ELSEVIER

Future Generation Computer Systems 18 (2002) v–viii

FGCS
FUTURE
GENERATION
COMPUTER
SYSTEMS

www.elsevier.com/locate/future

Guest editorial
Cluster computing

Rajkumar Buyya[a,1], Hai Jin[b,*,2], Toni Cortes[c,3]

[a] *School of Computer Science and Software Engineering, Monash University, Melbourne, Australia*
[b] *College of Computer, Huazhong University of Science and Technology, Wuhan 430074, PR China*
[c] *Department d'Arquitectura de Computadors, Universitat Politècnica de Catalunya, Barcelona, Spain*

Cluster computing can be described as a fusion of the fields of parallel, high-performance, distributed, and high-availability computing. Cluster computing has become a hot topic of research among academic and industry community including system designers, network developers, language designers, standardizing forums, algorithm developers, graduate students and faculties. The use of clusters as computing platform is not just limited to scientific and engineering applications; there are many business applications that can benefit from the use of clusters. There are many exciting areas of development in cluster computing with new ideas as well as hybrids of old ones being deployed for production as well as research systems. The aim of this special issue is to bring together original and latest work from both academia and industry on various issues related to cluster computing.

This research, and development, is being done in many areas but there are a few that are of special interest. The first one is, with no doubt, the network. Clusters are based on the communication between nodes and designing fast and low latency networks is a must for clusters to become the configuration of the future. Plenty of work being done in this are in both hardware (Infiniband, SCI, Myrinet, QSnet) and software (Low latency protocols such as VIA).

Allowing heterogeneity in both hardware and software (OS) is extremely important and is becoming one of the key issues in cluster research. This heterogeneity allows a better expandability (i.e. use old and new components) and scalability and has to be addressed at many levels. We need tools that allow us to develop applications for these configurations, as well as scheduling systems that offer mechanisms for exploitation of heterogeneity and that deliver better application performance. The papers selected for inclusion in this special issue address these issues.

File systems are also of great interest as clusters are being used for I/O intensive applications such as e-commerce, WEB servers or databases. This kind of file systems tight the I/O devices as well as offer high performance and a single-system image. Achieving these objectives is not easy and plenty of work is being devoted.

Clusters also provide an excellent platform for solving a range of parallel and distributed applications in both scientific and commercial areas. For scientific application, clusters can be used in grand challenge or supercomputing applications, such as earthquakes or hurricanes prediction, complex crystallographic and microtomographic structural problems, protein dynamics and biocatalysis, relativistic quantum chemistry of actinides, virtual materials design and

[*] Corresponding author.
E-mail addresses: rajkumar@csse.monash.edu.au (R. Buyya), hjin@hust.edu.cn (H. Jin), toni@ac.upc.es (T. Cortes).
[1] http://www.buyya.com.
[2] http://ceng.usc.edu/~hjin.
[3] http://www.ac.upc.es/homes/toni.

0167-739X/02/$ – see front matter © 2002 Published by Elsevier Science B.V.
PII: S0167-739X(01)00053-X

processing including crash simulations, and global climate modeling. For the commercial applications, cluster can be best used in e-commerce as superserver, which consolidate web server, ftp server, e-mail server, database server, etc. Clusters can also be used in data mining applications to provide the storage and data management services for the data set being mined and computational services required by the data filtering, preparation and mining tasks. Other commercial applications includes image rendering, network simulation, etc.

Due to the growing interest in cluster computing, the IEEE Task Force on Cluster Computing (TFCC) [8] was formed in early 1999. TFCC is acting as a focal point and guide to the current cluster computing community and has been actively promoting the field of cluster computing with the aid of a number of novel projects. With the support of TFCC, a series of special sessions, workshop, symposiums and conferences were held to guide R&D work both in academic and industrial settings. Further information on TFCC activities can be accessed from the web site: http://www.ieeetfcc.org/.

This special issue emerges from the seven best papers that we selected from the three conferences in the area of cluster computing. They are the first IEEE International Workshop/Conference on Cluster Computing[4] (Melbourne, Australia), CC-TEA 2001 (Las Vegas, USA), and the Asia–Pacific International Symposium on Cluster Computing (Beijing, China). The last two meetings have been merged to create a major IEEE international conference series (called CCGRID[5]) that focus on both cluster computing and grid[6] technologies. Grid technology will help in coupling multiple clusters in the same or different organizations to create computational grids (federated or hyper clusters) for solving large-scale multidisciplinary problems.

The papers that we have selected for inclusion in this special issue have received the higher review ratings for their research contributions. All the papers are extended and revised to reflect the latest research achievements in various areas of cluster computing. The topics of the papers cover different aspects of cluster computing ranging from lightweight (low-latency and high-bandwidth) communication protocols to highly optimized sorting algorithms implemented on clusters. Topic areas covered include rapid application development environments, operating system kernels, file systems, load-balancing mechanisms, communication subsystems, industry standard oriented user-level communication interfaces, and a number of parallel sorting schemes. The works presented have addressed mechanisms for handling heterogeneity in clusters, as this is an important issue for handling design issues such as expandability of clusters.

