DISTRIBUTED COMPUTING APPROACHES FOR SCALABILITY AND HIGH PERFORMANCE

MANJULA K A[1] , KARTHIKEYAN P [2]

Abstract:

Distributed computing is a science which solves a large problem by giving small parts of the problem to many computers to solve and then combining the solutions for the parts into a solution for the problem. This distributed computing framework suits to projects, which have an insatiable appetite for computing power. Two such popular projects are SETI@Home and Folding@Home. Different architectures and approaches for distributed computing are being proposed as part of the works progressing around the world. One way of distributing both data and computing power, known as grid computing, taps the Internet to put petabyte processing on every researcher's desktop. Grid technology is finding its way out of the academic incubator and entering into commercial environments. Cloud computing, which is a variant to grid computing, has emerged as a potentially competing approach for architecting large distributed systems. Clouds can be viewed as a logical and next higher-level abstraction from Grids.

*Keywords: Distributed Computing, SETI@Home, Folding@Home, Grid Computing, Cloud Comput*ing.

1. INTRODUCTION

Distributed computing refers to computational decentralization across a number of processors, which may be physically located in different components, subsystems, systems, or facilities [Palmintier, 2002]. This paper discusses Distributed computing concepts and practices.

The computing efforts, say projects to find more effective drugs to fight cancer and the AIDS virus, are so large, and require so much computing power to solve, that they would be impossible for any one computer or person to solve in a reasonable amount of time [Pearson, 2010]. Distributed computing is a means to overcome the limitations of single computing systems.

One flavor of distributed computing has received a lot of attention lately--an environment where we can harness idle CPU cycles and storage space of tens, hundreds, or thousands of networked systems to work together on a particularly processing-intensive problem. An innovative worldwide distributed computing project named SETI@Home, whose goal is to find intelligent life in the universe, has captured the imaginations and desktop processing cycles of millions of users and desktops. Another project, Folding@home uses novel computational methods coupled to distributed computing, to simulate problems in protein folding, millions of times more challenging than previously achieved. This paper gives descriptions of these two projects.

[1] Department of Information Technology, Kannur University, Kerala, India

[2] MES College of Enginerring, Kuttippuram, Kerala, India

This paper also discusses Grid Computing concepts, which is a form of Distributed Computing whereby resources of many computers in a network is used at the same time, to solve a single problem. Grid Computing is the use of hundreds, thousands, or millions of geographically and organisationally disperse and diverse resources to solve problems that require more computing power than is available from a single machine or from a local area distributed system [Lewis, 2010]. This technology has been applied to computationally intensive scientific, mathematical, and academic problems through volunteer computing, and it is used in commercial enterprises for such diverse applications as drug discovery, economic forecasting, seismic analysis, and back-office data processing in support of e-commerce and Web services.

Another topic of discussion in this paper is Cloud Computing. Compared to Grid Computing, Cloud Computing is relatively a newer distributed computing concept, which has become popular recently with the availability of environment like Amazon. Clouds leverages virtualization technology and that makes it distinguishable from Grids. With Cloud Computing, companies can scale up to massive capacities in an instant without having to invest in new infrastructure, which is beneficial to small and medium-sized businesses. Cloud computing users can avoid capital expenditure on hardware, software, and services when they pay a provider only for what they use.

2. DISTRIBUTED COMPUTING

Distributed computing utilizes a network of many computers, each accomplishing a portion of an overall task, to achieve a computational result much more quickly than with a single computer [Kumar & Kaur, 2007]. In addition to a higher level of computing power, distributed computing also allows many users to interact and connect openly. Distributed computing refers to the means by which a single computer program runs in more than one computer at the same time. In particular, the different elements and objects of a program are being run or processed using different computer processors. Distributed computing extends traditional computing by allowing computational components to be distributed across a heterogeneous network and seamlessly interoperating with each other to perform a task [Mangalwede & Rao, 2009]. Distributed computing can be deemed as an attempt to produce a virtual supercomputer out of hundreds or thousands of individual computers.

