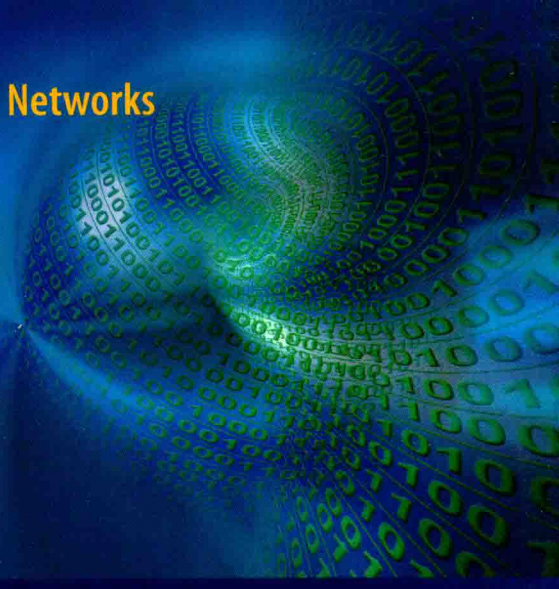
Computer Communications and Networks



Xiaoyu Yang
Lizhe Wang
Wei Jie *Editors*

Guide to e-Science

Next Generation Scientific Research and Discovery

Springer

Xiaoyu Yang • Lizhe Wang • Wei Jie
Editors

Guide to e-Science

Next Generation Scientific Research and Discovery

Springer

Editors
Dr. Xiaoyu Yang
Reading e-Science Centre
Harry Pitt Building
University of Reading
3 Earley Gate, Whiteknights
Reading, RG6 6AL
UK
kev.x.yang@gmail.com

Dr. Lizhe Wang
Pervasive Technology Institute
Indiana University
2719 East 10th Street
Bloomington, IN 47408
USA
lizhe.wang@gmail.com

Series Editor
Professor A.J. Sammes
Bsc, Mphil, PhD, FBCS, CEng
Centre for Forensic Computing
Cranfield University
DCMT, Shrivenham
Swindo SN6 8LA, UK

Dr. Wei Jie
Fac. Professional Studies
Thames Valley University
School of Computing
St. Mary's Road TC372
Ealing, London W5 5RF
United Kingdom
wei.jie@tvu.ac.uk

ISSN 1617-7975
ISBN 978-1-4471-2658-4 ISBN 978-0-85729-439-5 (eBook)
DOI 10.1007/978-0-85729-439-5
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Cover design: deblik

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Computer Communications and Networks

For other titles published in this series, go to
www.springer.com/series/4198

The **Computer Communications and Networks** series is a range of textbooks, monographs and handbooks. It sets out to provide students, researchers and non-specialists alike with a sure grounding in current knowledge, together with comprehensible access to the latest developments in computer communications and networking.

Emphasis is placed on clear and explanatory styles that support a tutorial approach, so that even the most complex of topics is presented in a lucid and intelligible manner.

Foreword

The way we carry out scientific research is undergoing a series of radical changes as a result of the digital revolution. Traditional scientific approaches are finding it increasingly difficult to solve complex problems and grand challenges without broadening horizons to exploit the use of new digital technologies that support collaborative working on one hand, and facilitate clever approaches to working with the huge quantities of data that can now be generated through modern experimental systems and computer simulations performed on supercomputers and grid computing facilities. We can throw into the mix of new emphasis on data sharing and open data, with access provided by web service technologies, new approaches to constructing and sharing workflow methods, and Web 2.0 technologies that enable communities, large and small, to document and share research outcomes and new knowledge. Simultaneously we are seeing greater emphasis on researchers sharing their primary data and analysis tools to accompany publication, enabling other researchers to reproduce and re-use their results. We are already past the point where researchers can analyse data by hand in the point-by-point that used to characterise research, which means that automated computational and data analysis systems need to be robust and flawlessly accurate. These requirements call for a comprehensive supporting cyberinfrastructure for modern scientific research, and e-Science will play an important role in addressing this challenge.

At the outset of the UK e-Science programme, the Director General of the Research Councils, John Taylor, wrote, "e-Science is about global collaboration in key areas of science and the next generation computing infrastructure that will enable it." This definition recognised that challenging scientific problems will increasingly be addressed by large teams working in different institutes – even in different countries – who will need access to a comprehensive cyberinfrastructure that should not only give access to massive computing capabilities, but also enable easy sharing of data and information in an era where huge quantities of data can be generated, and facilitate personal and team interactions required for genuine collaborative working.

