CHAPTER 1

Introduction

This chapter presents an overview of grid computing. It includes the rationale, emergence, scope and primary issues of grid computing. Scope of the present research work, research objectives, methodology and organization of the thesis is also included at the end of this chapter.

1.1 Grid Computing

Grid computing (Foster and Kesselman, 2003, Foster, 2002, Leinberger and Kumar, 1999) is a source of seamless and high performance computing power. It is the technology to apply the massive resources belonging to different organizations across the globe together in a network to solve the complex, resource or data intensive scientific, engineering or commercial problems. The primary goals of the grid are to make the effective use of the distributed computing resources, make available a high capacity collaborative computing infrastructure to approach the complex problems in different domains and provide a low cost access to high end resources. The rationale behind grid computing is the

- Emergence of complex resource intensive computing applications which are difficult to configure and maintained by the individual organizations.

- Economic constraints of individual, business and scientific organizations while investing in the resources (hardware/software). Especially the resources which are seldom required.
Emergence of reliable and high speed wide area communication and information sharing infrastructure to setup the distributed computing infrastructure.

Emergence of multi-institutional collaborative problem solving environment

Word wide participation in the development of grid infrastructure, protocols and standards etc.

The provision to fulfill the small scale (few processing cycles only) as well as large scale (supercomputer level or more) computing requirement at one place by creating a resource pool in computing grid.

Commitment to transform the computing resources and processes into household utility just like electric power or telephone, where users pay on usage basis without knowing the computing sources.

Unprecedented resource access in minimum cost

Recognizing grid as a promising infrastructure of next generation, many governments, research communities and educational institutions around the world has put in great efforts to initiate a number of research projects for the development of grid infrastructure and grid middleware. With this, grid computing has now developed to a stage that its initial scope of solving the high end scientific, engineering and commercial problems has been extended to fulfill also the day-to-day resource needs in offices and homes. The resources (hardware/software) and services are being presented as abstract services to be accessed and used over network through service oriented grid computing. Grid users are provided an unprecedented resource access in low cost (Iamnitchi and Foster, 2001). While exploring, the researchers have defined the computing grid in many ways, such as
• Grid is a coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations (Das et al., 2002, Krauter et al., 2002, Foster et al., 2001)

• A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high end computational capabilities (Foster, 2002).

• Grid is an infrastructure that enables the integrated and collaborative use of high end computers, networks, databases, and scientific instruments owned and managed by multiple organizations (Abramson et al., 2002).

• Grid enable the sharing, selection, and aggregation of a wide variety of resources including supercomputers, storage systems, data sources, and specialized devices that are geographically distributed and owned by different organizations for solving large-scale computational and data intensive problems in science, engineering, and commerce (Baker et al., 2002).

1.2 Emergence of Grid Computing

Although, the research communities became attentive towards the exciting technology of grid computing in mid 90s (Gentzsch, 2005, Chetty and Buyya, 2002, Foster et al., 2001), but the grid like scenario was envisioned long back in 1968 by Licklider et al. In 1969, Kleinrock forecasted the use of computing resources as utilities just like the present electric and telephone utilities at homes and offices. These early predictions could not be realized for long due to the lack of required networking infrastructure. Emergence of Internet and Web as the high performance Wide Area Network (WAN)
and information sharing infrastructure paved the way to realize the computing grid. Internet and Web has been the basis on which grid infrastructure is realized and discovered. Internet, originally called ARPANET was launched only in 1969 by the US Department of Defense’s Advanced Research Project Agency (DARPA) and was restricted to a few nodes until 1989 (Chetty and Buyya, 2002). Ethernet idea was flown in 1973 and in 1976, it came into practice. ARPANET in the form of Internet could be expanded exponentially only in 1989 with the academic oriented research on the infrastructure. After the invention of TCP/IP models and protocols in 1974, many regional, heterogeneous local area networks were combined with the ARPANET and internetwork communication became feasible. Another breakthrough in computing was the invention of Web in 1989. Web has provided an unprecedented means of information sharing across wide areas through Internet.