In a distributed computing setup, there are one or more servers, which contain the blueprint for the coordinated program efforts, the information needed to access member computers, and the applications that will automate distribution of the program processes when such is needed. It is also in the distributed computing administrative servers that the distributed processes are coordinated and combined, and they are where the program outputs are generated.

Distributed system is the concept central to distributed computing. A distributed system is an application that executes a collection of protocols to coordinate the actions of multiple processes on a network, such that all components cooperate together to perform a single or small set of related tasks. A distributed system can be much larger and more powerful given the combined capabilities of the distributed components, than combinations of stand-alone systems. Reliability is a difficult goal to achieve because of the complexity of the interactions between simultaneously running components.

To be truly reliable, a distributed system must have the following characteristics [Rizvi et al., 2010]:

- *Fault-Tolerant*: It can recover from component failures without performing incorrect actions.

- *Highly Available*: It can restore operations, permitting it to resume providing services even when some components have failed.
- *Recoverable*: Failed components can restart themselves and rejoin the system, after the cause of failure has been repaired.
- *Consistent*: The system can coordinate actions by multiple components often in the presence of concurrency and failure. This underlies the ability of a distributed system to act like a non-distributed system.
- *Scalable*: It can operate correctly even as some aspect of the system is scaled to a larger size.
- *Predictable Performance*: The ability to provide desired responsiveness in a timely manner.
- *Secure*: The system authenticates access to data and services

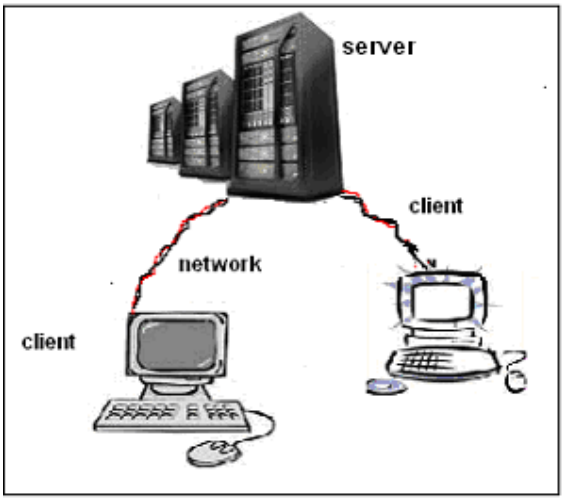

Fig 1: Distributed Computing

Distributed computing is being gradually accepted as a computing method. One popular project, that uses this framework is SETI@Home, a project dedicated to finding signs of extraterrestrial life. Another distributed computing project Folding@Home is dedicated to looking for a cure to cancer.

3. DISTRIBUTED COMPUTING PROJECTS

3.1 SETI@Home

The distributed computed project SETI stands for Search for Extra-Terestrial Intelligence. SETI@home is a scientific experiment that uses Internet-connected computers in the Search for Extraterrestrial Intelligence (SETI). Anyone in the world can participate by running a free program that downloads and analyzes radio telescope data. SETI@home searches through data from the radio telescope at the Arecibo Observatory in Puerto Rico, looking for narrow-bandwidth radio signals that might be taken as evidence of extraterrestrial technology. Radio telescope signals consist primarily of noise (from celestial sources and the receiver's electronics) and man-made signals such as TV stations, radar, and satellites [SETI@home]. There are great challenges in searching across the sky for a first transmission that could be characterized as intelligent, since its direction, spectrum and method of communication are all unknown beforehand.

3.2 Folding@Home

Folding@home is a distributed computing project which uses novel computational methods coupled to distributed computing, to simulate protein folding problems millions of times more challenging than previously achieved [Folding@home]. Proteins are biology's workhorses or its nano machines. Before proteins can carry out these important functions, they assemble themselves, or fold. The process of protein folding, while critical and fundamental to virtually all of biology, in many ways remains a mystery. Moreover, perhaps not surprisingly, when proteins do not fold correctly, there can be serious effects, including many well-known diseases, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, and Parkinson's disease. People from throughout the world, download and run software to band together to make one of the largest supercomputers in the world. Every computer takes the project closer to its goals. Fig.2 shows the growing number of participating CPUs over years.

