ADAPTIVE INTELLIGENT GRID SCHEDULING SYSTEM

SIRILUCK LORPUNMANEE

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

JUNE 2010

Dedicated to my beloved mother **Peksim Sae-Khow,**
and father **Liang Kuang Sae-Lor**

ACKNOWLEDGEMENTS

This thesis could not be successfully completed without the kindness of both of my supervisors. First and foremost to my late supervisor, **Associate Professor Dr. Mohd Noor Md Sap** for introducing me to data mining research area. I want to express my deep gratitude to my supervisor for his support and guidance throughout my research. I am very grateful for the inspirational discussions that led to this thesis to success. The special thanks to my co-supervisor, **Professor Dr. Abdul Hanan Abdullah**, who has a good guidance during my experimentation period, and for his appreciated suggestions. Additionally, I thank him for helping me in correcting my thesis.

I would also like to thank all the staff in the Faculty of Computer Science and Information Systems at Universiti Teknologi Malaysia for their support specially in providing facilities and materials required for my research study.

I must express my sincerest and heartiest thanks to **Miss Chaupaga Panpakdeevong** for the best wishes, continuous love, encouraging, and for the confidence she gave me throughout my study. In addition, I would like to thank Suan Dusit Rajabhat University for the financial support during my study. Without them, this thesis could not have been written.

Finally, my graduation would not be achieved without best wishes from my mother **Peksim Sae-Khow,** and father **Liang Kuang Sae-Lor** specifically for the wonderful life they gave me, and greatest love and willpower which I will always remember until the day I meet them. Last but not least my special gratitude and respect to my relatives and friends for their help and encouragement.

ABSTRACT

Grid technologies are established to share the large-scale heterogeneous resources over multiple administrative domains for processing the application. In these technologies, the grid scheduling problem is crucial that must be solved in order to achieve multiple objectives within different stakeholders (end-users, owner resources and administrators) preferences. The aim of this research is to design and implement the Adaptive Intelligent Grid Scheduling System (AIGSS) in order to achieve multiple objectives named Makespan Time, Grid Efficiency and Total delayed jobs. The popular meta-heuristic algorithms, namely *Ant Colony Optimization* (ACO) and *Tabu Search* (TS) algorithms are proposed and developed to maintain the selecting appropriate grid resource to execute each job within the different job inter-arrival times and grid resources. Additionally, the clustering technique named *Fuzzy C-Means* (FCM) algorithm is proposed for clustering the groups of grid resources as well as jobs based on the degree of characteristic similarity. Moreover, a popular discrete event simulation tool, namely, GridSim toolkit and Alea simulation, is extended by developing the service modules on top of it. Therefore, the experiment is simulated as realistic grid environment in order to measure the proposed system. The experimental results show that the AIGSS provides reasonable multiple objectives to stakeholders within different job inter-arrival times and machines in grid system. In addition to the experimental results, the proposed system performs better than the other algorithms for different goals of each stakeholder. The performance of AIGSS is compared with the common and heuristic algorithms such as *First-Come-First-Serve* (FCFS) with Optimization, *Earliest Deadline First* (EDF), *Minimum Tardiness Earliest Due Date* (MTEDD), *Minimum Completion Time* **(**MCT), *Opportunistic Load Balancing* **(**OLB), *MIN-MIN*, *Hill Climbing*, *EASY Backfilling*, *Simulated Annealing* (SA), and *Tabu Searching* (TS).

