The Fair-based Dispatching Architecture in Cloud Computing

Mao-Lun Chiang, Chuan-Bi Lin*, Jian-Wei Li, Chia-Chi Chang, Tz-Han Chen
Department of Information and Communication Engineering, Chaoyang University of
Technology 168 Wufeng District, Taichung County, Taiwan 413, R.O.C

{mlchiang; cblin*; lijw; ccchang; s10030608}@cyut.edu.tw

*: Corresponding author

ABSTRACT

As the network technology continues to grow at a high rate of speed, traditional
network topology is improved by the novel distributed topology such as cloud
network. With a new concept proposed by Google at 2007, the integrated services can
satisfy the huge demand of Internet application. However, a huge amount of service
demands can make overloading and omission of tasks for service providers. Because
the tasks need to be re-scheduled frequently, the costs are increased for users and the
workload of each node be unbalance. Therefore, there is a need of fair and
non-preemptive dispatching algorithms that provide high throughput under several
distributions of tasks in a queued cloud computing infrastructure. Based on the above
reason, the Fair-based Dispatching Architecture in Cloud Computing (FDACC) is
proposed to enhance the dispatching performance and reduce the search cost of the
cloud computing environment by fair deployments.

Keyword: Cloud Computing, Dispatching Algorithm, Queues, Scheduling.

1. INTRODUCTION

In general, the Internet applications and computer hardware grow rapidly resulting in
a large overhead for Internet Service Provider (ISP). To satisfy the huge demands on
Internet applications, the computational ability of hardware and network bandwidth
need to be improved. In view of this, a basic architecture is proposed by Google based
on the concept of grid computing (Buuyal et al., 2009) in a distributed system, which
may be deemed as a variant form of grid computing, known as “cloud computing”
(Grossman, 2009; Shafer et al., 2010; Amazon WEB Service, 2009; Application
Delivery Networking, 2009; Gartner, 2009; What is Cloud Computing?, 2009; ZXTM
for Cloud Hosting Providers, 2009), as both are extensions of distributed computing.
“Grid computing” focuses on the integration of multiple heterogeneous platforms.
The major difference between cloud computing and grid computing is that each task
of a grid computing contains a large quantity of data and usually be used in academic
research. Contrarily, cloud computing environment contains a small quantity of data
and vary with services. Therefore, the cloud computing has capability of extensibility
to support the different service requirements. For example, a company can rent the
ability from the cloud when it needs short-term resources, such as Amazon EC2
Most emerging industries prefer to use cloud computing to save costs and improve service ability on demand. In general, the cloud computing environment can be divided into three major parts, Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Server (IaaS) (Buyyal et al., 2009; Zhong et al., 2010). There are some services on web 2.0 to be provided in SaaS such as Google docs (More Google Product, 2009), Youtube (Youtube, 2012), Citrix (Citrix, 2010), and so on. In PaaS, a development platform can be used to develop the program, such as the Windows Azure Platform (Windows Azure Platform, 2010). Subsequently, the resource infrastructure can be provided in IaaS, such as Amazon EC2, Zeus (Amazon Web Services, 2009; Zeus, 2009). Therefore, the construction costs of physical equipment can be reduced.

Unfortunately, as a large number of nodes are used to provide services in cloud computing environment, the dispatching costs for searching the adaptive service node are heavy. For the above reason, a FDACC is proposed in this paper, to enhance the dispatching performance and reduce the search cost of the cloud computing environment by fair deployments.

The rest of this paper is organized as follows: Section 2 illustrates the related works of this paper. The detail of protocol we proposed is shown as Section 3. Section 4 illustrates out an example in detail. Subsequently, the conclusion will be presented in Section 5.

2. RELATED WORK

Cloud computing is a new distributed system concept that has been implemented by businesses such as Google, IBM, Microsoft, Yahoo, and Amazon (Amazon Web Services, 2009). Google provides various applications on their internet platform such as Gmail and YouTube (More Google Product, 2009; Youtube, 2012). In addition, Google provides free storage capacity for each user. The big and powerful Google search engine allows users to find multiple results from different file types on the Internet. Besides, Amazon provides many applications through Amazon Web Services (AWS) as cloud computing environment (Amazon Web Services, 2009); for example, Elastic Compute Cloud (EC2), Simple Storage Service (S3), and CloudFront (beta version). With AWS, users can requisition computing power, storage, and other services and gain access to a suite of elastic IT infrastructure services on demand. The AWS allows users to rent the infrastructure or application services platform as a cloud computing environment in accordance with the amount of cost.

