Cluster Comput (2009) 12: 353–356
DOI 10.1007/s10586-009-0104-0

GUEST EDITORIAL

Recent Research Advances in e-Science

Xiaoyu Yang · Lizhe Wang · Gregor von Laszewski

Received: 14 July 2009 / Accepted: 2 September 2009 / Published online: 23 September 2009

Keywords e-Science · Grid computing · Cloud computing

1 e-Science overview

The term e-Science was initially coined by John Taylor, the Director General of the United Kingdom's Office of Science and Technology in 1999:

> "*e-Science is about global collaboration in key areas of science, and the next generation computing infrastructure will enable it*".

This definition is based on the recognition that many challenging scientific problems today are conducted in large teams potentially distributed globally, and need access to advanced and modern cyberinfrastructure to not only foster the collaborative aspects, but also to meet the computational, data, and network quality of service requirements needed to engage in such research endeavors.

The fundamental principle in e-Science is based on the trend that procedures and practices of traditional way in which science is conducted are undergoing radical change. This change is based on the inclusion of modern cyberinfrastructure as part of the science, or experiment environment which includes not only the already ubiquitous high end computers, storage and network infrastructure, but also emerging Web technologies. Together they formulate an essential infrastructure to support various research missions in many areas of science (e.g. particle physics, earth science, bio-informatics). This allows the exploration of previously unknown problems via simulation, generation and analysis of large amount of data, and global research collaboration.

e-Science is inherently interdisciplinary allowing and promoting synergistic activities between different scientific disciplines rather than just between a single discipline and computer science. Grid computing is one of the major enabling technology contributors to make the e-Science vision a reality. It also includes using parallel computing and various distributed computing technologies such as SOA, collaborative computing, workflow, ontology and semantic Web to develop middleware, services and applications. In the past few years, e-Science has made significant progress in areas such as supporting infrastructure, resource management and scheduling, data management, Grid middleware, and practice of e-Science in scientific and engineering disciplines. For this special issue we solicited papers that fit the theme "Recent Research Advances in e-Science". From the submitted papers, we included 6 papers that target the following research issues in e-Science:

- Infrastructure interoperability,
- Large scale data management,
- Simulation process automation, and
- e-Science practice and application development.

2 e-Science infrastructure and interoperability

Currently several production e-Science infrastructures have been deployed which are built upon different middleware, as summarised in the paper "*Research Advances by Using*

X. Yang (✉)
Earth Sciences Department, University of Cambridge,
Cambridge, UK
e-mail: kev.x.yang@gmail.com

X. Yang
e-mail: xy231@cam.ac.uk

L. Wang · G. von Laszewski
Service Oriented Cyberinfrastructure Laboratory, Rochester
Institute of Technology, Rochester, USA

Interoperable e-Science Infrastructures” [1] by Riedel et al. For example, different middleware is used by the following efforts:

- the Enabling Grids for e-Science (EGEE) [2] infrastructure uses the gLite [3];
- the TeraGrid [4] infrastructure and UK National Grid Service (NGS) [5] use the Globus middleware [6];
- the Distributed European Infrastructure for Supercomputing Applications (DEISA) [7] uses the UNICORE [8] middleware;
- the Open Science Grid (OSG) [9] uses the Virtual Data Toolkit (VDT) [10] (which also includes Globus);
- the NorduGrid [11] uses the Advance Resource Connector (ARC) middleware; and
- the Chinese CNGrid [12] uses Grid Operating System (GOS) [13] middleware.

However, according to [1] these e-Science infrastructures are not interoperable. In order to address the e-Science infrastructure interoperability issue, Riedel et al. have proposed an Infrastructure Interoperability Reference Model (IIRM) that is intended to increase the interoperability of production e-science infrastructures. Riedel et al. stated in the paper that the Open Grid Services Architecture (OGSA) concept is “too broad to be realistically implementable in today’s production e-science infrastructures”. IIRM does not aim to replace the OGSA model, but can be regarded as a trimmed down version of OGSA in terms of functionality and complexity. It has been practically demonstrated in targeting two e-health use cases. The Grid Interoperation Now (GIN) [14] community group of Open Grid Forum (OGF) has created a working group “Production Grid Infrastructure (PGI)” to investigate the standardization of the IIRM. This paper presents an overview of the approach and discusses how this research impacts the Grid standards.

3 e-Science data and information management

Data and Information Management is a problem that is becoming increasingly common as more computational resources are made available and more data is produced as part of large scale collaborative activities. One of the motivations for the data management is based on the fact that although computing resources are increasingly more powerful, they are nevertheless not cheap. The consequence is that there is a financial imperative to manage and re-use simulation data. Therefore new solutions to facilitate the running of simulations must have data archiving, discovery and access capabilities built in from the outset.