ABSTRAK

Teknologi grid dibangunkan untuk membolehkan sumber seperti organisasi maya berkongsi sumber heterogen yang berskala besar ke atas pelbagai domain pentadbiran untuk memproses aplikasi. Melalui teknologi ini, masalah penjadualan grid adalah kritikal dan mesti diselesaikan bagi mencapai pelbagai objektif oleh pemegang amanah yang berbeza (pengguna akhir, pemilik sumber dan pentadbir). Matlamat kajian ini adalah untuk mereka bentuk dan melaksanakan Sistem Penjadualan Grid Pintar yang Adaptif (AIGSS) bagi mencapai pelbagai objektif iaitu *Makespan Time*, *Grid Efficiency* dan *Total delayed jobs*. Algoritma meta-heuristik yang terkenal seperti algoritma *Ant Colony Optimization* (ACO) dan *Tabu Search* (TS) dicadangkan dan dibangunkan untuk menyelenggara sumber grid yang dipilih untuk melaksanakan setiap kerja dalam sela masa ketibaan dan sumber grid yang berbeza. Tambahan lagi, teknik pengelompokan iaitu algoritma *Fuzzy C-Means* (FCM) dicadangkan untuk mengelompokkan sumber grid dan juga kerja berdasarkan kepada persamaan ciri-cirinya. Selain itu, penambahan ke atas peralatan simulasi peristiwa diskrit yang popular iaitu *GridSim toolkit* dan *Alea simulation* turut dilaksanakan dengan membangunkan modul perkhidmatan pada kedua-dua peralatan simulasi tersebut. Oleh itu, ujikaji yang dilaksanakan adalah simulasi sebenar persekitaran grid bagi menguji sistem yang dicadangkan. Hasil ujikaji menunjukkan bahawa Sistem Penjadualan Grid Pintar yang Adaptif (AIGSS) mampu mencapai pelbagai objektif yang ditetapkan oleh pemegang amanah dalam sela masa ketibaan dan mesin yang berbeza dalam sistem grid. Hasil ujikaji itu juga menunjukkan sistem yang dicadangkan mampu dilaksanakan dengan lebih baik berbanding algoritma lain bagi matlamat yang berbeza untuk setiap pemegang saham. Prestasi AIGSS telah dibandingkan dengan algoritma biasa dan algoritma heuristik seperti *First-Come-First-Serve* (FCFS) dengan pengoptimuman, *Earliest Deadline First* (EDF), *Minimum Tardiness Earliest Due Date* (MTEDD), *Minimum Completion Time* **(**MCT), *Opportunistic Load Balancing* **(**OLB), *MIN-MIN*, *Hill Climbing*, *EASY Backfilling*, *Simulated Annealing* (SA), dan *Tabu Searching* (TS).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVATIONS

p_{ij}^{k}	-	The probability with an ant *k* currently at city *i* chooses to go to an unvisited city *j*
$\eta_{ij} = 1/d_{ij}$	-	A heuristic information value, where d_{ij} is the distance between cities *i* and *j*
τ_{ij}	-	The pheromone trail value between cities *i* and *j*
α, β	-	Two parameters which determine the relative influence of heuristic information and pheromone trail
N_i^k	-	The feasible neighborhood connected to city *i* with respect to ant *k*
ρ	-	The pheromone evaporation rate
$\Delta\tau_{ij}^{bs}$	-	An amount of pheromone that only one ant constructed the shortest tour since the beginning of the trail
τ_0	-	An initial amount of pheromone
L^{bs}	-	The length of the globally best tour from the beginning of the trail
q	-	A random variable uniformly distributed in [0 … 1]
q_0	-	A parameter and its value is $0 \leq q_0 \leq 1$
J	-	The city $(J \in N_i^k)$
$X = \{x_1, x_2, ..., x_n\}$	-	A set of given data
x_i	-	The *i*th of d-dimensional measured data