This edited book comprises chapters authored by international e-Science experts and practitioners, presenting readers with e-Science practices and applications of how various technologies and tools can be employed to build essential infrastructures

to support a new generation of scientific research. The book is organised by grouping the features of modern scientific research into five themes: (1) Sharing and Open Research; (2) Data-intensive e-Science; (3) Collaborative Research; (4) Automated Research, Reusability, Reproducibility and Repeatability; and (5) e-Science: easy Science. The different chapters within each topic provide introductions, descriptions and discussions of relevant e-Science methodologies, architectures, tools, systems, services and frameworks that are designed to address a range of different requirements. The expert authors also share their experiences and lessons learned in their e-Science work.

The book is a timely contribution to e-Science communities. I believe readers, especially researchers and developers of successive generations of e-infrastructure, will find this book useful.

Department of Earth Sciences
University of Cambridge
UK

Professor Martin Dove

Preface

Next generation scientific research has radically changed the way in which science is carried out. With the assistance of modern e-infrastructure as part of science, or experiments including high-performance computing capabilities, large-capacity data storage facilities and high-speed network infrastructure, the exploration of previously unknown problems can now be solved via simulation, generation and analysis of large amount of data, sharing of geographically distributed resources (e.g. computing facilities, data, script, experiment plan, workflow) and global research collaboration.

A term "science 2.0" or "research 2.0" is now emerging which has outlined the features of next generation scientific research: large-scale, team-based global research collaboration; open research and open data; collective intelligence; knowledge and resources sharing; etc. This needs an advanced environment and e-Science infrastructure which can meet the new features and requirements. For example, in collaborative research that involves diversity of participants, security, trust and privacy are becoming increasingly essential; faced with the deluge of scientific data, diverse data integration, heterogeneous data interoperability, domain information and knowledge representation and retrieval, linked data and structured data presents critical challenges; resource sharing involves the exchange, selection and aggregation of geographically distributed resources and the development of innovative and high-performance oriented applications; automatic, reproducible, reusable and repeatable research concerns the auto-coordination of various tasks involved in a research study, and provenance of the research result.

According to the features and requirements for next generation scientific research, this book is structured into five themes, which demonstrate how e-Science methods and techniques can be employed to facilitate next generation scientific research and discovery from the following aspects:

- Part I: Sharing and Open Research
- Part II: Data-Intensive e-Science
- Part III: Collaborative Research
- Part IV: Research Automation, Reusability, Reproducibility and Repeatability
- Part V: e-Science, Easy Science

Part I: Sharing and Open Research

"Science has always been a social process." Next generation scientific research promotes the open research and sharing of resources and knowledge. Grid computing technology is one of the key enabling technologies in resource sharing which aims at the synergy of the distributed high computing resources. Grid middleware, such as Globus Toolkit, gLite and UNICORE, provides an effective approach to share usually the high-performance computing resources (e.g. super computers, clusters). This formulates the mainstream of grid computing in e-Science. However, high-performance computing is not always available in many institutions; on the other hand, there exists much idle computing power. In order to address this need, we can also use peer-to-peer (P2P) grid computing technology to build a P2P Grid for the sharing of idle computing resources. For example, in the P2P Grid, labs can donate their idle computational resources in exchange for accessing other labs' idle resources when necessary. Part I contains four chapters which discuss the use of mainstream grid computing and the P2P Grid in e-Science.

Chapter 1 describes the development of an e-infrastructure which integrates the data grid and computing grid to facilitate the hydrology environmental science. It allows a wide range of hydrological problems to be investigated and is particularly suitable for either computationally intensive or multiple scenario applications. The chapter discusses the grid infrastructure system integration and development, visualisation of geographic information from grid outputs and implementation of hydrological simulations based on the infrastructure. Also, the chapter investigates the adaption of cloud computing into scientific research by extending the computing grid to utilise the Amazon EC2 cloud computing resources. Users can carry out a complete simulation job from job submission to data management and metadata management based on the tools available in the infrastructure.

