Grid Computing-A Next Level Challenge with Big Data

C.Chandhini, Megana L.P

Abstract— This paper provides an extensive survey of grid computing, the need for grid computing, Distributed Caching and how distributed caching can be implemented in grid computing in order to solve the problem of big data. Each one of these topics is explained in detail. Information regarding grid security has also been mentioned about. And finally the applications which use grid computing are highlighted. e formatted further at IJSER. Define all symbols used in the abstract. Do not cite references in the abstract. Do not delete the blank line immediately above the abstract; it sets the footnote at the bottom of this column. Don't use all caps for research paper title.

Index Terms— LAN-Local Area Network, PC-Program Counters, ROI-Return on Investment, RFID-Radio Frequency ID, OLAP-Online Analytical Processimg, SSL-Secure Sockets Layer, CAD-Computer Aided Design and Drafting.

1 INTRODUCTION

IN the last few years there has been a rapid exponential increase in computer processing power, data storage and communication. But still there are many complex and computation intensive problem, which cannot be solved by supercomputers. These problems can only be met with a vast variety of heterogeneous resources. The increased use and popularity of the internet and availability of high-speed networks have gradually changed the way we do computing. Imagine several million computers from all over the world, and owned by thousand different people. Imagine they include desktops, laptops supercomputers, data vaults and instruments like mobile phones, telescopes, meteorological sensors. Now imagine that all of these computers can be connected to form a single, huge and super-powerful computer! This huge, sprawling global computer is what many people dream "The Grid" will be.

2 WHAT IS GRID COMPUTING

Grid computing is a model of distributed computing that uses geographically and administratively disparate resources. In grid computing individual users can access computers and data transparently, without having to consider location, operating system, account administration and other details. In grid computing the details are abstracted and the resources are virtualized.

In Grid Computing all the intended LAN are connected to each other through internet. If any client needs to perform some task on the grid, first it sends query to the domain in order to get information of the master server. Domain gives information of master server. Then client sends the request to the master server to perform the task. Master server divides the task into subtasks and distributes them to all nodes across the internet. Local servers at each node receive the assigned task, find the available systems on the LAN and distribute the assigned task to available systems. Local PC's perform the task and send their individual results to local servers. Local servers merge the collected result and send it back to the master servers. Master server combines the gathered results and then either send it to the client or the intended database.

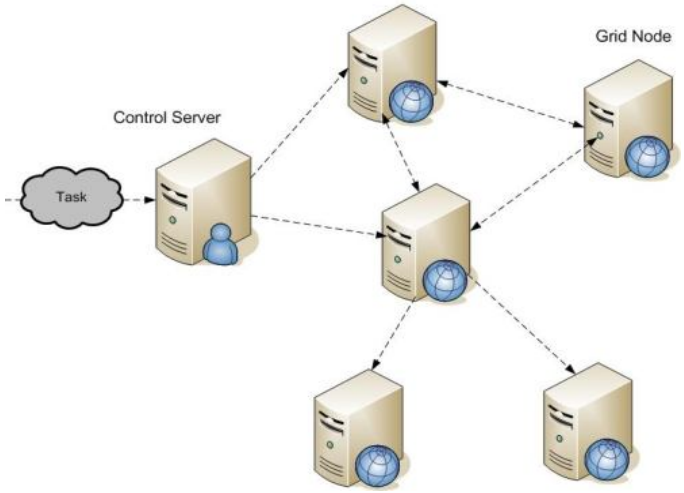


Figure 1 Grid

3 WHY GRID COMPUTING

Since the last decade there has been a substantial increase in commodity computing and network performance, mainly as a result of faster hardware and sophisticated software. These commodities technologies and fast networks have been used to develop high-performance computing systems, called clusters, to solve resource intensive problems in a number of application domains. But these systems have been found incapable of handling massive data processing and storage.

