International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 6 No. 1 Aug. 2014, pp. 63-67

http://www.ijisr.issr-journals.org/

# Cluster Computing for Processing Large Data Sets

***Asma Khatoon***

Software Engineering,
Department of Computing and Technology, Iqra University Islamabad, Pakistan



**ABSTRACT:** In this research we propose a cluster for handling the existing processing problems in distributed environments. The volume of data transmitted through online service provider facility is already measured in terabytes. A single customer support outlet sends approximately 2 GB data in one day. Normal digital computers face instances like resources, cost and processing time for the processing of such large amount of data. Here we propose Cluster computing as solution of processing speed problem in processing huge amount of data transmitted. Experiment results shows that processing speed is directly proportional to computational power.

**KEYWORDS:** Cluster Computing, Processors, Large Data Sets, Real Time, Speed, Distributed Environments, System, Transmitted.

## 1 INTRODUCTION

With the expansion of business in an organization, we need increased processing power, reduced cost, and scalability to process gigantic sized data we get every day from online service provider facility. Also, there is need for easy file sharing, and 24/7 availability. With the traditional network installed, there are limitations of processing power, scalability and availability. In this research we have conducted experiments on a medical advisory company. To answer patient's queries and provide them quick response in real time, the biggest problem is to search and share information, at a rapid pace, anytime from anywhere. We have been facing main challenge of processing speed in existing system. Currently we have installed mainframe server with 60 GHz processing speed.

Cluster computing is an efficient way for an organization for the processing of a very large amount of data. That gives same data processing speed as we will obtain from super computers to process large amount of data in a very cheap cost. In this process we connect many computers with one another simultaneously, through a local area network, to get high processing data speed as we have obtain from super computers within our cost limit. We can make cluster of computers using existing computer systems connected through network or local area network. There are used multiple computing sources to make an efficient computer cluster and these are connected with each other in a cluster to make a single powerful system for the processing of large amount of data. The multiple computing elements used in computer cluster are storage devices and processors respectively. We make computer clusters, because of parallel processing speed. It can become thrice as fast as current speed. Cluster computing is a very cheap technology which accelerates data processing speed. In this technology, desktop computers are used.

Following are main reasons why we choose cluster computing as a solution to computing gigantic data of our company.

- In cluster, multiple processors can execute in parallel resulting increase in speed to a greater extent.
- Cluster computers provide a high speed computation facility.
- Computer clusters is a very cheap technology in accordance with the cost point of view. Because it provides multiple applications and functionality which utilizes existing resources.
- Cluster computing technology is the best resource to utilizes existing applications.
- It is a flexible structure because we can add more elements in it if we feel need to do this.

**Corresponding Author:** Asma Khatoon 

Our research is backed up by experimental results we conducting at medical advisory company. We installed a cluster of computers having 8 Dual Core 128 bit 8.2-GHz Pentium processors. We examined results obtained and compared with the results from existing system and we found that cluster computers can solve problem of processing gigantic data in real time hence cost effective and time saving.

This paper comprises of introduction in which we discussed problems with the current system, our proposed solution, merits and demerits of proposed system. In related work we described how other people used systems like our proposed system and the analysis of success and failure they achieved. In fourth section of the paper we discussed framework overview of the proposed system which is followed by technique used for our experiment and experimental results. Conclusion, future work and references are given at the end.

The research paper has organized as; Section I elaborates Introduction .Section II elaborates related work. In this section, we provide some related work in the area of computer clustering. Section III elaborates frame work overview. In this section, a typical framework of middleware has been proposed for cluster computing to process patient's data. Section IV elaborates technique we have used in this. Section V discusses the Experimental Results and the paper is concluded in section VI.