While many computing Grids supporting e-Science are now in production, e-Science data Grids and associated tools are required to become more mature in addition to several general purpose tools for building data Grids such as the Storage Resource Broker (SRB) and its successor iRODS [15]. Bui et al. in the paper “*Experience with BXGrid: A Data Repository and Computing Grid for Biometrics Research*” [16] report their experiences in building a biometric application data Grid, namely, “BXGrid”, which is currently in daily production use by active biometrics research group. The “BXGrid” is a typical e-Science application Grid developed by working together between an information systems research group and a biometrics research group. This paper summarizes groups’ experience in building an e-Science data Grid. It identifies ten practical lessons that worked for the group including e-Science application analysis, design and implementation such as “Get a prototype running right away”, “Ingest provisional data, not just archival data”, and “Don’t use an XML representation as an internal schema”, which have great value to e-Science practice and associated application development.

The paper “*Cell Approximation Method in Quorum Systems for Minimizing Access Time*” [17] provides an interesting discussion for data replication methods. A quorum consensus method is a valuable tool for managing replicated data and finding meaningful patterns from large amount of data [17]. The paper discusses two methods: Cell Approximation Major Voting (CAMV) and Cell Approximation Grid (CAG) in quorum systems for managing replicated data.

4 Simulation in e-Science

Grid-based simulation usually requires defining sequences of activities such as meta-scheduling, job submission, file transfer, analysis and simulation, and data harvesting. This results in the need for process automation, which can integrate the activities required for Grid-based computation without human interaction. Thus workflow technology can be used in e-Science for the automation of a process where documents, information or tasks are passed from one participant to another to be processed, according to a set of procedural rules. We include two papers in this special issue to report recent research in area of process automation and task coordination, and parallel task scheduling.

Workflow can provide high level abstractions required for controlling and coordinating the flow of tasks and data between distributed resources in e-Science. However, existing workflow languages (e.g. Taverna, Kepler) have their specific features hence provide potential challenges for modeling and comparing workflows across different systems [18]. Curcin et al. in the paper “*Analyzing scientific workflows with Computational Tree Logic*” present a generic process model that can be applied to any of these languages [18].

QoS assured resource allocation to the Grid users is currently receiving considerable attentions and resource reservation in advance is a mechanism which can assure the availability of resources over a period of time in the future. Zhang et al. propose a theoretical system model for scheduling offline parallel tasks in multi-cluster Grid with QoS guarantees based on this mechanism [19].

5 e-Science practice and application development

e-Science is interdisciplinary and brings together researchers from science and IT/computing background. This means e-Science application development is not trivial as it ought to be. The skills required include application of parallel computing and distributed computing, as well as the understanding of domain specific knowledge (e.g. earth sciences, bioinformatics) which is significant.

The paper "*Evolutionary Computation and Grid Computing to Optimize Nuclear Fusion Devices*" [20] uses evolutionary algorithms to address an important issue in the area of nuclear fusion reactor design. The combination of evolutionary algorithms and Grid computing is not novel, but the key value of this paper is that authors have managed to solve the problem of determining optimal reactor configurations through reducing computation time by an order of magnitude. It highlights a domain specific example in e-Science where multiple disciplines are used to find suitable solutions.

As an application Grid, the development of "BXGrid" [16] has identified 10 practical lessons in building an e-Science data Grid, which has been discussed in Sect. 3.

6 Discussions and outlook

Current production e-Science infrastructures (e.g. EGEE, TeraGrid) are at this time not interoperable, and each of them appears to be still isolated. Hence building a global e-Science infrastructure is becoming increasingly critical. The IIRM model described in this special issue may provide a starting point for a simplified model to contribute to building a global e-Science infrastructure which interconnects the existing regional production e-Science infrastructures.

Cloud computing and virtualization technology advances also raise challenging issues to e-Science. According to Wang and von Laszewski [21], Cloud computing can be defined as "*A computing Cloud is a set of network enabled services, providing scalable, QoS guaranteed, normally personalized, inexpensive computing platforms on demand, which could be accessed in a simple and pervasive way*". Cloud computing has many differences to Grid computing technology. For example, the Grid infrastructure in nature is a decentralized system which contains heterogeneous resources (e.g. hardware/software configurations, access interfaces and management). It can span across geographically distributed sites and lack central control. Computing Clouds operate like a central compute server with single access point. Cloud infrastructures could span several computing centers, which contain homogeneous resources operated under central control [21]. The next step of e-Science will accommodate using Cloud infrastructure.

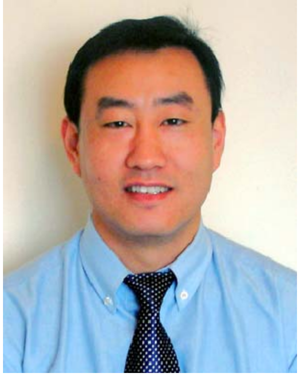
The Future Internet [22] is emerging and will have a great impact on addressing challenging issues faced by e-Science (e.g. petascale data management, next generation Grid computing). The current Internet was designed in 1970 and it has grown beyond its original expectations. This can be attributed to many factors including increased demand for performance, availability, security, and reliability. It gradually reaches a set of fundamental technological limits and is impacted by operational limitations imposed by its original design objectives and architecture. The concept of Future Internet consists of four pillars which are: (i) Internet by and for people, (ii) Internet of Contents and Knowledge, (iii) Internet of Things, and (iv) Internet of Services [22]. The Future Internet infrastructure will provide a common foundation to these pillars. We believe the Future Internet can help address current challenging issues that e-Science is facing, but will also raise additional challenges to e-Science.