Therefore, for such challenges which revolve around data managing its access, distribution, processing and its storage, computational infrastructure, coupling wide-area distributed

- *Chandhini Chandersekar is currently pursuing Bachelor's degree program in Computer Science engineering in Velammal Institute of Technology-Affiliated to Anna University, India, PH-00919791020767. E-mail: chandhini1993@gmail.com*
- *Megana Lakshmi Padmanabhan is currently pursuingBachelor's degree program in computer science engineering inVelammal Institute of Technology-Affiliated to Anna University, India, PH-00919444750165. E-mail: megana93@gmail.com*

resources such as databases, storage servers, high speed networks, supercomputers and clusters for solving large scale problems have been developed known as Grid Computing. Another reason for the growth of grid computing is the potential for an organization to reduce the capital and operating cost of its computing resources, while maintaining the computing capabilities it requires. This is because the computing resources of most organisations are vastly underutilized, but are necessary for certain operations. Thus the ROI on computing investments can increase through participation in grid computing even though networking and service costs may also increase to some extent.

4 TYPES OF GRID COMPUTING

Grid Computing can be used in a variety of ways to address various kinds of application requirements. Often grids are categorized by the type of solutions that they best address.

4.1 Computational Grid

A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to high-end computational capabilities. This grid provides secured access to huge pool of shared processing power suitable for high throughput applications. The computational grids offer a convenient way to connect many devices which helps in reducing the power consumed and also increases the speed of the systems. Computational grids are used by many organisations namely health maintenance, a material science collaboratory, computational market economy,government etc.

4.2 Data Grid

Data grids provide an infrastructure to support data storage, data discovery, data handling, data publication, and data manipulation of large volumes of data actually stored in various heterogeneous databases and file systems. It can also be defined as a system composed of multiple servers that work together to manage information and related operations such as computations in a distributed environment. An In-memory to achieve very high performance, and uses redundancy to ensure the resiliency of the system and the availability of the data in the event of server failure.

4.3 Scavenging Grid

A scavenging grid is most commonly used with large numbers of desktop machine. Machines are scavenged for available CPU cycles and other resources. Owners of the desktop machines are usually given control over when their resources are available to participate in the grid.

4.4 Collaboration Grid

With the advent of the internet, there has been an increased demand for better collaboration. Such advanced collaboration is possible using the grid. For instance, persons from different companies in a virtual enterprise can work on different components of CAD project without even disclosing their proprietary technologies.

4.5 Network Grid

A network grid provides fault-tolerant and high-performance communication services. Each grid node works as a data router between twon communication points, providing data-chaining and other facilities to speed up the communications between such points.

4.6 Utility Grid

This is the ultimate form of the grid, in which not only data and computation cycles are shared but software of just about any resources are shared. The main services provided through utility grids are software and special equipments. For instance, the applications can be run on one machine and all the users can send their data to be processed to that machine and receive the result back.

5 A CAUSE FOR CONCERN-BIG DATA

5.1 Introduction

While there is a certainty that more and more data will be stored over time, the question is how much data and how fast? The amount of data in our world is exploding and ballooning in size everyday. The amount of information now is measured in zettabytes and petabytes. Over the next decade the number of electronically stored piece of data or files that encapsulate the information in the digital universe and the number of servers managing the world's data stores will grow by ten times!. The data that's growing too big, moves too fast and doesn't fit in the structures of the database architectures. The practice of acquiring, analyzing and interpreting ridiculously huge data is something the technology and business world is much excited about! Such large, complex and unstructured data difficult to handle process and store is what we term as the "Big Data".

5.2 Big Data

Data becomes "big data" when it basically outgrows the current ability to process it, and cope with it efficiently. These are large pools of data that are difficult to be captured, communicated, aggregated, stored and analyzed. Such datasets have size beyond the ability of typical database software tools to capture, store and manage.

5.3 Examples of Big Data

1. RFID systemsgenerate upto 1000 times the data of conventional bar code systems.
2. 10,000 payment card transactions are made every second around the world.
3. Walmart handles more than one million customer transactions an hour.
4. 340 million tweets are sent per day. That's nearly 4,000 tweets per second.
5. Facebook has more than 901 million active users generating social interaction data.
6. More than 5 billion people are calling, texting, tweeting and browsing websites on mobile phones.

