Future Generation Computer Systems 29 (2013) 1211–1215

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Editorial

Model-driven provisioning of application services in hybrid computing environments☆

1. Introduction

Today, Small and Medium Business Enterprises (SMEs), governments, and universities face accelerated business change, more intense domestic and global competition and increased IT demands. They try to meet new demands through rapid implementation of innovative and inclusive business models while at the same time lowering IT barriers. These demands call for a dynamic computing model that supports rapid innovation for services and their delivery. Cloud computing [1–3], which can be an important component of such a model, is a recent advance wherein IT-related functionalities (e.g., CPU, applications, storage, network, etc.) are provided "as a service" to end-users under a usage-based payment model. In a Cloud computing model, end-users (such as SMEs, governments, and universities) can leverage virtualized application services probably on the fly based on fluctuating requirements and, in doing so, they avoid worry about infrastructure details such as where these resources are hosted or how they are managed. The new computing environment, buoyed by recent advances in the Cloud computing, has resulted in hybrid computing environments comprised of virtualized resources, applications, usage-based payment models, and networked devices. The benefit of such an environment is efficiency and flexibility, through creation of a more dynamic computing environment, where the supported functionalities are no longer fixed or locked to the underlying infrastructure. This offers tremendous automation opportunities in a variety of computing domains including, but not limited to, e-Government, e-Research, e-health and e-Business.

Resource provisioning means the selection, deployment, and run-time management of software (e.g., database servers, load-balancers, etc.) and hardware resources (e.g., CPU, storage, network, etc.) for ensuring guaranteed performance for applications. The diversity and flexibility of the IT functionalities (dynamically shrinking and growing computing systems) offered by the evolving hybrid environments, combined with the magnitudes and uncertainties [4] of its components (application workload, resource failure, and malicious activity), pose difficult problems in effective provisioning and delivery of application services. Computing systems managing hybrid environments must deal with the highly transient behaviours of end-users (e.g., arrival pattern, profile, etc.), resources (e.g., availability, performance, reliability, etc.), and networks (e.g., bandwidth, throughput, latency, etc.), all of which can be difficult to predict. The problem is further complicated by ever-increasing scale, performance, and energy saving requirements.

To counter these challenges, there is need to develop analytical models for each component that operates as a part of hybrid computing environments. These models will be important because they allow adaptive system management by establishing useful relationships between high-level performance targets (specified by operators) and low-level control parameters that system components can control or monitor. Consequently, there is a need to develop models for predicting behaviour and performance of different types of applications services and resources to adaptively transform service requests. Broad range of analytical models and statistical curve-fitting techniques such as multi-class queuing models and linear regression time series can be applied for this purpose. These models will drive and possibly transform the inputs to a service provisioner, which will improve the efficiency of the system. Such improvements will better ensure the achievement of performance targets (e.g., response time, throughput, fairness, etc.), while reducing costs due to improved utilization of resources. It will be a major advancement in the field to develop a robust and scalable system monitoring infrastructure to collect real-time data and re-adjust these models dynamically with a minimum of data and training time. We believe that these models and techniques are critical for the design of stochastic provisioning algorithms across large hybrid environments where resource availability is uncertain.

The rest of this editorial paper is organized as follows: Section 2 discusses the research challenges associated with provisioning applications in hybrid computing environments; Section 3 summarizes the research contributions that were accepted for this special issue; Section 4 concludes the paper with some future remarks.

2. Research challenges

The diversity and flexibility of the IT functionalities (dynamically shrinking and growing computing systems as shown in Fig. 1) considered by hybrid computing environments, combined with the magnitudes and uncertainties of its components (e.g., workload, cloud resources, end-users, etc.), pose difficult problems in effective provisioning and delivery of applications. Most applications consist of multiple components such as directory infrastructure, identity management, and databases. As shown in Fig. 1, these components are consolidated in a central location inside the enterprise data centre. Any attempt to migrate application component from enterprise premise to public Cloud environment is extremely complicated. In particular, following research challenges [5–8] must be overcome if hybrid computing systems have to become primetime.