v_j	-	The d-dimension center of the cluster
C	-	Fuzzy cluster
$V = \{v_1, v_2, ..., v_C\}$	-	The cluster centers
$U = (\mu_{ij})_{NxC}$	-	The degree of membership of vector x_i in the cluster jth
$d_{ij} = \|x_i - v_j\|$	-	The Euclidean distance (d_{ij}) between x_i and v_j
ε	-	Termination threshold
MPPs	-	Massively Parallel Processors computer
VOs	-	Virtual organizations
QoS	-	Quality of Service
SMP	-	Symmetric Multiple Process computer
COW	-	Cluster Of Workstations
GIS	-	Grid Information Service
GRIS	-	Grid Resource Information Server
GIIS	-	Grid Index Information Server
MDS	-	Globus Monitoring and Discovery System
NWS	-	Network Weather Service
MCT	-	Minimum Completion Time
MET	-	Minimum Execution Time
KPB	-	K-Percent Best
OLB	-	Opportunistic Load Balancing
BLBD	-	Based on Load Balancing and Demand
UDA	-	User Directed Assignment
MECT	-	Minimum Execution Completion Time
ACO	-	Ant Colony Optimization
AS	-	Ant System
EAS	-	Elitist strategy for Ant System
MMAS	-	Max-Min AS

ACS	-	Ant Colony System
BACO	-	Balanced Ant Colony Optimization
TS	-	Tabu Search
SA	-	Simulated Annealing
GA	-	Genetic Algorithm
FCM	-	Fuzzy C-Means Algorithm
TSP	-	Travelling Salesman Problem
ETC	-	Expected Time to Complete
FCFS	-	First Come First Served
EDD	-	Earliest Due Date
EDF	-	Earliest Deadline First
ERD	-	Earliest Release Date
MTEDD	-	Minimum Tardiness Earliest Due Date
EASY	-	Extensible Argonne Scheduling sYstem
SDSC	-	San Diego Super Computing Center
MIPS	-	Million Instructions Per Second
MI	-	Million Instructions
PEs	-	Processing Elements or CPUs
SPEC	-	Standard Performance Evaluation Corporation
OS	-	Operating System
UML	-	Unified Modeling Language
DAG	-	Directed Acyclic Graphs

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Preliminary

Generally, computing systems have developed starting from a single computing system and have eventually grown into Massively Parallel Processors (MPPs), computing network of workstations, distributed computing systems and now, grid computing systems. In fact, the development of computation systems always improves the ability of the systems to solve large complex problems and process a large amount of data. The high performance systems are able to solve large complexity problems and process a large amount of data. But they are restricted to costly systems available and privileged locations such as super computing centres. Therefore, the resources are often provided in a single organization. Recently, the explosion of the Internet and high-speed networks, mainly as a result of faster hardware and more sophisticated software, helps workstations to connect with one another. According to these technologies, the distributed computing systems are formed as a single system, unified computing resources, which lead to the popular heterogeneous distributed computing systems known as grid computing systems (Foster and Kesselman, 1999). The term grid is usually known as an electrical power grid which provides dependable, consistent, pervasive, and also transparent access to electricity but not irrespective of its resources.

Grid computing systems are the latest heterogeneous distributed computing systems concept and technology. According to Foster and Kesselman (1999); Foster *et al.* (2002); Mark *et al.* (2002), grids are recently proposed to enable the selection and sharing of large resources, and integration of various resources including supercomputers, workstations, storage systems, data sources, networks systems, software, specialized devices and even people. These technologies therefore have led to geographically distributed computing systems and multiple domain spanning resources in solving the large-scale computational and data intensive problems in science, engineering and business. Virtual organizations (VOs) are initially defined by Foster *et al.* (2001), where the researchers had tried to address and summarize the grid computing systems according to the current technologies and their framework.

1.2 Grid Computing Environments: An Overview

Most Grid computing systems, in the research and academic context, emphasize increasing the efficiency as well as leveraging high performance computing systems. Currently, the increasing performance of networking technologies and personal computers have produced powerful distributed computing systems by connecting PCs and the other devices together on the different networks. With the large gathering of connected distributed computing systems with its various resources sharing, the grid computing systems are the next evolutionary level of the heterogeneous distributed computing systems. It usually comprises of geographically heterogeneous distributed resources with each local management system such as Operating System (OS), queuing systems and resource owners (CPU, I/O, memory, etc.), policies and applications (engineering, scientific and commercial) with different requirements (CPU, memory, storage, etc.) connected with the network intensively. Consequently, it is geographically distributed with different time zones and heterogeneous in nature. The grids are made of a virtual organization (VO) (Foster *et al.*, 2001), who is the multiple stakeholder (resource owner and end users) aiming for different goals, objectives, strategies and demand patterns (Buyya *et al.*, 2000).