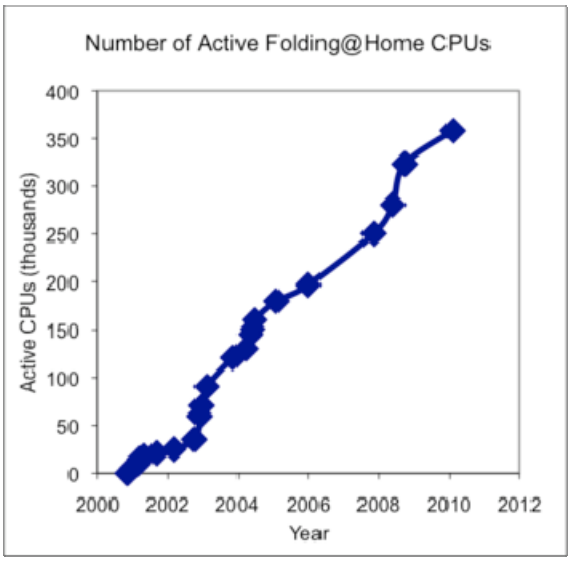

Fig. 2: Active CPUs vs Time of Folding@Home Project

4. GRID COMPUTING

The main concept of Grid Computing is to extend the original ideas of the Internet to sharing widespread computing power, storage capacities and other resources [Yang et al.,2009]. The term, Grid Computing, has become one of the latest buzzwords in the IT industry. Grid Computing can be thought of as distributed and large scale Cluster Computing and as a form of network distributed parallel processing. This innovative approach of computing leverages on existing IT infrastructure to optimize computing resources and manage data as well as computing workloads. Grids are collections of heterogeneous computation and storage resources scattered along distinct network domains. Grids provide tools that allow users to find, allocate and use available resources [Amador et al., 2009]. Grid middleware provides users with seamless computing ability and uniform access to resources in the heterogeneous grid environment [Buyya & Venugopal, 2005]. Structure of a grid is depicted in Fig.3.

A grid consists of the following nodes [SAS(R) 9.2]:

- A grid control server-a machine that distributes jobs to machines on the grid. A grid control server can also do work allocated to the grid.
- One or more grid nodes-a machine or machines that run a portion of the work allocated to the grid.

Grid computing appears to be a promising trend for three reasons [Berman et al., 2003]:

- Its ability to make more cost effective use of a given amount of computer resources.
- As a way to solve problems that cannot be approached without an enormous amount of computing power.
- It suggests that the resources of many computers can be cooperatively and perhaps synergistically harnessed and managed as collaboration toward a common objective.

