such as user's payment budget and divisible resource demand. The paper by Sheng Di, Cho-Li Wang and Franck Cappello entitled "Adaptive Algorithm for Minimizing Cloud Task Length with Prediction Errors" analyzes in-depth an optimal algorithm for minimizing task execution length with divisible resources and payment budget. This work derives the upper bound of cloud task length, by taking into account workload prediction errors. With such state-of-the-art bounds, the worst-case task execution performance is predictable, which can improve the quality of service in turn. The authors design a dynamic version for the algorithm to adapt to the load dynamics over task execution progress, further improving the resource utilization. A cloud prototype over a real cluster environment with 56 virtual machines was built to evaluate the proposed algorithm with different levels of resource contention. Cloud users in this cloud system are able to compose various tasks based on off-the-shelf web services. Experiments show that task execution lengths under the proposed algorithm are always close to their theoretical optimal values, even in a competitive situation with limited available resources.

Understanding the characteristics and patterns of workloads within a cloud computing environment is critical in order to improve resource management and operational conditions while QoS guarantees are maintained. Simulation models based on realistic parameters are also urgently needed for investigating the impact of these workload characteristics on new system designs and operation policies. The paper by Ismael Solis Moreno, Peter Garraghan, Paul Townend and Jie Xu entitled "Analysis, Modeling and Simulation of Workload Patterns in a Large-Scale Utility Cloud" presents a comprehensive analysis of the workload characteristics derived from a production cloud datacenter that features over 900 users submitting approximately 25 million tasks over a time period of a month. This work focuses on exposing and quantifying the diversity of behavioral patterns for users and tasks, as well as identifying model parameters and their values for the simulation of the workload created by such components. The derived model is implemented by extending the capabilities of the CloudSim framework and is further validated through empirical comparison and statistical hypothesis tests. The authors illustrate several examples of this work's practical applicability in the domain of resource management and energy-efficiency.

Workflows have been frequently used to model large-scale scientific problems. However, existing works fail to either meet the user's QoS requirements or to incorporate some basic principles of cloud computing such as elasticity and heterogeneity of the computing resources [5]. The paper by Maria A. Rodriguez and Rajkummar Buyya entitled "Deadline based Resource Provisioning and Scheduling Algorithm for Scientific Workflows on Clouds" proposes a resource provisioning and scheduling strategy for scientific workflows on infrastructure as a service (IaaS) clouds. The authors present an algorithm based on the meta-heuristic optimization technique, Particle Swarm Optimization (PSO), which aims to minimize the overall workflow execution cost while meeting deadline constraints. The proposed heuristic was evaluated using the simulator CloudSim and various well-known scientific workflows of different sizes. The results show that the proposed approach performs better than the current state-of-the-art algorithms.

Rule-based scheduling algorithms have been widely used on many cloud computing systems because they are simple and easy to implement [4]. The paper by Chun-Wei Tsai, Wei-Cheng Huang, Meng-Hsiu Chiang, Ming-Chao Chiang, and Chu-Sing Yang entitled "A Hyper-Heuristic Scheduling Algorithm for Cloud" presents a novel heuristic scheduling algorithm, called hyper-heuristic scheduling algorithm (HHSA), to find better scheduling solutions for cloud computing systems. The diversity detection and improvement detection operators are employed by the proposed algorithm to dynamically determine which low-level heuristic is to be used in finding better candidate solutions. This study compares the proposed method with several state-of-the-art scheduling algorithms, by having all of them implemented on the simulator CloudSim and the real system Hadoop. The results show that the proposed methods can significantly reduce the makespan of task scheduling, on both CloudSim and Hadoop.

6 Conclusion

All presented papers address either original research in cloud or utility computing technologies or propose novel application models in the various mobile, ubiquitous, and services fields. They trigger further related research and technology improvements in application of cloud computing and communication services.