5.4 3v's of Big Data

VOLUME, VARIETY and VELOCITY are the three persistent features which describe data in terms of size, speed at which they can be acquired and queried and the wide range of formats and file types generating these data. These are the three main landmarks on the basis of which big data is defined.

1. Volume- It is almost impossible to understand the shear amount of data being generated in various sectors. Many factors contribute to the increase in data volume-transaction based data stored through the years, text data constantly streaming in from

social media and increasing amount of sensor data. Excessive data volume creates issues including storage, how to determine relevance amidst the large volumes of data and how to create value from the data that is relevant.

2. Variety- Variety refers to many different data and file types that are important to manage and analyze more thoroughly, but for which traditional relational databases are poorly suited. Data today comes in all types of formats- from traditional databases to hierarchical data stores created by end users and OLAP systems to text documents, email, meter-collected data, video, audio, stock ticker data and financial transactions.
3. Velocity- The frequency of data generation and data delivery is very fast and furious which varies with different types of data. The rate of change in the data and how quickly it must be used to create real value is what velocity refers to. Traditional techniques are poorly suited to storing and using high-velocity data.
4. Variability- In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage.
5. Complexity- When huge volumes of data are dealt, it comes from multiple sources. It is quite an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or the data can quickly spiral out of control.

5.5 Challenges

The real issue is not about acquiring large amounts of data. It's about what has to be done with the big data. The concern is now about the amount of amassed data that is becoming so large that it is difficult to find the most valuable pieces of information. Organisations have been limited to using subsets of their data and are constrained to simplistic analyses because the sheer volumes of data overwhelm their processing platforms.

1. What if the data volume gets so large and varied and there are no ways to deal with it?
2. Should all the data have to be stored?
3. Should all the data be analyzed?
4. How to find out which data points are really significant?

6 DISTRIBUTED CACHING

6.1 What id Distributed Caching

Caching is a well-known concept. It is a technique to store data in memory so that application is not required to fire a query over the network to communicate with database or service. Communicating over the network always add lots of additional overheads ultimately increase the overall processing time for the request. When using caching, the data is always available in ready to use format within the memory, so that the overhead of network response time and processing time for the request can be reduced ultimately increase the performance. A collection of caches spread across multiple machines and multiple locations that functions as a single cache for the individual applications is known as "distributed caching".

6.2 Need for Distributed Caching

Caching has always been a stand alone mechanism which is no more workable in most environments because applications now run on multiple servers and in multiple processes within each server.

In memory distributed caching is a form of caching that allows the cache to span multiple servers so that it can grow in size and in transactional capacity. It is also a very scalable because of its architecture. It distributes its work across multiple servers but still gives you a logical view of a single cache. For application data, a distributed cache keeps a copy of a subset of the data in the database. This is meant to be a temporary store, which might mean hours, days or weeks. In a lot of situations, the data being used in an application does not need to be stored permanently.

Distributed caching has become feasible now for a number of reasons. First, memory has become very cheap, and computers can be stuffed with many gigabytes at throw away prices. Second, network cards have become very fast, with 1Gbit now standard everywhere and 10Gbit gaining traction. Finally, unlike a database server, which usually requires a high-end unlike a database server, which usually requires a high-end machine, distributed caching works well on lower cost machines, which allows to add more machines easily.

6.3 Advantages of employing Distributed Caching

1. Offloading feed handler transmissions
2. Network connection and bandwidth preservation
3. Service consolidation
4. Failover and recovery

7 DISTRIBUTED CACHING IN GRID FRAMEWORK

In a distributed server environment, the main shortfall of object caching is not capable of offering linear scalability. The other difficulty is synchronization complexity. The complexity increases because consistency between the cached data's state and the data source must be ensured. Otherwise, the cached data can fall out of sync with the actual data, which leads to data inaccuracies.