The project to link geographically dispersed supercomputers, distributed computing systems and to connect other devices, leads to the start of conceptual grid computing environments. Currently, the fast growing of the Internet has taken us far beyond its original intention. Wide-area distributed supercomputers and workstation computers linked with other devices have been a popular infrastructure of grid computing environments, which a number of applications can take advantage of it. Nevertheless, the advantages of the grid infrastructure are not only in its application but also the collaborative engineering, huge data exploration, high throughput computing and high performance computing systems.

The grid can be viewed as an open system that integrates computational and collaborative environment. The users interacting with grid computing systems can potentially use the enormous number of services that are available in the grid. It is similar to the way that electrical devices can draw power from electrical power grid. The different users across a virtual organization (VO) can get the opportunity to use grid computing system for solving a large application problem.

1.3 Problem Background

The grid computing systems are one of the heterogeneous distributed computing systems that have been mentioned before. A common requirement in these systems is to utilize resources located within different administrative domains and provide services for the requesting application. The utilization of grid resources also provides an opportunity for collaborative computing to solve large complex problems and process a large amount of data. The complicated computing needed by the grid resource consumer (or grid application) and the resource owner lead to complicated resources sharing. Therefore, the application needs to take into account of the resources behaviour that often demands assurances of the level and type of service provided by the resources.

According to the dynamic nature of grid, the system workload of each grid resources cannot be determined in advance and also it is impossible for grid users and grid system administrators to manually find and specify all the needed grid resources during the arrival of applications. To achieve the goals and potentials of computational grids, robust mechanisms for allocating resources to applications, an efficient and effective scheduling system is fundamentally important.

The problem of scheduling jobs assigned to resources has been extensively studied as a basic problem in traditional parallel and distributed systems. For example, symmetric multiple process machine (SMP), massively parallel processors computers (MPP), cluster of workstations (COW) and even the new conceptual heterogeneous distributed computing systems such as grid computing system have been studied. The different objectives in scheduling algorithm are designed to meet these aims. Unfortunately, all of these scheduling algorithms such as FCFS (First Come First Served) and STF (Shortest Task First) (Karatza, 2000) in traditional parallel and distributed systems do not usually perform well in the new circumstances such as grid system environment in particular (Berman, 1998).

1.3.1 The Grid Scheduling Problem

According to the characteristics of grid environment which are currently presented, a number of applications are turning to grid environment to achieve their computational resources and share large data storage. Existing scheduling systems are now based on distributed computing systems. For example in Casavant and Kuhl (1988), the detailed description of different categories of scheduling classified a large number of distributed scheduling approaches according to the taxonomy. Additionally, the researches of Gehring and Reinefeld (1996); Weissman (1996); Weissman (1999); Xing and Xiaodong (1997) of scheduling systems for distributed were focused on maintaining load balancing, throughput of the systems, and turnaround times.

From the point of view of these scheduling, it is well known that the complexity of a general scheduling problem is NP-Complete problem (El-Rewini *et al.*, 1994; Fernandez-Baca, 1989; Tracy D. Braun *et al.,* 2001) which has been presented by several researches. Besides the grid scheduling problem issues have been studied by many researchers as referred in Nabrzyski *et al.* (2003). In Fibich *et al.* (2005), these researchers have studied and proposed the model of grid scheduling by extending from Graham's classification of scheduling problems (R.L. Graham, 1979). In brief, they made an effort to propose the model of grid scheduling problem, its model may consist of environment, characteristics, constraints and the objective. Therefore, the different job and machine characteristics in the natural grid are the issue in order to achieve the performance grid.