☆ This special issue compiles seven technical contributions that significantly advance the state-of-the-art in model-driven provisioning of application services in hybrid computing environments.

0167-739X/$ – see front matter
doi:10.1016/j.future.2013.01.007

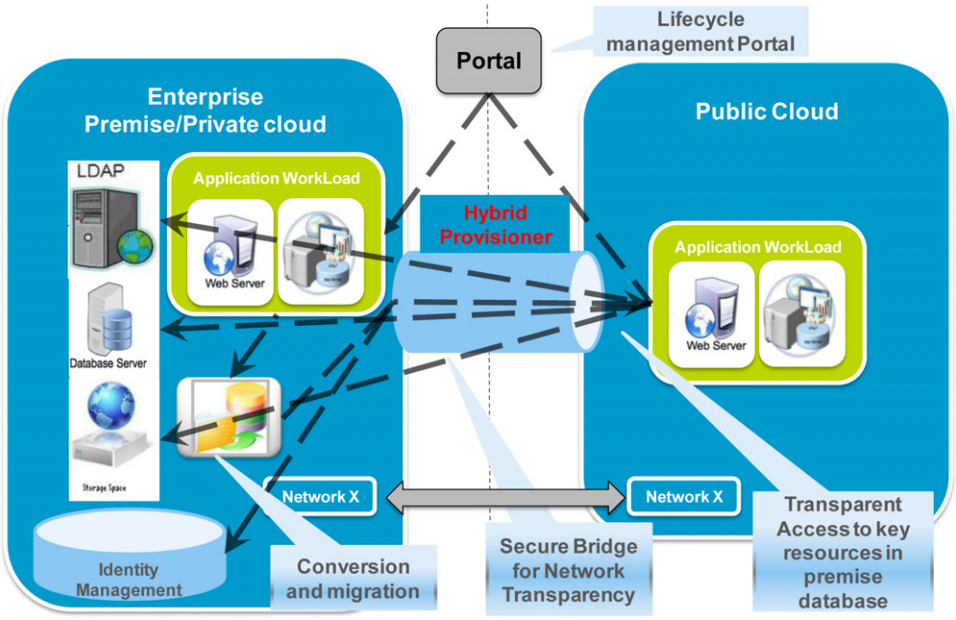

Fig. 1. An example hybrid computing environment consisting of enterprise specific private Cloud and its integration with public Clouds such as AWS, Azure, etc. Key to realizing hybrid computing environment is the design and development of "Hybrid Provisioner" software system. The main functionalities that Hybrid Provisioner needs to support includes: (i) end-to-end application data security and compliance; (ii) network QoS optimization via traffic compression, TCP optimizations, and de-duplications; (iii) easy-to-use portal interface for interacting with cloud resources and application components; and (iv) application QoS optimization while handling uncertainties.

— *Application-centric QoS Optimization*: There are a number of technical challenges related to hosting applications in hybrid computing environments consisting of in-house and public resources. Existing in-house application components (see Fig. 1) need to be significantly re-engineered in order to properly inter-operate with the virtualized Cloud resources. Mainly, technical challenges involved with porting in-house application components to cloud resources (such as CPU resources and appliance images), while ensuring that Quality of Service (QoS) targets are met under uncertainties. These challenges [8] include optimized cloud resource selection, configuration, deployment, monitoring, and control. Moreover, techniques required for efficiently porting applications to a hybrid environment depend on the application type, purpose, QoS targets, and Service Level Agreement (SLA). Notably, QoS targets vary across application types. For example, QoS targets for e-Research applications are different from static, single tier web applications (e.g., web site serving static contents) or multi-tier applications (e.g., on demand audio/video streaming). Based on application types, there is always a need to manage QoS and SLA differently.