Distributed cache technology using data grid solves both shortfalls. Most important is the linear scalability through data grid partitioning. In the data grid, the data is spread out over all servers in such a way that no two servers are responsible for the same piece of cached data. This means that the size of the cache and the processing power associated with the management of the cache can grow linearly with the size of the cluster. When there is a request for cached data, the resonse can be accomplished with a "single hop" to another server, if the data object is not found in local cache. This means when more servers are adding to the grid, the performance of the response doesnot decrease.

The data grid also allow to configure number of backups

for the caches, when the primary cache fails, one of the back-ups will takeover, this provides the failover for clustering technologies.

8 Examples of Distributed Caching in Grid Computing

1. Memcached is a high-performance, distributed caching system. Although application-neutral, its's most commonly used to speed up dynamic web applications by alleviating database load. Memcached is used on Live Journals, Slashdot, Wikipedia and other high-traffic sites.
2. GemFire Enterprise is in-memory distributed data management platform that pools memory across multiple processes to manage application objects and behaviour. Using dynamic replication and data partitioning techniques, Gemfire Enterprise offers continuous availability, high performance and linear scalability for data intensive applications without compromising on data consistency, even under failure conditions. In addition to being a distributed data container, it is an active data management system that uses an optimized low latency distribution layer for reliable asynchronous event notifications and guaranteed message delivery.
3. Platform Symphony is the most powerful enterprise-class SOA management softwarefor running application services on scalable, shared, heterogeneous grid. It accelerates a wide variety of parallel applications, quickly computing results while making optimal use of available infrastructure.

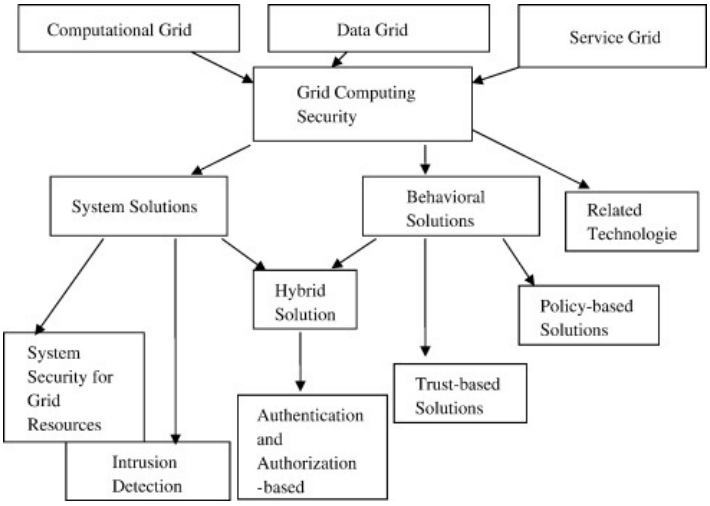
9 A Solution to Big Data

In spite of grid computing being advantageous in many ways, experts have not yet found an exact solution for the computer database to deal with large volumes of data. The process of introducing distributed caching into the grid environment might help in solving the issue of big data storage, management and processing. This would also help in increasing the speed of the systems working on this server. Hence the idea of combining the concepts of distributed caching and grid computing into a single framework will help to increase the efficiency and capability of future computing systems.

10 Security Issues of Grid Computing

Security issues become very important when resources and data are shared in huge amount within the organisations. Data flowing across the different nodes of the grid is very much valuable for its owner, so it should go only to those who are intended to receive it. And therefore there is enormous concern about data and application both during its flow across the Internet. The first concern is mainly because I is possible for someone to tap your data and possibly modify it is possible for someone to tap your data and possibly modify it on its path. The second concern is that when you use others computers in the grid, it is possible that the owners of those computers may read your data. These can be addressed by sophisticated encryption techniques both during transmission and also during their representations/storage on external system in order to authenticate the users. The Grid Security Ifrastructure (GSI) uses SSL certificates for authentication.