Grid scheduling system is usually defined as the process of making scheduling decisions that involves resources across multiple administrative domains. This process can also include searching module that gathers the different resources in multiple administrative domains. Its performance assigns or schedules an application or a single job to use multiple resources at a single site or multiple sites. For an application, to a set of application such as a parameter sweep, a single job can be anything that needs resources to execute, for example CPU request, network bandwidth as well as data storage request. For the resource, it is anything that can be scheduled, for instance, a machine, CPU speed, CPU requirement, disk space, QoS network, and so forth. The difference between grid scheduler and local resource scheduler is typically the owners control over the resources. Since the functional local resource schedule is usually responsible for scheduling and also managing resources at a single site or perhaps only for a single cluster or resource. These are, therefore, the lowest-level and lower-level scheduling as discussed within Uwe and Ramin (2004). Particularly, grid scheduler has no control over a set of resources and a set of job assigned to them, even the state of their information. In fact, the dynamic resource systems and grid users in the natural grid are willing to deal with each other in order to allocate the coordination of the resources for their application or job. As a result, this lacks of ownership, control over the resources and also to use manually find, reserve, and allocate all the resources needed to execute an application or a job. These issues lead to address the challenges of problem.

The grid account of a high diversity system in general, in which comprises of various applications, middleware component, and resources. The point of view of functional scheduling involves three main phases (Jennifer, 2004), which comprises of resources discovery, resource selecting and scheduling, and job execution. By these phases, to lack of information of the status of available resources, especially the dynamic heterogeneous distributed computing systems such as the nature of grid, can lead to inefficient and ineffective scheduling. Accordingly, grid information service (GIS) is an indispensable component in the nature grid. The main functional Grid Information Service (GIS) is usually discovering, gathering and providing information to grid scheduler system. The information within GIS may be comprised of CPU capacities, memory size, storage size, network bandwidth, software availabilities, work load of each resource in particular period and so forth. Currently, Globus Toolkit (http://www.globus.org/toolkit/) is open source for building Grid systems and applications. It also provides Globus Monitoring and Discovery System (MDS) (Czajkowski *et al.*, 2001) for providing information service; likewise, Network Weather Service (NWS) (Wolski *et al.*, 1999) also provides the information service. These are two mature examples of GIS.

Therefore, the information of grid system plays highly important role in helping the grid scheduler system for selecting, making scheduling decision along with criteria estimation. The lack of detailed information causes the poor performance of the scheduler system. Because of the diversity of job and resource (machines) characteristics in the natural grid, the job consists of arrival time, a number of CPUs requirement, size of job and its due date, as well as the machine consists of CPU speed, a number of CPU and its workload, involve the making scheduling decision in order to achieve the objective. Also, a number of machine and job, the different objectives of the stakeholders (administrators, resource owners, end users) and the types of job cause the different scheduling decision making. Similarly, scheduling algorithm within grid scheduler system is also an indispensable module. The successful grid scheduler system needs intelligent algorithm and accommodates all of the information for selecting, making scheduling decision that allocates jobs or applications to proper resources belong to criteria estimation. The intelligence algorithm usually needs correct information so that it can make scheduling decision well. In order to achieve the objectives and potentials of computational grids, the requirement is not only the precise provided information but also the intelligent algorithm. The successful grid scheduler system therefore should consider and accommodate all of these issues.

1.3.2 The Grid Scheduling Algorithms

The grid scheduling system must make resources selection decisions in an environment where it does not have fully control over local resources on grid sites in nature. Additionally, it has significantly to deal with high rapidly changing, regularly unpredictable input data, causing ambiguity and incompleteness of information. Furthermore, the performance and the availability characteristics of the individual resources within the resources pool and also the status of grid applications may change from time to time.

The diverse characteristics of grid applications and grid resources have been therefore considered to significantly impact on the design of scheduling algorithm system. To achieve the computational grid of performance goals, the grid information service plays a highly important role. Moreover, the intelligent scheduling algorithm should be considered and accommodated all of these issues.