In early 1970s, with the advent of networking the entire computing scenario got changed. Distributed computing was experimented to harness the computing cycle on the idle machines through the networks. Evolution of high speed, high bandwidth networks, powerful computers and workstations led to the emergence of clusters to fulfill the needs of high performance computing users in low costs (Chetty and Buyya, 2002). Ubiquity of Internet and Web, high power computing resources and sophisticated software gave new dimensions to parallel distributed computing where instruction level parallelism was drawn to minimize the application run time and maximize the throughput.

Researchers in mid 1990s explored the ways to aggregate distributed resources across multiple domains, which has led to the emergence of computing grids. Instead of
assembling all the computational resources needed to solve resource intensive
application at one place, communication intensive solutions are tried where the
resources from multiple domains are integrated and used collaboratively using high-
speed networks (Buyya and Venugopal, 2005).

1.3 Grid Architecture and Research

Figure 1.1 presents the layered grid architecture and its components (Buyya, 2002). Lowest layer in the architecture is ‘Grid Fabric’ which corresponds to the grid infrastructure. It includes network infrastructure, communication protocols, hardware and software resources, storage repositories, local resource managers etc. The next two layers (Core Grid middleware and User Level middleware) correspond to the ‘grid middleware’. Grid middleware layers rests on top of the grid infrastructure. Grid infrastructure is made accessible and operative through grid middleware. It includes the routines and service of resource management, scheduling, execution management, security, authentication, authorization etc. The top layer is the ‘Application Layer’. This layer includes the functionality of application submission and configuration in grid through grid middleware. Grid applications use grid middleware to access the low level hardware and software resources.

In 2011, Huang and Yang presented a broader view of grid architecture using three layers as shown in Figure 1.2. ‘Core Grid middleware’ layer and ‘User Level Middleware’ layer is replaced with a common grid middleware layer.

Research in grid computing may be broadly classified as the grid middleware research and development of grid infrastructure (Buyya and Venugopal, 2005). Grid
middleware research aims to enable the best possible use of grid for high end computing applications as well as routine resource access. Grid infrastructure development involves the setting up of hardware, software and administrative procedures to enable the exploitation of grid computing for application configuration and execution. The widely acclaimed middleware initiatives are Globus toolkit and Gridbus project, SRB, Condor Project, AppLeS and Ninf. Major grid infrastructure development initiatives around the world are TeraGrid, GriPhyN, PPDG, EGEE, LCG, NAREGI, APAC Grid, K* Grid and GARUDA. To explore an extended list of worldwide grid initiatives refer (http://www.gridcomputing.com/).

![Layered grid architecture and its components](http://www.gridcomputing.com/)

Figure 1.1: Layered grid architecture and its components (Buyya, 2003)
Figure 1.2: A general layered grid architecture (Huang and Yang, 2011)

Figure 1.3: View of grid computing in different perspectives (http://adarshpatil.com)
1.4 Scope and Need of Grid Computing

At the core, grid computing provides the means to share and integrate the existing resources between organisations to create a powerful virtual computing infrastructure. It has a wide scope and need in different perspectives. A broader view on its scope and need is presented in Figure 1.3. Scope of grid computing scales from a low level access to a remote resource to setup a high level collaborative and high performance computing infrastructure.

a) Source of High Performance Computing

Complexity of applications in science, engineering and commerce has been increased many times over the period. Computing infrastructure (desktop systems, clusters, supercomputers, network and scientific instruments etc.) within an organization is unable to handle the growing computing needs of the complex applications in different fields. Therefore, the technologists are forced to device the mechanisms of deploying the computing infrastructure to solve the complex applications beyond the distributed parallel computing on local area networks of PCs/workstations and clusters to distributed computing on high-end computers across multiple organizations. Grid computing which is analogous to electrical power grid (Chetty and Buyya, 2002) fulfills the large scale resource requirement of the complex applications by ensuring access to large scale resources irrespective of their sources. Therefore, the applications which were otherwise difficult to configure with the limited resources in an organization are greatly benefitted from grid. A recent example is Worldwide Large Hadron Collider Computing Grid (WLCG) (www.wlcg.web.cern.ch) established to
create high capacity computing infrastructure for the storage and analysis of the large data emanating from Large Hadron Collider (LHC) experiment. LHC experiment was expected to produce more than 15 million Gigabytes of data each year. The storage and analysis of the massive data was carried out through WLCG, where more than 140 computer centers are connected in 33 countries. Similarly in India, GARUDA is first national grid initiative to bringing together academic, scientific and research communities for developing their data and compute intensive applications (http://www.garudaindia.in/html/about_garuda.aspx). It aggregates the computational nodes, storage and scientific instruments distributed across the country.