  Future research efforts also need to focus on the development of models for forecasting end-user behaviours to understand their effect on application components in terms of CPU, storage, I/O, and network bandwidth requirements. These models will allow the Hybrid Provisioner to better understand workload demands to improve resilience to uncertainties. Further, private and public Cloud resource QoS models need to be developed that allow the provisioner to predict the effects of a provisioning schedule (e.g., type and mix of cloud resources that host application components such as web servers and database servers) on QoS targets. These models will help us understand and predict what mixes of private and public cloud resources are most suitable for hosting the given application types.

— *Scalable QoS Monitoring*: Since the components (e.g., web server, database server, storage resources, identity management system, etc.) that contribute to hybrid computing environments are distributed, there is a need to develop monitoring techniques that can scale to large system size while gracefully dealing with network, resource, and application component uncertainties. We claim that centralized approaches to QoS monitoring are not an appropriate solution due to concerns of scalability, performance, and reliability arising from the management of multiple application component queues and the expected large volume of service requests. The research challenge here is what type of network model (e.g., hierarchical, unstructured peer-to-peer, structured peer-to-peer [9,10], etc.) should be applied in architecting QoS monitoring services such that it scales with increase in application size (number of components), hybrid computing service pool (number of CPU resources provisioned), and workload patterns (e.g., number of application users, request rate, etc.).

— *End-to-End Security and Privacy*: Integrating a public cloud computing environment to on-premise computing environment will result in application data flow [11] via the public Internet. Hence, it is mandatory to develop techniques that can ensure security and privacy of data during transit as well as at rest (e.g., stored over a storage resource such as AWS S3 or AWS RDS). Though traditional firewalls can be applied in this case, defining and maintaining optimal chaining rules may impose challenging scenarios. As cloud environments and on-premise computing resources run on independent networks, it may cause incompatible network policies and unaligned IP addressing.

— *Easy-to-Use Portal Interface*: It is clear that hybrid computing environments will span multiple physical boundaries [11]. Management of application components must be made independent of the physical cloud of the hosting resource (in-house or public cloud). Hence, there is a need to develop application provisioning portal interfaces that allow seamless access and administration of applications in the way regardless of whether its components are hosted on-premise in-house or in the public cloud data centre. The portal should expose web-based interfaces for deploying resources, selecting application components, integrating components, configuring automated storage backups, etc. It should hide the nuances of the underlying

virtualization environments, heterogeneous resource configurations and abstractions. From the end-user perspective, the system and application should behave as though they are still being delivered via on-premise enterprise data centre.

3. Summary of contributions

In this section, we present the summary of the papers that were accepted for publication in this special issue that address some of the challenges discussed earlier.

For Cloud-based services to support enterprise class production workloads, Mainframe like predictable performance is essential. However, the scale, complexity, and inherent resource sharing across workloads make the Cloud management for predictable performance difficult. As a first step towards designing Cloud based systems that achieve such performance and realize the service level objectives, Rahul, Francesco, Vijay, and Kishore in their paper titled "*Modelling and Performance Analysis of Large Scale IaaS Clouds*" develop a scalable stochastic analytic model for performance quantification of Infrastructure-as-a-Service (IaaS) Cloud. Specifically, they model a class of IaaS Clouds that offer tiered services by configuring physical machines into three pools with different provisioning delay and power consumption characteristics. Performance behaviours in such IaaS Clouds are affected by a large set of parameters, e.g., workload, system characteristics and management policies. Thus, traditional analytic models for such systems tend to be intractable. To overcome this difficulty, they propose multi-level interacting stochastic sub-models approach where the overall model solution is obtained iteratively over individual sub-model solutions. By comparing with a single-level monolithic model, they show that their approach is scalable, tractable, and yet retains high fidelity. Since the dependencies among the sub-models are resolved via fixed-point iteration, they prove the existence of a solution. Results from our analysis show the impact of workload and system characteristics on two performance measures: mean response delay and job rejection probability.