Therefore, the cloud computing is a style of computing where massively scalable IT-related capabilities are provided to multiple external customers “as a service” using internet technologies. The cloud providers achieve a large, general-purpose computing infrastructure, and they have to use virtualization of infrastructure for different
customers and provide the multiple application services, such as ZXTM and ZEUS Web Server (ZWS) (Application Delivery Networking, 2009; Application Traffic Management, 2009; Windows Azure Platform, 2010; Zeus, 2009; ZXTM for Cloud Hosting Providers, 2009), as shown in Fig 1.

![Figure 1](image_url)

Figure 1. The resource of cloud infrastructure with virtualization

However, the distinguishing features of a cloud computing environment are a number of applications can be provided specific due to its computational ability. The huge costs of search for adaptive service node and management are necessary. Therefore, how to build the fair resource deployments and the efficient task dispatches in a cloud computing system can be considered.

3. The framework of cloud computing and definitions

This section describes a two-layer cloud architecture, the Fair-based Dispatching Architecture in Cloud Computing (FDACC), as shown in Figure 2. The front end of this architecture is Input Queues (IQs), which manages and allocates tasks to specific queue. The back end of this cloud architecture is Server Queues (SQs), which major function is to receive the tasks allocated from the upper layer IQs into the queue and assign the task to server respective. Subsequently, the IQs can execute the proposed Fair-based scheduling algorithm (FSA) by the current queue status. The dispatching algorithm of this architecture is shown as follows.
3.1 The **Fair-based Scheduling Algorithm (FSA) in cloud computing**

The FSA follows request-accept steps and show as follows:

**Request:** Non-full occupancy SQs send a request to their destined IQs. Requests include the number of tasks in the SQs.

**Accept:** If an IQ arbiter \( a_i \) receives any requests, it accepts the one with the least occupancy from the SQ. Ties are broken arbitrarily.

After the IQ arbiter \( a_i \) accepts a request from the SQ, the IQ will send the tasks to the SQ and the value of SQ\(_j\) counter is updated as follows:

If \( \text{SQ}_j(t) \geq 0 \), \( \text{SQ}_j(t+1) = N \), where \( N \) is the number of full occupancy SQ\(_j\), and SQ\(_j\) does not send a request to IQ at time slot \( t+1 \).

**Figure 3. An example of FSA**

Figure 3 shows an example of a matching in the FSA. The SQ values are shown as server queue contents. In this example, we assume that the full sizes are the number of 8 at each SQ. In the request phase, server queues 0, 1, and 3 send a request to IQ with the number of tasks in the SQs. SQ\(_2\) does not send a request to IQ as its queue is full. The input arbiter selects a request in a fair-based fashion among all requests with the value of non-full occupancy SQs, as shown by accept phase. IQ selects the request
from SQ3 over the requests from SQ0 and SQ1. After the match is completed, IQ sends tasks to SQ3. Noted that at time slot $t+1$, the SQ3 value becomes 8, where is the number of full occupancy as shown in the figure.

4. CONCLUSIONS

Due to a huge amount of service demands, it can make overloading and omission of tasks for service providers. The costs are increased for users and workload of each node be unbalance because the tasks need to be re-scheduled frequently. Therefore, a novel architecture is proposed in this paper, Fair-based Dispatching Architecture in Cloud Computing (FDACC), to determine task eligibility in the dispatching process for the cloud computing environment. Also, we proposed the fair-based Scheduling Algorithm (FSA) to enhance the dispatching performance and reduce the search cost of the cloud computing environment by fair deployments.

ACKNOWLEDGMENT

This work was supported in part by the Taiwan National Science Council under Grant NSC99-2221-E-324-041-MY3.

REFERENCES


Application Traffic Management, Application Security, October 2009,