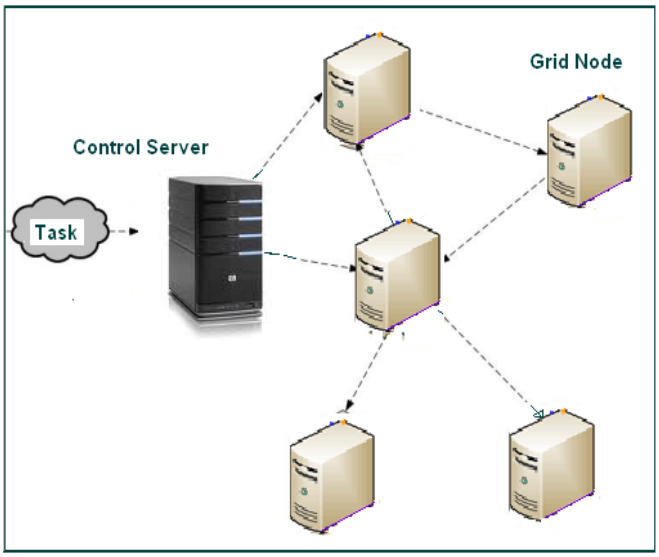

Fig. 3: Grid Computing

Grid computing connects computers that are scattered over a wide geographic area, allowing their computing power to be shared. Just as the World Wide Web enables access to information, computer grids enable access to computing resources. These resources include data storage capacity, processing power, sensors, visualization tools and more. Thus, grids can combine the resources of thousands of different computers to create a massively powerful computing resource, accessible from the comfort of a personal computer and useful for multiple applications, in science, business and beyond. Grids could allow the analysis of huge investment portfolios in minutes instead of hours, significantly accelerate drug development, and reduce design times and defects. Larger bodies of scientific and engineering applications stands to benefit from grid computing, including molecular biology, financial and mechanical modeling, aircraft design, fluid mechanics, biophysics, biochemistry, drug design, tomography, data mining, nuclear simulations, environmental studies, climate modeling, neuroscience/brain activity analysis, astrophysics [Kaufman et al., 2003].

5. CLOUD COMPUTING

Cloud Computing evolves from grid computing and provides on-demand resource provisioning. Cloud computing has emerged as potentially competing approach for architecting large distributed systems [Myerson,

2009]. Cloud Computing is the convergence and evolution of several concepts from virtualization, distributed application design, grid, and enterprise IT management to enable a more flexible approach for deploying and scaling applications.

To deliver a future state architecture that captures the promise of Cloud Computing, architects need to understand the primary benefits of Cloud Computing [Bennett et al., 2009]:

- Decoupling and separation of the business service from the infrastructure needed to run it (virtualization).
- Flexibility to choose multiple vendors that provide reliable and scalable business services, development environments, and infrastructure that can be leveraged out of the box and billed on a metered basis—with no long term contracts.
- Elastic nature of the infrastructure to rapidly allocate and de-allocate massively scalable resources to business services on a demand basis.
- Cost allocation flexibility for customers wanting to move capital expenditure into operating expenditure.
- Reduced costs due to operational efficiencies, and more rapid deployment of new business services.

Cloud computing encompasses any subscription-based or pay-per-use service that, in real time over the Internet, extends IT's existing capabilities. Cloud computing users can avoid capital expenditure on hardware, software, and services when they pay a provider only for what they use.

Cloud computing eliminates the costs and complexity of buying, configuring, and managing the hardware and software needed to build and deploy applications, these applications are delivered as a service over the Internet (the cloud). Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services [Fig.4]. Cloud computing incorporates infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS) as well as Web 2.0.

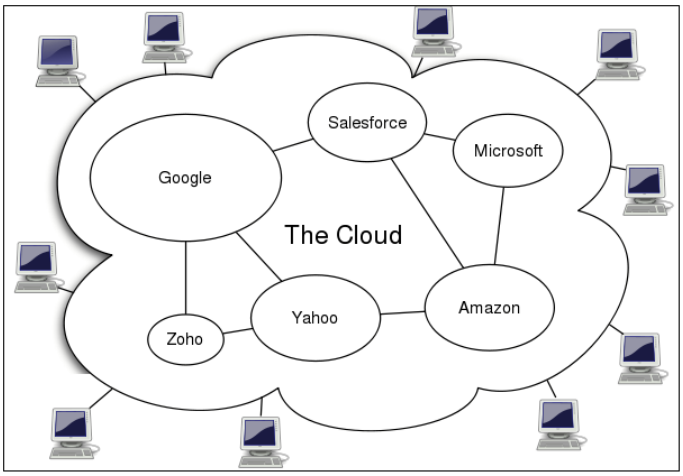

Fig. 4 Cloud Computing

Cloud computing is massively scalable, provides a superior user experience, and is characterized by new, internet-driven economics [Myerson, 2009].

