Lecture Notes in Computer Science 5951

Mumtaz Siddiqui Thomas Fahringer

Grid Resource Management

On-demand Provisioning, Advance Reservation, and Capacity Planning of Grid Resources

Springer

Authors

Mumtaz Siddiqui
Institute for Computer Science, University of Innsbruck
Technikerstraße 21a, 6020 Innsbruck, Austria
E-mail: Mumtaz.Siddiqui@uibk.ac.at

Thomas Fahringer
Distributed and Parallel Systems Group
Institute of Computer Science, University of Innsbruck
Technikerstraße 21a, 6020 Innsbruck, Austria
E-mail: tf@dps.uibk.ac.at

Library of Congress Control Number: 2009942714

CR Subject Classification (1998): C.2, F.2, D.2, H.3, I.2.11, C.4, I.6, K.6.5

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN	0302-9743
ISBN-10	3-642-11578-0 Springer Berlin Heidelberg New York
ISBN-13	978-3-642-11578-3 Springer Berlin Heidelberg New York

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12833689 06/3180 5 4 3 2 1 0

dedicated to our families
for their love, encouragement, and steadfastness

Preface

The Grid computing paradigm has evolved tremendously since the beginning of this millennium. It enables a flexible, secure and coordinated resource sharing among dynamic collections of individuals and institutions. The evolution has occured in various sectors – different Grid infrastructures such as computational Grid, data Grid and knowledge Grid are available to fulfill the diverse requirements of scientific communities. Today various application development and execution environments are present to create new Grid-enabled applications and to execute them onto the Grid. During this timeframe, the major focus of the research community remained on the development of an effective Grid middleware that could lead to a robust operating and management system. The middleware provides services for security, scheduling, information dissemination and resource provisioning.

Despite all these efforts, resource allocation remained a challenge for building an effective Grid resource management, as Grid applications are continuously evolving over time. In the Grid, resource management intends to make high-performance computational resources available on-demand to anyone from anywhere at anytime without undermining the resource autonomy. This is still an art due to non-dedicated heterogeneous resources distributed under multiple trust domains spanning across the Internet under the dynamic Grid environments.

In this monograph, we address the challenges of providing an effective Grid resource management by applying various techniques such as automatic brokerage, dynamic allocation, on-demand resource synthesis, advance reservation and capacity planning. In contrast to the conventional computing environments, applying these techniques in the Grid is all but straightforward. Nevertheless, in the Grid, application of such techniques is of paramount importance for resource management with effective provisioning, better utilization and an optimal allocation.

First of all, we identify the requirement of automatic resource brokerage in the Grid. The development of the Grid with diversified applications competing for scarce resources accentuates the need for an automatic resource brokerage – this is required to enable the Grid to shield its middleware complexities and to lead toward an invisible, intuitive, and robust runtime environment. Conventional research in the area of Grid resource management mainly focuses on job scheduling with manual or semi-manual resource allocation where resources must be prepared in advance with required applications and system software. This results in non-portable applications hard-coded to a specific Grid environment. Furthermore, it undermines the possible availability of a wide range of logical resources such as software services and tools, thus restricting Grid users to a limited set of physical resources. However, the importance of logical resources is accentuating with the evolution of Grid applications. Some advanced Grid programming environments allow application developers to specify Grid application components at a higher level of abstraction, which then require a dynamic mapping between high-level resource descriptions and actual deployments.

Advance reservation has been largely ignored in the Grid. A Grid application is mostly distributed in nature and runs in a workflow paradigm where dynamic mappings of application components may require advance allocation of resources so that resources could be available as the workflow execution progresses. Advance reservation improves behavioral predictability and on-demand provisioning QoS. A service-level agreement is required to ensure terms and conditions to be agreed upon during negotiation. In the Grid, advance reservation of resources is a challenging task due to the dynamic behavior of the Grid, multi-constrained contending applications, under utilization concerns, and lack of support for agreement enforcement. The under-utilization of resources is a major issue with advance reservation. For a workflow application, resources must be available even under the worst possible conditions for a successful execution. As a result, resource allocation time has to be much longer than average execution time, resulting in computing power wastage.

Nowadays a huge collections of logical resources such as software services and tools are independently and freely available in the Grid, distributed among different physical resources. In order to utilize these logical resources a manual composition is required to build a Grid-enabled application. The manual composition is not only a time-consuming process but it also requires a domain-specific knowledge. Thus, a large collection of logical resources are left unused.

This monograph renders boundaries of Grid resource management, identifies research challenges, proposes new solutions and introduces new techniques with implementations in the form of a Grid resource management system called *GridARM*. GridARM is part of Askalon—a Grid application development and execution environment. The system is designed and developed as a scalable distributed resource manager that delivers resources on-demand, works for resource providers, and optimizes resource utilization with better load distribution and capacity-planning strategies. Optimal load distribution among resources is done according to their proportional share in the Grid.

Novel techniques are introduced for on-demand provisioning, advance reservation, and capacity planning. On-demand provisioning becomes possible with automatic deployment, resource synthesis, and advance reservation. In contrast to existing resource managers, *GridARM* covers logical resources as well by enabling advanced Grid programming environments to specify application components at higher level of abstractions and mapping them to actual deployments dynamically. Henceforth, it simplifies abstract descriptions and separates them from concrete deployments. We further exploit Semantic Web technologies for the Grid to specify explicit definitions and unambiguous machine-interpretable resource descriptions for intelligent resource matching and automatic resource synthesis capabilities.

We introduce a smart negotiation protocol to make optimal resource allocation for a better resource utilization. Advance reservation is supported with a practical solution for agreement enforcement. In order to address underutilization concerns, we introduce capacity planning with multi-constrained optimized resource allocations. We model resource allocation as an *on-line strip packing problem* and introduce a new mechanism that optimizes resource utilization and other QoS parameters while generating contention-free solutions.

Furthermore, this monograph introduces a new mechanism for automatic synthesis of logical resources by applying ontology rules. The synthesis process generates new compound resources that can be provisioned as new or alternative options for negotiation and advance reservation. This is a major advantage compared to other approaches that only focus on resource matching. The newly generated compound activities provide aggregated capabilities that otherwise may not be possible; this leads toward an automatic generation of complex workflow applications. In addition, we introduce semantics in capacity planning for improving optimization in resource allocation.

The newly introduced techniques and proposed solutions are already integrated in Askalon Grid runtime environment and deployed in the Austrian Grid. The book also demonstrates the effectiveness of the system through well-performed experiments.

June 2009

Mumtaz Siddiqui
Thomas Fahringer

Contents

Part I Overview