Provisioning for Internet-facing applications such as cloud backup or cloud network games is sensitive to wide-area network metrics such as a round trip time, bandwidth, or loss rates. In order to optimize the quality of the service provision in hybrid clouds, it is highly valuable for the hybrid clouds to collect detailed network metrics between participating nodes of the hybrid clouds. However, since nodes can be large-scale and dynamic, the network metrics may be diverse for different cloud resources; it is challenging to increase the generality, scalability, accuracy, and the robustness of the measurement process. To solve this problem, Yongquan, Yijie, and Ernst in their paper titled "*A General Scalable and Accurate Decentralized Level Monitoring Method for Large-scale Dynamic Service Provision in Hybrid Clouds*" propose a novel distributed level monitoring method, called Hierarchical Performance Measurement (HPM), satisfying these requirements. For each kind of network metric, HPM represents the degree of pairwise closeness with discrete level values inspired by the hierarchical clustering tree. HPM maps probed metric to discrete levels based on an existing distributed K-means clustering method that helps maximize the similarity of the network metric in the same level, which therefore optimizes the matching between pairwise levels and the real-world pairwise proximity. Furthermore, for scalability reasons, HPM computes the pairwise levels with decentralized coordinates. Each node independently maintains its low-dimensional coordinate based on a novel decentralized implementation of the Maximum Margin Matrix Factorization method, which optimizes the mapping between the network metrics and the level values. Simulation results for the round trip time, bandwidth, loss, and hop count metric confirm that HPM converges fast, is robust to parameter settings, scales well with increasing levels or system size, and adapts well to diverse metrics. A prototype deployment on the PlanetLab platform shows that HPM not only converges fast, but also incurs modest bandwidth costs. Finally, applying HPM to optimize the service provision of hybrid clouds shows that HPM can achieve close to optimal solutions.

Hosting a multi-tier application using cloud-based hardware resources requires deploying components of the application stack across CPU resources to provide the application's infrastructure while considering factors such as scalability, fault tolerance, performance and deployment costs. In the paper titled "*Performance Implications of Multi-Tier Application Deployments on Infrastructure-as-a-Service Clouds: Towards Performance Modelling*"; Wes Lloyd et al. present results from an empirical study which investigates implications for application performance and resource requirements (such as CPU, storage and network) resulting from how multi-tier applications deployed to cloud environments. They investigate the implications of: (1) component placement across CPUs, (2) CPU memory size, (3) CPU hypervisor type (KVM vs. XEN), and (4) CPU placement across physical hosts (provisioning variation). All possible deployment configurations for two multi-tier application variants are tested. One application variant was computationally bound by the application middleware, the other bound by geospatial queries. The best performing deployments required as few as CPUs VMs, half the number required for CPU-level service isolation, demonstrating potential cost savings when components can be consolidated. Resource utilization (such as CPU time, disk I/O, and network I/O) varied with component deployment location, CPU memory allocation, and the hypervisor used (XEN or KVM) demonstrating how application deployment decisions impact required resources. Isolating application components using separate CPUs produced performance overhead of ∼1%–2%. Provisioning variation of CPUs across physical hosts produced overhead up to 3%. Relationships between resource utilization and performance were assessed using multiple linear regressions to develop a model to predict application deployment performance. Their model explained over 84% of the variance and predicted application performance with mean absolute error of only ∼.3 s with CPU time, disk sector reads, and disk sector writes serving as the most powerful predictors of application performance.

Recall that Cloud-hosted applications run on virtualized physical servers deployed across multiple data centres. The task or application request from the same or different group of end-users can be abstracted as a Virtual Network (VN) request, which are supported by the same underlying substrate network and thus share its resources. Therefore, efficient mapping techniques that intelligently use the substrate network resources are critical. Current research only considers the case when the VN requests are static. However, end-user demands and the corresponding VN requests can change dynamically. In their paper titled "*A Cost Efficient Framework and Algorithm for Embedding Dynamic Virtual Network Requests*", Gang, Hongfang, Vishal, and Lemin address the issue of how to optimally reconfigure and map an existing VN while the VN request changes. They first model this problem as a mathematical optimization problem with the objective of minimizing the reconfiguration cost by using mixed integer linear programming. Since the optimal problem is NP-hard, they also propose heuristic algorithms for solving it efficiently. They validate and evaluate their framework and algorithms by conducting extensive simulations on different realistic networks under various scenarios, and by comparing with existing approaches. Their simulation results show that the proposed approach outperforms existing solutions.