6. GRID COMPUTING vs CLOUD COMPUTING

Cloud Computing and Grid Computing do have a lot in common; both are scalable. Both computing types involve multitenancy and multitask, meaning that many customers can perform different tasks, accessing a single or multiple application instances [Myerson, 2009]. Cloud and grid computing provide service-level agreements (SLAs) for guaranteed uptime availability of, say, 99 percent. At the same time Cloud Computing and Grid Computing do have differences. One major difference being that while grids are typically used for job execution, clouds are more often used to support long-serving services. Grids provide higher-level services that are not covered by Clouds - services enabling complex distributed scientific collaborations (i.e. virtual organisations) in order to share computing data and ultimately scientific discoveries.

Grid systems are designed for collaborative sharing of resources belonging to different admin domains, while Clouds at the moment expose the resources of one domain to the outside world. Grid systems support the execution of end-users applications as computational activities; A typical computational activity once accepted by a Grid endpoint, is locally handled by a batch system as a batch job; Clouds are mainly used for the remote deployment of services.

Grids provide more domain-specific services; Grids are moving towards the adoption of virtual machine technologies, but the usage pattern will be the same (the submitted job is bound with the execution environment as VM image).Grid systems support large set of users organized in virtual organizations where Cloud systems support individual users.

A comparison of grid (EGEE Grid -Enabling Grids for EsciEnce Project) and cloud (Amazon cloud) is depicted in Fig.5, which points out to the power of cloud computing [Klems, 2008].

	EGEE Grid	*Amazon Cloud*
Target Group	Scientific Community	Business
Service	Short-lived batch-style processing (job execution)	Long-lived services based on hardware virtualization
SLA	Local (between EGEE project and the resource providers)	Global (between Amazon and users)
User Interface	High-level interfaces	HTTP(S), REST, SOAP, Java API, BitTorrent
Resource-side Middleware	Open Source (Apache 2.0)	Proprietary
Ease of use	Heavy	Light
Ease of Deployment	Heavy	Unknown
Resource Management	Probably similar	
Funding Model	Publicly funded	Commercial

Fig 5: Comparison of EGEE Grid and Amazon Cloud

The Fig 6 shows the result after analysing the trends in search volume and news reference volumes of computing terms Grid Computing and Cloud Computing. It is seen that the term Grid Computing, which has

been around for a while is seen trending downwards. But, the newcomer Cloud Computing, which made its full entrance into this trend analysis around 2007 and is rapidly gaining momentum