This special issue of *IEEE Transactions on Cloud Computing (TCC)* covers different aspects of utility and cloud computing both from theoretical and practical point of view. It is information, reference, and education source for professors, researchers, and graduate students interested in updating their knowledge in cloud and utility computing, resource provisioning and management, and novel application models for merging cloud and utility services and systems. After a large open call we received more than 80 submissions and 11 research papers were selected for publication by an international editorial review committee. Each paper was reviewed by at least three reviewers in two-cycles. As per IEEE policy, for a paper co-authored by EiC of *TCC*, Dr. Irena Bojanova, UMUC, USA served as the Proxy-Editor-in-Chief in processing the reviewing independently.

The guest editors would like to express sincere gratitude to Dr. Rajkumar Buyya, EiC, *TCC*, for giving us the opportunity to work on and prepare this special issue. Next, we would like to acknowledge the help in getting Best Papers submissions from the UCC 2013 PC Chairs, Dr. Omer Rana, Cardiff University, UK and Dr. Manish Parashar, Rutgers University, USA, the CloudCom 2013 PC Chair, Dr. Siani Pearson, HP, UK and the IC2E 2013 PC Chair, Dr. Hui Lei, IBM Watson Research. In addition, we are deeply indebted to the numerous reviewers for their professional effort, insight, and help to put together these selected articles. Last but not least, we are grateful to all authors for their contributions and for undertaking two-cycle revisions of their manuscripts, without which this special issue could not have been produced. We hope that this special issue will be a good addition to the area of next generation cloud and utility science and technology.

References

[1] R. Buyya, C. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Comput. Syst.*, vol. 25, no. 6, pp. 599–616, June 2009.

[2] Y. Dong, D. Xu, Y. Zhang, and G. Liao, "Optimizing network i/o virtualization with efficient interrupt coalescing and virtual receive side scaling," in *Proc. IEEE Int. Conf. Cluster Comput.*, 2011, pp. 26–34.

[3] L. Ferretti, F. Pierazzi, M. Colajanni, and M. Marchetti, "Security and confidentality solutions for public cloud database services," in *Proc. 7th Int. Conf. Emerging Security. Inf., Syst. Technol.*, 2013, pp. 36–42.

[4] J. Li, M. Qiu, Z. Ming, G. Quan, X. Qin, and Z. Gu, "Online optimization for scheduling preemptable tasks on IaaS cloud systems," *J. Parallel Distrib. Comput.*, vol. 72, no. 5, pp. 666–677, 2012.

[5] M. Mao and M. Humphrey, "A performance study on the VM startup time in the cloud," in *Proc. 5th IEEE Int. Conf. Cloud Comput.*, 2012, 423–430.

Irena Bojanova is a professor and a program director, at Information and Technology Systems Department, University of Maryland University College (UMUC). Her research interests are in the area of cloud computing, mobile computing, and the Internet of Everything (IoE). She is, a co-editor of *Encyclopedia of Cloud Computing*, Wiley (to be published 2014), an associate editor of IEEE *Transactions on Cloud Computing*, an associate editor in chief and editor of the IT Trends department of *IEEE IT Professional*, and an associate editor of the *international journal of Big Data Intelligence*. You can read her cloud computing blog at www.computer.org/portal/web/Irena-Bojanova. She is a IEEE senior member, a general co-chair of the IT Professional Conference, and the founding chair of IEEE CS Cloud Computing STC.

Ching-Hsien Hsu is a professor in the Department of Computer Science at Chung Hua University, Taiwan; and a distinguished chair professor at Tianjin University of Technology, China. His research includes cloud computing and big eta intelligence, high-performance computing, parallel and distributed systems, ubiquitous/pervasive computing and intelligence. He is the editor-in-chief of *International Journal of Grid and High Performance Computing,* and *International Journal of Big Data Intelligence*. He received six times outstanding research award since 2005 from Chung Hua University. He is an elected member of the Phi Tau Phi Scholastic honor society, a senior member of the IEEE, the regional director of the Future Technology Research Association (FTRA), and the standing director of Taiwan Association of Cloud Computing (TACC).