Hybrid cloud computing is gaining popularity among mobile users. The ABI Research predicts that the number of mobile

cloud computing subscribers is expected to grow from 42.8 million (1.1% of total mobile users) in 2008 to 998 million (19% of total mobile users) in 2014. Despite the hype achieved by mobile cloud computing, the growth of mobile cloud computing subscribers is still below expectations. According to the recent survey conducted by International Data Corporation, most IT Executives and CEOs are not interested to adopt such services due to the risks associated with security and privacy. The security threats have become a hurdle in the rapid adaptability of the mobile cloud computing paradigm. Significant efforts have been devoted in research organizations and academia to build secure mobile cloud computing environments and infrastructures. In spite of the efforts, there are a number of loopholes and challenges that still exist in the security policies of mobile cloud computing. In the paper titled "*Towards Secure Mobile Cloud Computing: A Survey*", Abdul, Laiha, Samee, and Sajjad et al. present the literature review that: (a) highlights the current state of the art work proposed to secure mobile cloud computing infrastructures, (b) identifies the potential problems, and (c) provides a taxonomy of the state of the art.

Hybrid computing environments is not only beneficial to SMEs and business applications, but also to scientists, who can now run data-intensive scientific applications by leveraging its vast storage and computation capabilities. During the scheduling of such scientific applications for execution, various computation data flows will happen between the controller and CPU resource instances. Amongst various QoS metrics, data security is always become one of the greatest concerns to scientists because their data may be intercepted or stolen by malicious parties during those data flows, especially for less secure hybrid cloud systems. An existing typical method for addressing this issue is to apply Internet Key Exchange (IKE) scheme to generate and exchange session keys, and then to apply these keys for performing symmetric-key encryption which will encrypt those data flows. However, the IKE scheme suffers from low efficiency due to its low performance of asymmetric-key cryptological operations over a large amount of data and high-density operations which are exactly the characteristics of scientific applications. In paper titled "*CCBKE -Session Key Negotiation for Fast and Secure Scheduling of Scientific Applications in Cloud Computing*", Chang, Xuyun, and Chi propose Cloud Computing Background Key Exchange (CCBKE), a novel authenticated key exchange scheme that aims at efficient security-aware scheduling of scientific applications. Their scheme is designed based on randomness-reuse strategy and Internet Key Exchange (IKE) scheme. Theoretical analyses and experimental results demonstrate that, compared with the IKE scheme, their CCBKE scheme can significantly improve the efficiency by dramatically reducing time consumption and computation load without sacrificing the level of security.

The last decade has witnessed an explosion of the interest in technologies of large simulation with the rapid growth of both the complexity and the scale of problem domains. Modelling & simulation of crowd is a typical paradigm, especially when dealing with large crowd. In the paper "*Hybrid Modelling and Simulation of Huge Crowd over a Hierarchical Grid Architecture*" Dan Chen et al. construct a hybrid computing infrastructure for simulation of evacuating tens of thousands of pedestrians in an urban area. The simulation infrastructure can facilitate a large crowd simulation comprising models of different grains and various types in nature. A number of agent-based and computational models residing at two distinctive administrative domains operate together, which successfully presents the dynamics of the complex scenario at scales of both individual and crowd levels. Experimental results indicate that the proposed hybrid modelling and simulation approach can effectively cope with the size and complexity of a scenario involving a huge crowd.