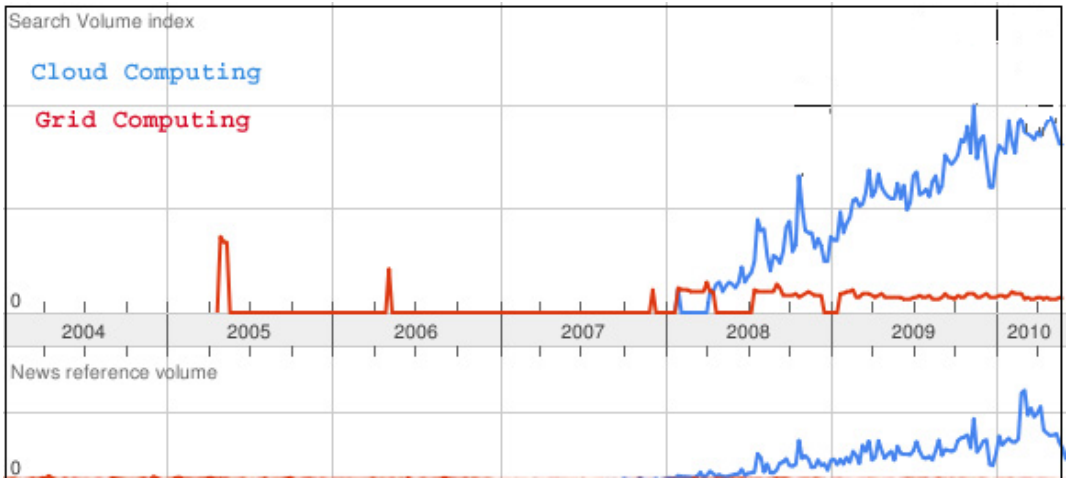

Fig.6. Analysis of Trends in Cloud Computing and Grid Computing

7. CONCLUSION

Distributed computing deals with the development of applications that execute on different computers interconnected by networks. There are many interesting projects proceeding around the world that uses novel computational methods coupled to distributed computing, to solve problems like, simulating protein folding or searching life in universe, millions of times more challenging than previously achieved Grid computing is a form of distributed computing. Computational Grids act as popular platforms for deploying large-scale and resource-intensive applications. Another related technology, the newly emerging IT delivery model—Cloud Computing—can significantly reduce IT costs & complexities while improving workload optimization and service delivery. Cloud Computing is found to be massively scalable, provides a superior user experience, and is characterized by new, internet-driven economics. Grid Computing and Cloud Computing resemble on some respects, but there are differences too. Clouds can be viewed as a logical and next higher-level abstraction from Grids [Jha et al., 2008].

REFERENCES

[1] Amador, G., Alexandre, R., & Gomes, A., 2009: Re-engineering Jake2 to work on a grid. Retrieved from http://www.av.it.pt/conftele2009/Papers/96.pdf
[2] Bennett, S., Bhuller, M., & Covington, R., 2009: Oracle White Paper in Enterprise Architecture – Architectural Strategies for Cloud Computing. Retrieved from: www.oracle.com
[3] Berman, F., Fox, G. & Hey, A. J. G., 2003: Grid Computing: Making the Global Infrastructure a Reality - John Wiley and Sons.
[4] Buyya, R., & Venugopal, S., 2005:A Gentle Introduciton to Grid Computing and Technologies. CSI Communications, July, 2005.
[5] Folding@home: http://folding.stanford.edu/
[6] Jha, S., Merzky, A., Fox, G., 2008: Programming Abstractions for Clouds. Proceedings of Conference Cloud Computing and its Applications, Chicago, 2008.
[7] Kaufman, J. H., Lehman, T. J. & Thomas, J., 2003: Grid Computing Made Simple. The Industrial Physicist, vol. 9(4), pp. 31-33.
[8] Klems, M., 2008: Comparative study: Grids and Clouds, Evolution or Revolution. Retrieved from: www.eu.egee.org
[9] Kumar, R. & Kaur, N., 2007: Jobscheduling in Grid Computers. Retrieved from http://www.rimtengg.com/ coit2007/ proceedings/pdfs/111.pdf
[10] Lewis, M., 2010: Grid Computing. Retrieved from; http://grid.cs.binghamton.edu. Grid Computing Research Laboratory, State University of New York (SUNY) Binghamton
[11] Mangalwede, S. R. & Rao, D. H., 2009: Performance Analysis of Java-based Approaches to Distributed Computing International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009, pages 556-559.

[12] Myerson, J.M., 2009: Cloud computing versus grid computing. Retrieved from: http://www.ibm.com
[13] Palmintier, B., Kitts, C., Stang, P., & Swartwout, M., 2002: A Distributed Computing Architecture for Small Satellite and Multi-Spacecraft Missions. SSC02-IV-6 -16th Annual AIAA/USU Conference on Small Satellites.
[14] Pearson, K., 2010: Distributed computing: http://distributedcomputing.info/ what is distributed computing?
[15] Rizvi, S. A. M., Hussian, Z, & Sharma, V., 2010: Distributed Media Player. Proceedings of the 4th National Conference; INDIACom-2010 Computing For Nation Development, February 25 – 26, 2010, New Delhi.
[16] SAS(R) 9.2 Intelligence Platform: Application Server Administration Guide, Retrieved from: http://support.sas.com/documentation/cdl/en/biasag/61237/HTML/default/viewer.htm#/documentation/cdl/en/biasag/61237/HTML/default/grid.htm
[17] SETI@home: http://setiathome.berkeley.edu
[18] Yang, C.T., Han, T.F., & Kan, H.C. 2009: G-BLAST: a Grid-based solution for mpiBLAST on computational Grids. Concurrency and Computation: Practice and Experience, vol. 21, no. 2, pp. 225-255.