b) Collaborative Computing

Computing resources own by different organizations or users are hardly utilized round the clock. For example in an educational institution, the desktop systems in software labs are kept in on position for 6-7 hours per day on an average. Even when a desktop system is on, its CPU utilization seldom reaches to 100%. Similar the case is with other computing resource such as network and storage. Grid computing provides a collaborative computing environment where the free resource cycles are used to serve the large scale computing applications. SETI@home project is the most popular example where free CPU cycles are utilized on the idle resource to analyze radio signals, searching for signs of extra terrestrial intelligence (www.setiathome.berkeley.edu). With over 2 million independent volunteer participants, SETI@home is the world’s largest collaboration of distributed resources
spawned across 234 countries. It has the ability to compute over 769 teraFLOPS by utilizing free CPU cycles.

c) **Service Oriented Computing**

Grid computing provides an environment where resources (hardware and software) may be offered as services to the external users. Web services based Service Oriented Architecture (SOA) has transformed the computing grid into service oriented grid (Srinivasan and Treadwell, 2005). Both hardware and software resources are virtualized in the grid environment as shown in Figure 1.4. The virtualized grid resources are accessed through standardized web services

![Figure 1.4: Grid computing environment (Srinivasan and Treadwell, 2005)](image.png)
d) **On Demand Computing**

Grid computing is an alternative low cost solution to the demanding applications. A large number of distributed resources scattered among different organizations are connected together to create a central pool of available resources. Individuals or organizations can use the resources (hardware or software) from the central pool to meet their demands. Thus the required resources are not required to be owned for the application configurations. Particularly the resources which are needed only once a while can be used on demand basis from the pool.

1.5 **Primary Issues of Grid Computing**

Characteristics, design principles and features of grid computing (Baker et al., in 2002) makes it difficult to exploit the full potential of grid computing. For example resource dynamism, heterogeneity, timing constraints, quality of service, application performance, service priority, site autonomy, fault tolerance, service reliability etc. are the issues needing more explorations to meet the essence of grid. Some of the primary issues of grid computing are outlined in this section.

a) **Resource Management and Scheduling**

Grid is a de-centrally controlled environment where resource owners enjoy full autonomy having their own access policies, cost models and priorities. Widely distributed grid environment is prone to failures. De-centrally controlled, heterogeneous and dynamic resources possess intermittent availability in grid. These characteristics make the process of resource management and scheduling a complex task. The resource
management and scheduling strategies are needed to be adaptive to the changing resource availability patterns as well as the changing resource demands of the applications in grid. Grid characteristics must also be taken into account while devising the resource management and scheduling strategies.

b) Security and Privacy

Computing grid is a worldwide distributed infrastructure where resource users can log into remote resources through open network. Resource user authentication mechanisms are desirable to establish the service relationship among resource providers and resources users. It has to be ensured that who is allowed to use and who is using the resources in grid environment. Further the privacy of grid clients is also needed to be ensured.

c) Trust in Grid

Trust is the basic component on which a relationship relies. It is a belief that the concerned entity will perform as expected. In computing grid dynamic service relationships are established among grid entities (resource providers and resource users). With the notion of trust among grid entities, service reliability can be maintained in computing grid.

d) Standardization

In spite of enormous research and development around the world in the field of grid computing, a single worldwide interconnected grid is elusive. Instead, many smaller
customized grids exist to cater the needs of a specific user group. To realize a worldwide grid, the existing grid technologies are to be made interoperable through conforming to the common standards. Standardizations of grid products have been already initiated by Global Grid Forum (GGF) (www.ggf.org). GGF has produced Open Grid Service Infrastructure (OGSI) specifications, which specify the standards for grid technologies. GGF has also produced Web Service Resource Framework (WSRF) to integrate web services within grid architecture. Although the major grid projects around the world are conforming to the specified standards but still the more is to follow to make the world wide acceptability of grid computing.