Operating systems already provide means to control who is allowed to access data. For example, on UNIX systems there is a support to set permissions such as only the owner of the re



source is permitte to access data.

Figure 2 Grid Computing Security

11 Applications of Grid Computing

The search for Extra Terrestrial Intelligence (SETI) project is one of the earliest grid computing systems to gain popular attention. The vast amount of computing capacity required for SETI radioa signal processing has led to a unique grid computing concept that has now been expanded to many applications. SETI@home is a scientific experiment that uses internet-connected computers to download and analyze radio telescope data for the SETI program. A free computer software program harnesses the power of millions of personal computers, and runs in the background using idle computer capacity. More than 5.2 million participants have ogged over 2 million years of aggregate computing time.

Grid computing is now being used for other applications that include biology, medicine, earth sciences, physics, astronomy, chemistry, chemistry and mathematics.

Another grid computing application is climateprediction.net, which investigates the approximations that have to be made in state-of-the-art climate models by running the models respond to variations in the approximations. These simulations should improve confidence in cimate change predictions that have long-term effects on the global economy

The military in many countries have already started developing the grid technology. The United States has traditionally used their most powerful computers for military applications. But this Virtual Organisation is unlikely to let other users access its grid.

Education involves student, teachers, mentors, parents and administrators and so is very natural application of grid technologies. E-libraries and e-learning centres are already benefitting from grid-based tools for accessing distributed students, resources and tutors.

Global enterprises and large corporations have sites, data, people and resources distributed all over the world. Grids will allow such organisations to carry out large-scale modeling or computing by stimultaneously using the resources at their many sites.

In medical field, access to a grid that could handle administrative databases, medical image archives and specialized instruments such as MRI machines, CAT scanners and cardio angiography devices, enhance the diagnosis procedures, speeds up analysis of complex medical images, and enables life-critical applications such as tele-robotic surgery and remote cardiac monitoring.

Another place where grid computing could be great use is in Governments and International Organisations. Problems like disaster response, urban planning and economic modeling are traditionally assigned to national governemnts or coordinated by International Organisations like the United Nations or the World Bank. These groups could use grid computing and share their data more simply and effectively.

12 CONCLUSION

Grid computing provides high performance data processing service with the help of the integrated computers connected to each other through local area network or through internet. It uses parallel processing and distributed systems technology which are the backbone of high performance computing. It has extensive use in many fields of science and engineering. It this paper we have focused on Grid Computing, Big Data and Distributed Caching. It also tells us about how big data could be solved using distributed caching in grid computing and the applications of grid computing. Grid computing with its characteristics provides cheap and efficient solution for all said issues. Grid computing has yet to be explored further to meet today's challenges.

REFERENCES

[1] Michael P. Cunnings and Jeffry C. Huskamp "New Horizons", 2005
[2] Gargi Shankar Verma, "Grid Computing" (Book Style)
[3] Jonathan Purdy "Data Grids and Service-Oriented Architecture", 2007
[4] Iqbal Khan, "Distributed Caching on the Path to scalabiulity", 2009
[5] Rajkumar Buyya and Srikumar Venugopal, "A Gentle Introduction to Grid computing and Technologies", 2005
[6] Brian Moon "Memcached: what is it and what does it do?", 2009
[7] Ashish Khandewal, "why distributed caching", 2011
[8] Elesevier Science Publishers, "Decision support systems", 2008
[9] James Manyika and Micheal Chui, "Big Data- the next frontier for innovation, completion and productivity", 2011
[10] Margaret Rouse, "Big Data", 2011
[11] O'Reilly Radar, "Big Data Now- current prespective" (Book Style)
[12] The Times of India, article- "Big Data to create a boom in job market"
[13] www.pansas.com
[14] http://azhar-paperpresentation.blogspot.in/2010/04/grid-computing_5337.html
[15] http://memcached.org
[16] http://www.oracle.com/us/technologies/big-data/index.html
[17] http://www.zdnet.com/how-is-big-data-faring-in -the-enterprise-7000002404/
[18] www.platform.com