Acknowledgements We would like to thank authors for their contributions, including those whose papers are not included in this special issue. We also would like to acknowledge thoughtful work from many reviewers who provided valuable evaluations and recommendations.

References

1. Riedel, M., Wolf, F., Kranzlmuller, D., Strelt, A., Lippert, T.: Research advances by using interoperable e-Science infrastructures, Clust. Comput. J. (2009) Special Issue Recent Research Advances in e-Science
2. EGEE portal, http://www.eu-egee.org/
3. gLite, http://glite.web.cern.ch/glite/
4. TeraGrid, http://www.teragrid.org/
5. National Grid Service, http://www.grid-support.ac.uk/
6. Globus, http://www.globus.org/
7. DEISA, http://www.deisa.eu/
8. UNICORE, http://www.unicore.eu/
9. Open Science Grid (OSG), http://www.opensciencegrid.org/
10. Virtual Data Toolkit, http://vdt.cs.wisc.edu/
11. Nordu Grid project, http://www.nordugrid.org/
12. China National Grid project, http://www.cngrid.org
13. Xu, Z., Li, W.: Vega Grid: a computer systems approach to grid research. In: Lecture Notes in Computer Science, vol. 3032, pp. 480–486. Springer, Berlin (2004)
14. Grid Interoperation Now, http://www.ogf.org/gf/group_info/view.php?group=gin-cg
15. Rajasekar, A., Wan, M., Moore, R., Schroeder, W.: A prototype rule-based distributed data management system. In: HPDC Workshop on Next Generation Distributed Data Management, May, 2006

16. Bui, H., Kelly, M., Lyon, C., Pasquier, M., Thomas, D., Flynn, P., Thain, D.: Experience with BXGrid: a data repository and computing grid for biometrics research. Clust. Comput. J. (2009) Special Issue Recent Research Advances in e-Science
17. Lee, Y.J., Kim, H.Y., Lee, C.H.: Cell approximation method in quorum systems for minimizing access time. Clust. Comput. J. (2009) Special Issue Recent Research Advances in e-Science
18. Curcin, V., Ghanem, M.M., Guo, Y.: Analysing scientific workflows with computational tree logic. Clust. Comput. J. (2009) Special Issue Recent Research Advances in e-Science
19. Zhang, J., Luo, J.: A comparison of utility-oriented algorithms for scheduling parallel tasks in multi-cluster grid. Clust. Comput. J. (2009) Special Issue Recent Research Advances in e-Science
20. Antonio, G.I., Miguel, A.V., Francisco, C.M., Miguel, C.M., Enrique, M.R.: Evolutionary programming and grid computing to optimise nuclear fusion devices. Clust. Comput. J. (2009) Special Issue Recent Research Advances in e-Science
21. Wang, L., von Laszewski, G., Kunze, M., Tao, J.: Cloud computing: a perspective study, J. New Gener. Comput. (2009, to be published)
22. Future, Internet.: Available at www.future-internet.eu/fileadmin/documents/reports/Cross-ETPs_FI_Vision_Document_v1_0.pdf



Xiaoyu Yang is a post-doctoral Research Associate in the Department of Earth Sciences at the University of Cambridge, UK. He is also an affiliated Software Engineer in Cambridge e-Science Centre, and a Senior Member of Wolfson College, University of Cambridge. His technical interests include SOA, systems engineering, Grid computing & e-Science, Cloud computing, service-oriented workflow, and product lifecycle information management. He has an M.Sc. degree in IT and a Ph.D. degree in Systems Engineering from De Montfort University, UK. Contact him at kev.x.yang@gmail.com or xy231@cam.ac.uk.

Lizhe Wang is an Assistant Director of Service Oriented Cyberinfrastructure Laboratory at Rochester Institute of Technology. Dr. Wang got his bachelor and master degree from Tsinghua University, China and doctoral degree from University Karlsruhe, Germany. Dr. Wang's research interests include cluster computing, Grid computing, Cloud computing and Green computing.

Gregor von Laszewski currently is the Director of the Service Oriented Cyberinfrastructure Laboratory at Rochester Institute of Technology (RIT), an associate professor of the Ph.D. program at RIT and holds a guest appointment in the computer science department. He worked between 1996 and 2007 for Argonne National Laboratory where he was last a scientist and a fellow of the Computation Institute at University of Chicago. He received a Masters Degree in 1990 from the University of Bonn, Germany, and a Ph.D. in 1996 from Syracuse University in computer science. His research interests include Grid computing, e-Science and Cloud computing.