From the definition of grid system in nature (Mark *et al.,* 2002), they showed the necessary components to form grid system and the services of grid were also discussed by details. The theory, algorithm of scheduling (Pinedo, 2002) characterized the classification of scheduling problem in term of a number of job and machines, their characteristics and optimization criterion. By these characteristics, they were described by several types, for example the arrival time of job, the processing time of job, the due date of job, machines in parallel with different speeds, unrelated machines, and so on. For optimization criterion, the makespan time, average complete time and maximum lateness, etc. were shown by details. These constraints almost satisfy in traditional scheduling problems while grid scheduling is significantly a new scheduling problem which differs from the traditional scheduling in the diverse characteristics of circumstance. However, grid scheduling problem can get benefits from several traditional scheduling methodologies. These methodologies have achieved successful results in a wide variety of scheduling problems.

One of the first taxonomies of scheduling algorithm in distributed computing systems is categorized by Casavant and Kuhl (1988). They mainly categorized the problem domain into static and dynamic scheduling. They also presented the groups of algorithm which related to the kind of domain problem, optimal, sub-optimal, approximate, and heuristic for example. However, due to the grids scheduling, which is a special kind of scheduling system, scheduling algorithm in the grids therefore is subset of this taxonomy. Additionally, another research in Tracy D. Braun *et al.* (1998) made effort to present taxonomy of a matching and scheduling heuristics. The major different parts of their taxonomy for classifying scheduling heuristic have shown in three parts; (1) the models used for applications and communication request, (2) the models used for target hardware platforms and (3) the characteristics of mapping heuristics.

Dynamic heuristics for a class of independent tasks presented in reference (Muthucumaru Maheswaran *et al.*, 1999) mentioned that two types of different mapping heuristics of their research within different heuristic algorithms. And there were five different heuristic algorithms in online mode (immediate mode); Minimum Completion Time (MCT), Minimum Execution Time (MET), Switching Algorithm (SA), K-Percent Best (KPB) and Opportunistic load balancing (OLB). While the other heuristic algorithms such as Min-Min, Max-min and Sufferage were presented in batch mode. The results of their experiment were the most appropriate for a given situation. However, the degree of different machine characteristics, for example number of machine, number of CPU, load of CPU and also job characteristics such as deadline time, finish time, arrival time, requirement of CPU, etc. were not considered. Furthermore, their work made a common assumption to solve the problem. The performance evaluation of grid system such as make-span time was not adequate because grid computing systems usually play with many user parties, for example end-users, administrators.

Currently, the new approaches of the heuristic algorithms have been inspired by natural phenomena. These have been proposed to address the challenge of the grid scheduling algorithm. Several algorithms from natural phenomena have been introduced to form powerful heuristics and those have been proven to be highly successful. The meta-heuristic algorithms for grid scheduling, named Simulated Annealing (SA), Tabu Search (TS), and Ant Colony Optimization (ACO) have been applied by using each standard algorithm.

1.4 Problem Statement

In recent years, the dynamic scheduler system problems in grid environment have grown in the number of the studies. Many common goals of grid scheduler system such as to minimize the makespan time of a job set, tardiness time, as well as to maximize resource utilization, non-delayed jobs, and so on have been highly considered. Additionally, several criteria of scheduling problems, for example, minimizing the cost, communication delay, giving priority to certain users' processes, needing specialized hardware devices, the different resources and job characteristics, are included. Moreover, the soft constraints and hard constraints (see 3.4.1.3 in Chapter 3) play a highly important role. Unfortunately, scheduling algorithms in traditional parallel and distributed computing systems, which normally operate on homogeneous and dedicated resources such as computer clusters, supercomputer, cannot promise good potential for the grid environment because scheduling in grid environment is significantly complicated by the heterogeneous and dynamics nature of the grid.