The first paper entitled "*MPI–Delphi: An MPI Implementation for Visual Programming Environments and Heterogeneous Computing*" presents authors' efforts to integrate MPI for parallel programming with Delphi visual programming environment. MPI–Delphi interface makes it possible to manage a cluster of heterogeneous PCs. This interface is also suitable for some specific kind of programs, such as monitoring long execution parallel programs, or computational intensive graphical simulations.

"*PODOS — The Design and Implementation of a Performance Oriented LINUX Cluster*" reports the design issues of a performance-oriented operating system, PODOS, which harness the performance capabilities of a cluster-computing environment. PODOS added four new components to existing LINUX operating system, they are Communication Manager, PODOS File System, Resource Manager, and Global Inter-Process Communication. The custom designed communication protocol uses round-robin Transmission-Groups mechanism to multiplex packets across multiple network interfaces. PODOS file system builds an efficient file sharing environments on top of high-speed communication subsystem.

"*On a Scheme for Parallel Sorting on Heterogeneous Clusters*" discusses the parallel sorting algorithms and their implementations suitable for cluster architectures in order to optimize cluster resources. By carefully studying the various sorting algorithms on heterogeneous cluster, the authors concluded that the software challenges to better utilize the heterogeneous cluster platform are data decomposition techniques, scheduling and load balancing methods. The algorithm described in this paper combines very good property for load balancing.

[4] http://www.clustercomp.org.

[5] http://www.ccgrid.org.

[6] http://www.gridcomputing.com.

"*Cluster File Systems: A Case Study*" presents a scalable single-image file system, called COSMOS, designed for Dawning 2000 superserver. COSMOS provides location transparency and strong UNIX file-sharing semantics. COSMOS uses serverless design and introduces a dual-granularity cooperative caching. Other characteristics of COSMOS includes heuristic replacement algorithm, network disk striping, and distributed meta-data management. COSMOS provides high I/O bandwidth for large files and good I/O performance for small files and directory reads.

"*Load Balancing for Heterogeneous Clusters of PCs*" discusses the authors' experience of using the asymmetric load balancing approach in heterogeneous cluster environments. By using the LU benchmark from the NPB family, authors find that asymmetric load balancing approach is a general-purpose tool that can be used with any data decomposed regular problem, and with some extensions, it can be also used in irregular problem.

"*Directed Point: A Communication Subsystem for Commodity Supercomputing with Gigabit Ethernet*" studies the practical issues on the design of a new high performance communication subsystem, called Directed Point (DP). The DP abstraction model depicts the communication channels built among a group of communicating processes. It supports both point-to-point communication and various types of group operations. The API of DP combines features from BSD sockets and MPI to facilitate the peer-to-peer communication in a cluster. DP improves the communication performance by reducing protocol complexity through the use of directed message, by reducing the intermediate memory copies between protocol layers through the use of token buffer pool, and by reducing the context switching and scheduling overhead through the use of light-weight messaging calls.

"*VI Architecture Communication Features and Performance on the Giganet Cluster LAN*" presents the performance result study of Giganet cLAN VI architecture hardware implementation. VI architecture aims to close the performance gap between the bandwidths and latencies provided by the communication hardware and visible to the application by minimizing the software overhead on the critical path of communication. The focus of this study is to assess and compare the performance for different VIA data-transfer modes and specific features that are available to higher-level communication software like MPI. The features investigated in this paper include the use of send/receive vs. RDMA data transfers, polling vs. blocking to check the completion of communication operations, multiple VIs, completion queues, and scatter capabilities of VIA.

From the above discussion, it is clear that the papers included in this special issue cover a broad range of topics in cluster computing. We hope that the whole cluster computing community including researchers, developers, practitioners, and users find this special issue of use and interest.

We would like to take this opportunity to thank editors of the Future Generation Computer Systems, Peter Sloot and Doutzen Abma for inviting us to edit this special issue. In fact, we are honored by their invitation. We congratulate all the authors whose papers have been included in this special issue and we thank them for agreeing to extend their papers to reflect the latest state of the art in their topical area. In the references section below, we have included pointers to further information [7], particularly a list of books [1–6] that have been published recently. We hope that you will find this special issue interesting and useful. Happy reading!

References

[1] R. Buyya (Ed.), High Performance Cluster Computing: Systems and Architectures, Vol. 1, 1st Edition, Prentice-Hall, Englewood Cliffs, NJ, 1999.

[2] R. Buyya (Ed.), High Performance Cluster Computing: Programming and Applications, Vol. 2, 1st Edition, Prentice-Hall, Englewood Cliffs, NJ, 1999.