Chapter 2 discusses the current state and future perspective of the German National Grid Initiative (NGI), namely, D-Grid. It describes the current D-Grid e-Infrastructure in detail, and provides a discussion on how D-Grid's future may look like with virtualisation and cloud computing striding ahead. Particularly, it discusses the incorporation of service level agreements (SLA) to allow D-Grid service providers to deliver the service level objectives (SLO) assured services to service customers.

In Chap. 3, the authors share their experience in developing a P2P grid middleware called OurGrid and deploying it to build the OurGrid Community. The chapter describes the mechanisms that effectively promote collaboration and allow the assemblage of large P2P Grids from the contributions of thousands of small sites. The authors present a successful case study of using OurGrid, and summarise their lessons learned and experience.

Chapter 4 introduces an approach to grid and overlay network research for the sharing of computing resources to enable the simulation of increased complexity and speed. It presents a Peer4Peer platform that provides the main infrastructure for efficient peer-to-peer simulation over the Internet.

Part II: Data-Intensive e-Science

Scientific research can be regarded in some sense as activities around a data lifecycle (i.e. acquisition, transfer, storage, analysis/data mining, visualisation). It is now increasingly facing the challenges of data explosion. For example, the high-energy physics experiment of the large hardon collider (LHC) at CERN in 2007 produced a stream of data at 300 MB/s, which is equivalent to a stack of CDs as high as the Eiffel Tower every week. Modern sciences have stronger demands for effective data curation and management than ever before. In Part II, we include three chapters, which present different methods for data management in e-Science.

Chapter 5 discussed how the NASA Jet Propulsion Laboratory in California, USA, developed successful science data systems for highly distributed communities in physical and life sciences that require extensive sharing of distributed services and common information models based on common architectures. It demonstrated that a well-defined architecture and set of accompanied software can vastly improve the ability to develop roadmaps for the construction of virtual science environments.

In Chap. 6, the authors develop a tool integration framework, namely, Galaxy, which enables advanced data analysis that requires no informatics expertise. The Galaxy tool has also been deployed as Amazon Web Service in the Amazon Cloud for open access.

Large-scale cross-disciplinary scientific collaborations usually involve diverse data integration, heterogeneous data interoperability and domain information and knowledge representation and retrieval. Chapter 7 proposes an integrated ontology management and data sharing framework which builds upon the advancements in object-oriented database design, semantic web and service-oriented architecture to form the key data-sharing backbone. The framework is implemented to cater data-sharing needs for large-scale sensor deployment from disparate scientific domains. This enables each participating scientific virtual organisation (VO) to publish, search and access data across the e-infrastructure in a service-oriented manner, accompanied by domain-specific knowledge.

Part III: Collaborative Research

Collaborative research is a distinct feature in modern scientific research. Interoperability, security, trust and privacy are key elements in collaborative research. Part III contains four chapters that discuss the aspect of collaborative environment, e-infrastructure interoperability and security, trust and privacy.

Chapter 8 proposes a collaborative environment to support the scientific processes of a solar-enabled water production and recycling application. This environment can perform complex tasks such as distributed instrument control, data collection from heterogeneous sources, data archival, data analysis and mining,

data visualisation and decision support. It aims to address the issues of archival and management of multidimensional data from heterogeneous sensors and instruments, allowing efficient data sharing among different groups of scientists who are not computer experts, allowing different parties involved to publish and consume data and process, real-time decision support with control, etc.

Chapter 9 investigates challenges and provides proven solutions in the context of e-Science infrastructure interoperability. The chapter illustrates how an increasing amount of e-Scientists can take advantage of using different types of e-Science infrastructures jointly together for their e-research activities, and proposes seven steps towards interoperability for e-Science.

e-Infrastructure based on distributed computing could result in malicious intervention resulting in theft of the models and data that have significant commercial value. In order to tackle this problem, Chap. 10 proposes two distributed systems, one applicable for a computational system and the other for a distributed data system. A *configuration resolver* is used to maintain a list of trustworthy participants available in the virtual organisation. Users can then submit their jobs to the *configuration resolver*, knowing that their jobs will be dispatched to trustworthy participants and executed in protected environments. This security model was tested in the UK National Grid Service (NGS), and the performance overhead was measured.