In recent years also, the dynamic heterogeneous distributed computing systems such as grid environment and its service such as Grid Information Service (GIS) are proposed to address the challenges of grid scheduling system. The discussion in problem background has shown the techniques with some limitations in handling grid scheduling problems. Especially meta-heuristic algorithms as referred in Tracy D. Braun *et al.* (1999); Young *et al.* (2003); Zhihong *et al.* (2003), which have been growing number of studies. According to these studies, the researchers proposed the different meta-heuristics by several objectives named the makespan time, the response time, the sum of the price per unit time and resource average utilization. However, the several requirement goals in order to satisfy different stakeholders such as high performance, decreasing delayed jobs and increasing grid efficiency were not taken into consideration as multiple objectives in assigning the job to available machine in dynamic grid system.

After the study of the problem background, there are several issues that should be addressed. Hence, this research is carried out in order to answer the following questions:

The main question of the Grid Scheduling System is to share the large resources in grid environment within the soft and hard constraints (see 3.4.1.3 in Chapter 3) that leads to the exploration of the method in achieving the multiple objectives of the grids. This question also deals with an algorithm for providing adaptive, realistic estimation of execution cost of the jobs and current system information status. Therefore, the questions are as follow:

i) How does the Grid Scheduling System improve the performance of the grid system based on the soft and hard constraints?

ii) How to design and implement an efficient and effective Grid Scheduler System so that the grid system can achieve the multiple objectives?

iii) How to discover and allocate the optimal available grid resource for each job?

Secondly, accommodating the different job characteristics and grid resources (machines) within the rapidly changing system information are also considered in order to improve the high performance, high throughput and high system usage of the overall system. Therefore, the questions are as follow:

i) How to cluster the groups of the jobs based on the degree of similarity of job characteristics that should be assigned to the suitable grid resources?

ii) How to cluster the groups of the machines based on the degree of similarity of machine characteristics before executing the job?

iii) Do Grid Information Service (GIS) and the current system status information help to improve the performance of grid scheduling system?

Finally, exploring the algorithm and method for solving grid scheduling problems will be of further interest. Therefore, the questions are as follow:

i) Which scheduling algorithms give better results on grid environment?

ii) How to apply clustering technique and meta-heuristic algorithms to solve grid scheduling problems within the complexity of its environment?

1.5 Objectives of Research

The objectives of this research is to improve the performance, maintain system efficiency, and also accommodate several jobs into the grid computing systems with the use of the Adaptive Intelligent Grid Scheduling System (AIGSS). To deal with the multiple objectives of high performance, decreasing delayed job and increasing grid efficiency, multi-criteria plays a highly important role. The criteria can distinguish the two general groups that are related to particular stakeholders (end users and resources owners). For end users, time criteria such as makespan time, tardiness time, and due date are considered in this research, while grid efficiency and resource utilization are also taken into account. Therefore, the multiple objectives of this research are as follows:

i) To study, design and develop the Adaptive Intelligence Grid Scheduling System (AIGSS) architecture.
ii) To develop and apply the meta-heuristic algorithms in searching the appropriate grid resource to execute the jobs in the grid.
iii) To develop and apply data mining technique for clustering the jobs and grid resources based on the degree of similarity of their characteristics.
iv) To evaluate and compare the proposed solution with the other algorithms based on the performance, delayed jobs and efficiency of grid in this research so that the strength and weakness of them can be find.

1.6 Assumptions and Scope of the Research

The design, development and experiment involved in this research have a few limitations that are worth mentioning.

i) Grid Scheduler system is responsible for making scheduling decision based on the different jobs characteristics and a set of grid resources within their characteristics. In this approach, the local resource scheduling policies of the different grid resources (machines) are responsible for scheduling, managing and controlling over the resources at lower-level scheduling instance (Jennifer, 2004; Uwe and Ramin, 2004) that are only considered in space sharing fashion. Additionally, the set of grid resources across multiple administration domains is not considered because this may be a drawback when each implementation of its own local resource scheduling policies within different administrative domains. This research uses simulation based on GridSim Toolkit (Buyya and Manzur, 2002; Klus´aˇcek *et al.*, 2007), Alea simulation (Klus´aˇcek *et al.*, 2007) and also extends internal entities to more complex requirement. Moreover, the discrete-event simulation plays a highly important role in this research.

ii) In this research, the independent jobs are mainly concentrated and are also scheduled in non-pre-emptive manner.

iii) In view of the independent jobs concept, the network bandwidth of the grid system is not considered.

iv) The machine is only considered as grid resource in this research.

v) Finally, this research does not consider the grid resources failure, job loss, job migration and the precedence constraints for workflow jobs (Directed Acyclic Graph (DAG) applications (Yongcai *et al.*, 2007)).