4. Conclusions

Hybrid computing environment is making rapid progress in acceptance by SMEs, government agencies, universities and scientists. Such computing environment cannot only reduce management costs, but is pivotal to enabling next generation of applications on unprecedented scales. Research and development efforts in hybrid computing environments are still at infancy. More tangible efforts are needed, if hybrid computing environments has to become primetime. Hybrid computing environments has potential to improve many application areas including scalable programming (such as HPC and multicore), future networking and general application provisioning. The researchers not only need to overcome technological challenges, but also legal, economic, environmental, and standardization challenges. In this special issue, we have selected some research papers that aim to address these challenges. We hope that the readers will find the articles of this special issue to be informative and useful.

References

[1] M. Armbrust, et al., A view of cloud computing, in: Communications of the ACM Magazine, Volume 53, Issue 4, ACM Press, New York, USA, 2010, pp. 50–58.
[2] L. Wang, R. Ranjan, J. Chen, B. Benatallah, Cloud Computing: Methodology, Systems, and Applications, Edited Book, CRC Press, Taylor and Francis Group. Published October 3, 2011, P. 575, ISBN: 978-1439856413.
[3] R.L. Grossman, Y. Gu, M. Sabala, W. Zhang, Compute and storage clouds using wide area high performance networksbe, Future Generation Computer Systems 25 (2) (2009) 179–183. Elsevier Press.
[4] J. Schad, J. Dittrich, J. Quian, Runtime measurements in the cloud: observing, analyzing, and reducing variance, Proceedings of VLDB Endowment 3 (1–2) (2010) 460–471.
[5] R. Buyya, R. Ranjan, R.N. Calheiros, Intercloud: utility-oriented federation of cloud computing environments for scaling of application services, in: Proceedings of the 10th International Conference on Algorithms and Architectures for Parallel Processing-Volume Part I, ICA3PP'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 13–31.
[6] K. Alhamazani, R. Ranjan, F. Rabhi, L. Wang, K. Mitra, Cloud monitoring for optimizing the qos of hosted applications, in: 4th IEEE International Conference on Cloud Computing Technology and Science, IEEE Computer Society, 2012, p. 6.
[7] M. Zhang, R. Ranjan, A. Haller, D. Georgakopoulos, P. Strazdins, Investigating decision support techniques for automating cloud service selection, in: 4th IEEE International Conference on Cloud Computing Technology and Science, IEEE Computer Society, 2012, p. 6.
[8] R. Ranjan, B. Benatallah, Programming cloud resource orchestration framework: operations and research challenges, - arXiv Preprint arXiv:1204.2204, 2012.
[9] P. Trunfio, D. Talia, H. Papadakis, P. Fragopoulou, M. Mordacchini, M. Pennanen, K. Popov, V. Vlassov, S. Haridi, Peer-to-peer resource discovery in grids: models and systems, Future Generation Computer Systems 23 (7) (2007) 864–878.
[10] R. Ranjan, L. Chan, A. Harwood, S. Karunasekera, R. Buyya, Decentralised resource discovery service for large scale federated grids, in: 2007 IEEE International Conference on e-Science and Grid Computing, IEEE Computer Society, 2007, pp. 379–387.
[11] Citrix OpenCloud Bridge – Extending Your Existing Datacenter to the Cloud. http://www.citrix.com/site/resources/dynamic/salesdocs/citrix_opencloud_bridge.pdf (accessed december 2012).

Rajiv Ranjan is a Scientist in the CSIRO ICT Center, Information Engineering Laboratory, Australian National University, Canberra, where he is working on projects related to cloud and service computing. Previously, he was a Senior Research Associate (Lecturer level B) in the School of Computer Science and Engineering, University of New South Wales (UNSW). Dr. Ranjan has a Ph.D. (2009) in Computer Science and Software Engineering from the University of Melbourne. He completed Bachelor of Computer Engineering from North Gujarat University, India, in 2002. Dr. Ranjan is broadly interested in the emerging areas of cloud, grid, and service computing. The main goal of his current research is to advance the fundamental understanding and state of the art of

provisioning and delivery of application services in large, heterogeneous, uncertain, and evolving distributed systems.