[3] G. Pfister, In Search of Clusters, 2nd Edition, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[4] K. Hwang, Z. Xu, Scalable Parallel Computing — Technology, Architecture, Programming, McGraw-Hill, New York, 1998.

[5] T. Sterling, J. Salmon, D. Becker, D. Savarrese, How to Build a Beowulf, MIT Press, Cambridge, MA, 1999.

[6] B. Wilkinson, M. Allen, Parallel Programming Techniques and Applications Using Networked Workstations and Parallel Computers, Prentice-Hall, Englewood Cliffs, NJ, 1999.

[7] Cluster Computing Info Centre. http://www.buyya.com/cluster.

[8] IEEE Task Force on Cluster Computing. http://www.ieeetfcc.org.

Rajkumar Buyya is an Australian Govt. Research Scholar at the School of Computer Science and Software Engineering, Monash University, Melbourne, Australia. He was awarded the Dharma Ratnakara Memorial Trust Gold Medal for his academic excellence during 1992 by Mysore/Kuvempu University. He has authored three books, *Microprocessor x86 Programming, Mastering C++,* and *Design of PARAS Microkernel*. He has edited a popular two-volume book on *High Performance Cluster Computing*: Architectures and Systems (Vol. 1); Programming and Application (Vol.2) published by the Prentice Hall, USA. He has edited proceedings of six international conferences and served as guest editor for many research journals in the area of Parallel and Distributed Computing, Cluster Computing, and Grid Computing. He has contributed to the development of system software for PARAM supercomputers produced by the Centre for Development of Advanced Computing, India. At Monash University, he is engaged in R&D on next generation Internet/Grid computing technologies and its applications.

Rajkumar is a speaker in the IEEE Computer Society Chapter Tutorials Program and Foundation Chair of the IEEE Computer Society Task Force on Cluster Computing (TFCC). He has co-organised and chaired three major IEEE/ACM international conferences: IEEE TFCC International Cluster Computing Conference (IEEE ClusterComp'99, Melbourne, Australia), ACM/IEEE International Workshop on Grid Computing (Grid 2000, Bangalore, India), and IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2001, Brisbane, Australia). He has chaired few other international events held in USA (CC-TEA, Las Vegas, USA, 1997-2000), China, and Germany that have merged with IEEE conferences.

Rajkumar has lectured on advanced technologies such as Parallel, Distributed and Multithreaded Computing, Internet and Java, Cluster Computing, and Java and High Performance Computing in several international conferences and institutions located in Korea, Singapore, USA, Mexico, Australia, Norway, Spain, China, Canada, Germany, France, and India, of course. His research papers have appeared in international conferences and journals. For further information, please browse: http://www.buyya.com.

Hai Jin is a Professor of Computer Science and Engineering at Huazhong University of Science and Technology (HUST) in China. He received his Ph.D. in computer engineering from HUST in 1994. In 1996, he was awarded a scholarship by German Academic Exchange Service (DAAD) at Technical University of Chemnitz in Germany. He worked at the University of Hong Kong between 1998 and 2000, where he participated in the HKU Cluster project. He worked as a visiting scholar at the Internet and Cluster Computing Laboratory at the University of Southern California between 1999 and 2000.

Dr. Jin is a member of IEEE and ACM. He chaired the *2000 Asia-Pacific International Symposium on Cluster Computing (APSCC'01)*and *First International Workshop on Internet Computing and E-Commerce (ICEC'01)*. He served as program vice-chair of *2001 International Symposium on Cluster Computing and Grid (CCGrid'01)*. He has co-authored four books and published more than 50 papers. His research interests cover parallel I/O, RAID architecture, fault tolerance, and cluster computing. Contact him at hjin@hust.edu.cn.

Toni Cortes is an associate professor at Universitat Politècnica de Catalunya, Barcelona, Spain. He obtained his Ph.D. degree in computer science in 1997 from Universitat Politècnica de Catalunya. He is currently the coordinator of the single-system image technical area in the IEEE CS Task Force on Cluster Computing. Besides working in many academic projects, he has been cooperating in European industrial projects such as Paros, Nanos, and Dimemas. He organized (along with Rajkumar Buyya) the *Cluster Computing Technologies, Environments, and Applications* session organized as part of the PDPTA'99 and PDPTA'00 Conferences. He has also been program vice chair for the IEE Cluster 2000 conferences and has been a reviewer for important international conferences such as Cluster, CC-Grid, ISCA, ICS, SCCC, and international journals such as *IEEE Transactions on Computer*s, *Parallel and Distributed Computing Practice*s, ISCA International Journal of Computers, and *Software-Practice* & *Experienc*e. He has also published more than 20 papers in international journals and conferences (most of them about parallel I/O). He has also published a chapter on parallel I/O in *High Performance Cluster Computing* (1999). His research interests cover computer architecture, parallel I/O, RAID architecture design, high performance storage system, cluster computing, and operating systems.