Cloud computing has become a new computing paradigm as it can provide scalable IT infrastructures, QoS-assured services and customisable computing environments. As a new computing paradigm, cloud computing provides new methods and techniques for e-Science. However, cloud computing introduces new challenges with respect to security mainly caused by the unique characteristics inherited via virtual machine technology. Chapter 11 focuses on the challenges imposed on intrusion diagnosis for clouds. This chapter identifies the importance of intrusion diagnosis problem for clouds and the new challenges for this intrusion diagnosis. An appropriate solution is proposed to address these challenges and is demonstrated to be effective by empirical evaluation.

Part IV: Research Automation, Reusability, Reproducibility and Repeatability

A research study usually involves a series of tasks. For example, even a simple research study may involve three tasks which are data acquisition, data analysis and visualisation. In modern scientific research involving global collaboration, resource sharing and data deluge, a research study may include multiple complex tasks. For example, a computational experiment that spans multiple geographically distributed computation resources and analytical models involves sequences of tasks such as resource discovery, job submission, file staging, simulation, data harvesting and visualisation. Research process automation with less direct human control and research traceability are vital in scientific research. In order to address this need, workflow technology can be employed 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. A scientific workflow integrating all tasks required for a research study can be reusable, and the output from a scientific workflow can be reproducible and repeatable. This part includes four chapters discussing the scientific workflow, which will give you a guide of how workflow technologies are used in e-Science.

Chapter 12 discusses the requirements on scientific workflows, the state of the art of scientific workflow management systems as well as the ability of conventional workflow technology to fulfil requirements of scientists and scientific applications. In order to overcome the disadvantages of the conventional workflow, authors proposed a conceptual architecture for scientific workflow management systems based on the business workflow technology as well as extensions of existing workflow concepts. This can improve the ability of established workflow technology in scientific simulations.

Chapter 13 discusses the integration of *Kepler* workflow system into the *University of California Grid*. This architecture is being applied to a computational enzyme design process. The implementation and experiments validated how the Kepler workflow system can make the scientific computation process automated, pipelined, efficient, extensible, stable and easy to use.

Chapter 14 concerns the quality of a scientific workflow by service level agreement (SLA). This chapter describes related concepts and mapping algorithms to facilitate the resource reservation at each grid site and the user providing the estimated runtime of each sub-job correlated with a resource configuration. In particular, it describes several sub-optimisation algorithms to map sub-jobs of the workflow to the grid resources within an SLA context.

Considering the differences of using workflow in scientific research and the business world, where scientific workflows need to consider specific characteristics and make corresponding changes to accommodate those characteristics, Chap. 15 proposes a task-based scientific workflow modelling and performing approach for orchestrating e-Science with the workflow paradigm.

Part V: e-Science, Easy Science

This part mainly focuses on the application of e-Science in certain science domains.

Chapter 16 presents a robust face recognition technique based on the extraction of scale invariant feature transform (SIFT) features from the face areas. This technique has the potential to be employed in an e-infrastructure of a face recognition system with ATM cash machines.

Chapter 17 presents a framework for the metamodel-driven development of open grid services architecture (OGSA) based service-oriented architecture (SOA) for collaborative cancer research in the CancerGrid project. The authors extend the existing Z model and the generation technology to support OGSA in a distributed

collaborative environment. They built a generic SOA model combining the semantics of the standard domain metamodel and metadata, and the Web services resource framework (WSRF) standard. This model can then employed to automate the generation of the trial management systems used in cancer clinical trials.

The last chapter, Chap. 18 the introduces e-Science practice and application in the Computer Network Information Centre (CNIC), Chinese Academy of Science (CAS). This chapter introduces the information infrastructures supporting scientific research from five aspects, which include digital network and communication infrastructure, high-performance computing environment, scientific data environment, digital library and virtual laboratory. CAS proposed an e-Science model, and stated that e-infrastructure should apply information infrastructure and digital technology in every aspect of research activity to enable better research and advanced research patterns. The chapter also presents a collaborative, environmental and biological e-Science application conducted in the Qinghai Lake region, Tibetan Plateau, to show how various information and communication technologies can be employed to facilitate scientific research, providing a cyberinfrastructure for protecting wildlife and ecological environment and decision making. CNIC realised that e-Science is the way leading to next generation scientific research, and has been promoting e-Science practice and application systematically: by e-Science, to easy Science.

Reading e-Science Centre
University of Reading,
UK

Dr. Xiaoyu Yang

Acknowledgements

We would like to thank the authors for their contributions, including those whose chapters are not included in this book.

