Journal of Network and Computer Applications 65 (2016) 103–119

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Review

Survey of load balancing techniques for Grid

Deepak Kumar Patel [a,*], Devashree Tripathy [b], C.R. Tripathy [a]

[a] Department of Computer Science & Engineering, Veer Surendra Sai University of Technology, Burla, Sambalpur 768018, Odisha, India
[b] CSIR-Central Electronics Engineering Research Institute, Pilani 333031, Rajasthan, India

ARTICLE INFO

Article history:
Received 28 November 2014
Received in revised form
8 February 2016
Accepted 17 February 2016
Available online 24 February 2016

Keywords:
Grid computing
Distributed systems
Load balancing

ABSTRACT

In recent days, due to the rapid technological advancements, the Grid computing has become an important area of research. Grid computing has emerged a new field, distinguished from conventional distributed computing. It focuses on large-scale resource sharing, innovative applications and in some cases, high-performance orientation. A Grid is a network of computational resources that may potentially span many continents. The Grid serves as a comprehensive and complete system for organizations by which the maximum utilization of resources is achieved. The load balancing is a process which involves the resource management and an effective load distribution among the resources. Therefore, it is considered to be very important in Grid systems. The proposed work presents an extensive survey of the existing load balancing techniques proposed so far. These techniques are applicable for various systems depending upon the needs of the computational Grid, the type of environment, resources, virtual organizations and job profile it is supposed to work with. Each of these models has its own merits and demerits which forms the subject matter of this survey. A detailed classification of various load balancing techniques based on different parameters has also been included in the survey.

Contents

* Correspondence to: Veer Surendra Sai University of Technology, Burla, Sambalpur 768018, Odisha, India.
E-mail addresses: patel.deepak42@gmail.com (D.K. Patel), devashree.tripathy@gmail.com (D. Tripathy), crt.vssut@yahoo.com (C.R. Tripathy).

http://dx.doi.org/10.1016/j.jnca.2016.02.012
1084-8045/

1. Introduction

A *Grid* is a computing and data management infrastructure that provides the electronic underpinning for a global society in business, government, research, science and entertainment (Berman et al., 2003). A computational Grid constitutes the software and hardware infrastructure that provides dependable, consistent, pervasive and inexpensive access to high end computational capabilities (Foster and Kesselman, 1999; Foster, 2002). The Grid integrates networking, communication, computation and information to provide a virtual platform for computation and data management in the same way that the Internet integrates resources to form a virtual platform for information (Berman et al., 2003). The Grid can also be considered as a collection of distributed computing resources over a local or wide area network that appear to an end user as one large virtual computing system (Myer, 2003). The speedy development in computing resources has enhanced the performance of computing systems with reduction in cost. The availability of low cost, high speed networks, powerful computers coupled with the advances and the popularity of the Internet has led the computing environment to be mapped from the traditional distributed systems to the Grid environments (Rathore and Channa, 2014).

A *computational Grid* enables the effective access to high performance computing resources. It supports the sharing and coordinated use of resources, independently from their physical type and location, in dynamic virtual organizations that share the same goal (Rathore and Channa, 2011). Grid infrastructure provides us with the ability to dynamically link together resources as an ensemble to support the execution of large-scale, resource-intensive, and distributed applications (Berman et al., 2003). With its multitude of heterogeneous resources, a proper scheduling and efficient load balancing across the Grid is required for improving the performance of the system (Shah et al., 2007).

Load balancing has been discussed in traditional distributed systems literature for more than three decades. Various strategies and algorithms have been proposed, implemented, and classified in a number of studies. In those studies, the load balancing algorithms attempt to improve the response time of the user's submitted applications by ensuring maximal utilization of available resources. The main goal of this type of algorithm is to prevent, if possible, the condition in which some processors are overloaded with a set of tasks while others are lightly loaded or even idle (Hao et al., 2012). The process of load balancing algorithms in Grids can be generalized into the following four basic steps as shown in Fig. 1 (Yagoubi et al., 2006; Rathore and Channa, 2014).

Fig. 1. Basic load balancing steps.

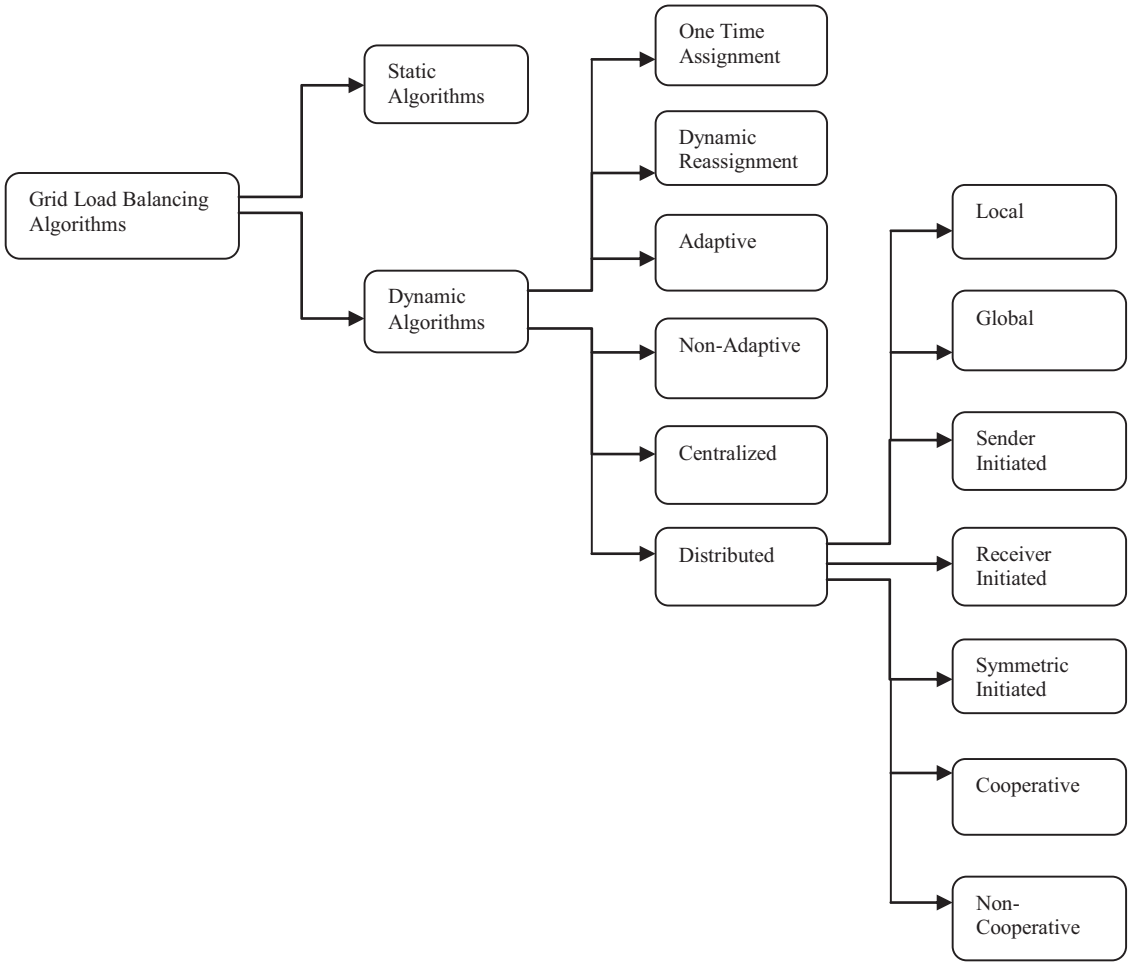

Fig. 2. Grid load balancing tree.