Dr. Ranjan has 54 research publications in journals with high impact factor (according to JCR published by ISI), in proceedings of IEEE's/ACM's premier conferences and in books published by leading publishers (7 Books, 18 journals, 24 conferences, and 5 book chapters). Dr. Ranjan has often been invited to served as Guest Editor for leading distributed systems and software engineering journals including Future Generation Computer Systems (Elsevier Press), Concurrency and Computation: Practice and Experience (John Wiley & Sons), and Software: Practice and Experience (Wiley InterScience). He was the Program Chair for 2010 and 2011 Australasian Symposium on Parallel and Distributed Computing and 2010 IEEE TCSC Doctoral Symposium. He serves as the editor of IEEE TCSC Newsletter. He has also recently initiated (as chair) IEEE TCSC Technical area on Cloud Computing.

Rajkumar Buyya is a Professor of Computer Science and Software Engineering; and Director of the Cloud Computing and Distributed Systems (CLOUDS) Laboratory at the University of Melbourne, Australia. He is also serving as the founding CEO of Manjrasoft Pty Ltd., a spin-off company of the University, commercializing its innovations in Grid and Cloud Computing. He has authored and published over 300 research papers and four text books. The books on emerging topics that Dr. Buyya edited include, High Performance Cluster Computing (Prentice Hall, USA, 1999), Content Delivery Networks (Springer, Germany, 2008), Market-Oriented Grid and Utility Computing (Wiley, USA, 2009), and Cloud Computing: Principles and Paradigms (Wiley, USA, 2011). He is one of the highest cited authors in computer science and software engineering worldwide.

Software technologies for Grid and Cloud computing developed under Dr. Buyya's leadership have gained rapid acceptance and are in use at several academic institutions and commercial enterprises in 40 countries across the world. Dr. Buyya has led the establishment and development of key community activities, including serving as foundation Chair of the IEEE Technical Committee on Scalable Computing and four IEEE conferences (CCGrid, Cluster, Grid, and e-Science). He has presented over 250 invited talks on his vision on IT Futures and advanced computing technologies at international conferences and institutions in Asia, Australia, Europe, North America, and South America. These contributions and international research leadership of Dr. Buyya are recognized through the award of "2009 IEEE Medal for Excellence in Scalable Computing" from the IEEE Computer Society, USA. Manjrasoft's Aneka technology for Cloud Computing developed under his leadership has received "2010 Asia Pacific Frost & Sullivan New Product Innovation Award". For further information on Dr. Buyya, please visit his cyberhome: www.buyya.com.

Surya Nepal is a Principal Research Scientist at CSIRO ICT Centre, Australia. His main research interest is in the development and implementation of technologies in the area of service-oriented architectures, web services, cloud computing, social networks and security, privacy and trust. He received his Ph.D. from RMIT University, Australia, ME from AIT, Thailand, and BE from NIT Surat, India. He has published several journal and conference papers in the areas of multimedia databases, web services and service-oriented architectures, and security, privacy and trust in collaborative environment, cloud computing and social networks. He is a programme committee member in many international conferences and has also edited books and special issues of many international journals.

Rajiv Ranjan*
CSIRO Information and Communication Technologies (ICT) Centre,
Information Engineering Laboratory,
Computer and Information Technology Building,
Australian National University,
North Road, Acton, ACT 2601, Australia
E-mail address: rranjans@gmail.com.

Rajkumar Buyya
Cloud Computing and Distributed Systems (CLOUDS) Laboratory,
Department of Computing and Information Systems,
The University of Melbourne, Australia

Surya Nepal
CSIRO Information and Communication Technologies (ICT) Centre,
Information Engineering Laboratory,
Computer and Information Technology Building,
Australian National University,
North Road, Acton, ACT 2601, Australia
19 January 2013
Available online 1 February 2013

* Corresponding editor.