We would like to express our gratitude to the Editorial Advisor Board members, Professor Martin Dove (University of Cambridge, UK), Dr. Andrew Martin (University of Oxford, UK) and Mr. Morris Riedl (Jülich Super Computing Centre, Germany), for their support and contributions to this book.

We also would like to acknowledge thoughtful work from many reviewers who provided valuable evaluations and recommendations.

Our special thanks go to Mr. Simon Rees and Mr. Wayne Wheeler from Springer for their assistance in the preparation of the book.

About the Editors

Xiaoyu Yang completed his postdoctoral research in e-Science at Earth Sciences Department of University of Cambridge, UK in 2008. He is currently working in Reading e-Science Centre, University of Reading, UK. Dr. Yang has research interests which include e-Science/e-Research, geoinformatics, Grid/Cloud computing, and distributed computing, etc. He worked in School of Electronics and Computer Sciences in University of Southampton, UK after his postdoctoral research at University of Cambridge. He earned MSc Degree in IT in 2001 and PhD degree in Systems Engineering in 2006 at De Montfort University, UK.

Lizhe Wang received his Doctor of Engineering from University Karlsruhe, Germany. He is currently a principal research engineer at Pervasive Technology Institute (PTI), Indiana University, USA. Dr. Wang's research interests include cluster and Grid computing, Cloud computing, multi-core system and energy-aware computing. He has published 3 books and more than 40 scientific papers. Dr. Lizhe Wang received his Bachelor of Engineering with honors and Master of Engineering both from Tsinghua University, China.

Wei Jie was awarded PhD in Computer Engineering from Nanyang Technological University (Singapore). He is currently a lecturer in computing at Thames Valley University (UK). Dr. Jie has been actively involved in the area of parallel and distributed computing for many years, and published about fifty papers in international journals and conferences. His current research interests include Grid computing and applications, security in distributed computing, parallel and distributed algorithms and languages, etc. He received his BEng and MEng in Beijing University of Aeronautics and Astronautics (China).

Contributors

Enis Afgan
Department of Biology and Department of Mathematics & Computer Science, Emory University, Druid Hills, GA, USA
eafgan@emory.edu

Ilkay Altintas
San Diego Supercomputer Center, UCSD, 9500 Gilman Drive, MC 0505, La Jolla, CA 92093, USA

Jörn Altmann
School of Information Technology, International University in Germany, Campus 3, 76646 Bruchsal, Germany
jorn.altmann@acm.org

Nazareno Andrade
Departamento de Sistemas e Computação, Laboratório de Sistemas Distribuídos, Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brazil
nazareno@dsc.ufcg.edu.br

Junaid Arshad
School of Computing, University of Leeds, Leeds LS2 9JT, UK
sc06ja@leeds.ac.uk

Dannon Baker
Department of Biology and Department of Mathematics & Computer Science, Emory University, Druid Hills, GA, USA

Hock Beng Lim
Intelligent Systems Center, School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore
limhb@ntu.edu.sg

C. Isabella Bovolo
School of Civil Engineering and Geosciences, University of Newcastle upon Tyne, Newcastle upon Tyne NE7 7RU, UK

Francisco Brasileiro
Departamento de Sistemas e Computação, Laboratório de Sistemas Distribuídos, Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brazil
fubica@dsc.ufcg.edu.br

Radu Calinescu
Aston University, Birmingham B4 7ET, UK

Chee Keong Chan
Intelligent Systems Center, School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore
eckchan@ntu.edu.sg

Jinjun Chen
Faculty of Information and Communication Technology, Swinburne University of Technology, Melbourne, Australia
jinjun.chen@gmail.com

Gen-Tao Chiang
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
gtc@sanger.ac.uk

Fook Hoong Choo
Intelligent Systems Center, School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore
efhchoo@ntu.edu.sg

Nate Coraor
Huck Institutes of the Life Sciences and Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, USA

Daniel Crawl
San Diego Supercomputer Center, UCSD, 9500 Gilman Drive, MC 0505, La Jolla, CA 92093, USA

Daniel J. Crichton
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA
daniel.j.crichton@jpl.nasa.gov

Wanchun Dou
State Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing 210009, China
douwc@nju.edu.cn

Martin T. Dove
Department of Earth Sciences, University of Cambridge, Cambridge CB2 3EQ, UK