1.7 Research Contributions

In this research, the artificial intelligence, knowledge discovery and data mining techniques in dynamic environment are mainly used to build Adaptive Intelligence Grid Scheduling System (AIGSS). This research work is different from many relevant areas of existing research which are described as follows:

i) The research builds the Adaptive Intelligent Grid Scheduling System (AIGSS) that provides a model consisting of knowledge information discovery, clustering technique and adaptive grid scheduling algorithm. Most of the existing systems cannot meet all modules.

ii) The research also designs and implements the Adaptive Intelligence Grid Scheduling Algorithm that helps make accurate decisions in assigning jobs to available grid resources within dynamic grid environment by using intelligent Ant Colony Optimization and Tabu Search algorithms in order to achieve the multiple objectives.

iii) To apply data mining technique for clustering of job and grid resource based on the degree of their characteristics similarity by using Fuzzy C-Means algorithm.

iv) Lastly, the proposed system known as the Adaptive Intelligence Grid Scheduling System (AIGSS) attempts to deal with different characteristics of the independent jobs and dynamic grid resources. Whole results are depended on the diversity of grid systems such as the different job inter-arrival times, grid resources as well as the fluctuation in CPU load.

1.8 Thesis Organization

The thesis has been organized to elaborate the major aspects of this complex multidisciplinary study. Following paragraphs give a brief overview of the thesis.

Chapter 2 presents a brief overview of taxonomies for grid scheduling schemes including scheduler organization and scheduling policies, and grid scheduling algorithms. Additionally, description of previous works that used the different techniques including the Meta-heuristic algorithms such as Ant Colony Optimization (ACO), Tabu Search (TS) and Simulated Annealing (SA) algorithms will also be mentioned. Moreover, Grid Information Service (GIS), the simulation tools for grid scheduling systems, and the modelling including architecture of GridSim are presented in this thesis.

Chapter 3 discusses the definitions of the grid scheduling problems. Also, the grid scheduling constraints, the definitions of job and grid resource characteristics, and the performance evaluation metrics are discussed and presented. Additionally, the demonstration of general description of the system design and development process in order to achieve objectives of this study is also discussed.

Chapter 4 presents the details of Adaptive Intelligent Grid Scheduling System (AIGSS) architecture. The main components consist of the Intelligent Job Characteristics Clustering Groups module, Adaptive Intelligent Grid Scheduling Algorithm module, and Resource Information Management and Knowledge Discovery module.

Chapter 5 presents the architecture model of extended simulation with all important classes based on GridSim toolkit and Alea Simulation. The UML diagram is provided to describe the details of all classes in the simulation. Additionally, the flowchart diagrams of the important modules consisting of the Intelligent Job Characteristics Clustering Groups module, the Resource Information Management and Knowledge Discovery module, and the Adaptive Intelligent Grid Scheduling Algorithm module are presented.

Chapter 6 presents the experiments and results that are shown by the different grid scenarios, for example different job inter-arrival times and a number of machines including different CPU speed and a number of its CPUs. The comparison with other algorithms will show that the proposed Adaptive Intelligent Grid Scheduling System (AIGSS) can achieve the multiple objectives of different requirements of each stakeholder.

The thesis concludes in Chapter 7 by expressing the experiments and results. It is strongly clear that the different scenarios in grid system can cause the different results within the other algorithms. However, the proposed system can produce the good results in comparison. The challenges and emerging trends identified for future research are also suggested in this chapter.