1.6 Scope and objectives of the Research work

Scope of the thesis is to devise the resource scheduling strategies in grid computing which takes into consideration the probable resource failure and variance (performance and availability) during service execution in order to improve the objective parameters like reliability of service, system throughput and execution performance (by reducing the execution delay). Therefore, the objectives of the research work are defined as

- Improve the reliability of computing service in grid by providing the stable availability of computing system in grid.
- Improve the system throughput by minimizing the failures of computing service in grid.
- Reduce the task execution makespan by minimizing the delay in its execution.
- Reduce the task execution overhead through passive adaptive replication and optimum resource utilization.
1.7 Research Methodology

- First of all, a review of the existing resource scheduling strategies in the context of thesis scope is conducted.
- Research gaps in the existing resource scheduling research are noticed.
- To fill the noticed research gaps in the existing resource scheduling strategies, resource scheduling algorithms are formulated.
- A Cooperative Computing System (CCS) model is proposed.
- Mathematical modeling of the proposed CCS model is specified. CCS state is modeled using continuous time Markov process. Reliability and availability analysis of the proposed system model is given using Markov modeling of a repairable component.
- Computing grid is simulated through a discrete event based simulator developed in ‘c’ language. The formulated algorithms are used to schedule the grid resources through CCS in the simulator.
- A set of 500 computing tasks is simulated and scheduled using the formulated scheduling algorithms.
- For comparison purpose, the existing resource scheduling strategies such as checkpointing and task migration, task replication and primary-backup task execution are also simulated.
- The objective parameters are recorded through the proposed and existing scheduling strategies by varying the system parameters of such as task duration, resource failure rate, task arrival rate, system load etc.
• Improvement of the objective parameters using the proposed scheduling strategies in comparison to the existing scheduling strategies in grid is presented through bar charts in MS EXCEL.

1.8 Thesis Organization

Rest of the thesis is organized as follows. Chapter 2 presents a survey of the existing resource scheduling strategies in grid computing. By exploring the research gaps in the existing resource scheduling strategies, the resource scheduling problem is formulated in this chapter.

Chapter 3 presents Cooperative Resource Scheduling (CRS) Strategy in grid computing. In CRS strategy computing task is executed in the cooperative environment by specifying a Cooperative Computing System (CCS). Simulation results are presented and discussed to evaluate and compare the proposed resource scheduling strategy with the existing strategies.

Chapter 4 presents Trust Oriented Cooperative Resource Scheduling (TOCRS) Strategy in grid computing. CCS is specified by considering the resource trust in grid. Simulation results are presented and discussed to evaluate and compare the proposed resource scheduling strategy with CRS and Trust Oriented Resource Scheduling (TORS).

Chapter 5 presents the Execution Backup Customization in CCS to complete the execution of task conforming to its delay tolerance as maximum, average or minimum.

Chapter 6 presents Priority Based Task Execution through CRS strategy in grid. CCS is customized to allocate multiple tasks. The allocated tasks are executed in the order of their priority which is established on the basis of their delay tolerance.
Chapter 7 presents conclusions of the work performed and also includes future scope of the work.

1.9 Chapter Summary and Conclusions

In this chapter, an attempt is made to introduce the concept of grid computing. The rationale of grid computing, its emergence, scope and primary issues are discussed while introducing grid computing. Layered grid architectures used in literature are also discussed to clearly specify the research scope in the field.

In order to formulate the research problem, scope and objectives of the research are stated in the chapter. Research methodology adopted for achieving the objective and organization of the thesis is also described in the chapter.

In next chapter the existing resource scheduling strategies are reviewed in the context of scope of the research work. Research gaps in the existing scheduling strategies are observed.