## 2 RELATED WORK

In this section, we provide some related work in the area of computer clustering. In Cluster Computing R&D in Australia, Asian Technology Information Program ATIP, the overview of clustering computing research and development in Asia Pacific region is given. The report itself is divided in three parts software research part, cluster computing projects and cluster systems; the main focus is on Australian academic research. It also gives us a summary of some reports which tells us about the research going on cluster computing in Australia. With the passage of time cluster computing is becoming very popular and cluster computing technology will be a primary area of research for computing community.

In Nexus a low level substrate is proposed which allows different framework to share same resources providing isolation among them which free them implements their own programming model. As new frame works and programming models emerge in a same cloud they still need to share resources and data sets. To share resources between different frameworks is a difficult job because each framework has its own way of performing job execution and resource management. In Nexus job execution management is isolated from resource management by providing a resource management layer over different frameworks. As slot is provided by Nexus on which each framework run its task and perform arbitrary work.

In Integrating IS Curriculum Knowledge through a Cluster-Computing Project, a cluster computing project is developed to expose business students to concepts of parallel and distributive computing and security networking. In designing this project, the concepts of OS, networking, parallel computing are used. The basic objective of Integrating IS Curriculum Knowledge through a Cluster-Computing Project is to introduce the concept of distributive and cluster computing in business.

## 3 FRAMEWORK OVERVIEW

In this section, a framework of cluster computing to process a patient's data is proposed. Data received from help desk and online system is given to application layer. The Application is the top most layer of a cluster computing architecture and user request is received by this layer. The request can be storage or computational job. The middleware is the 2nd Layer which is again divided into sub-layers the framework is given in figure 1.

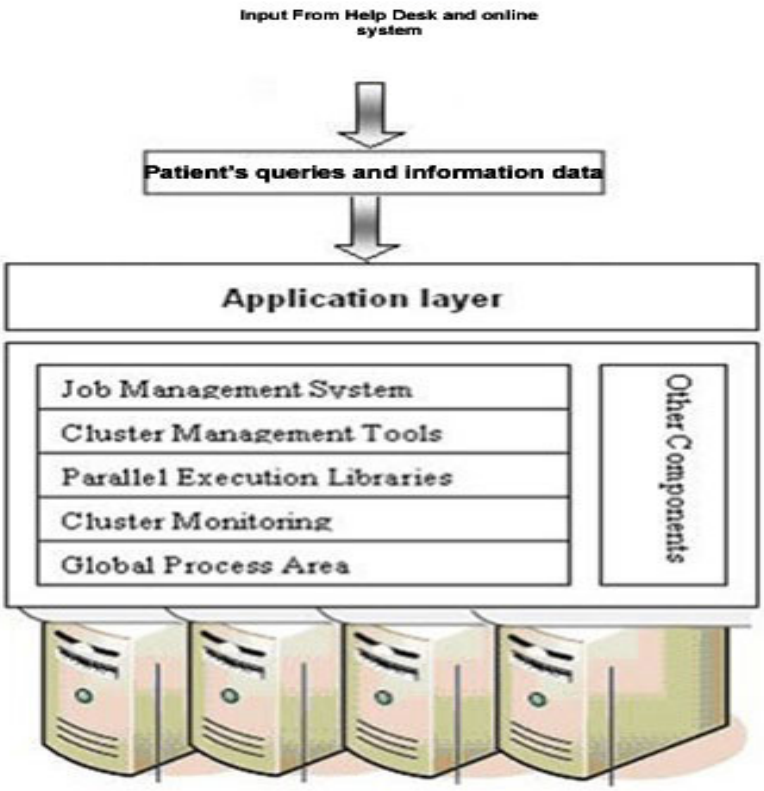


***Figure 1: Framework Overview***

## 4 TECHNIQUE

The user query for the certain information is taken up by first layer of computational cluster. i.e. top most layers these services are known as ontology services. The ontology services are treated by the Job Management System that first takes the refined job and then in next step divided it into independent subtasks and after that these sub tasks are divided among the multiple nodes of the Job Management System. In the above mentioned all the activity the process of recursion is assumed to be involved in the breakdown of the major problems into sub tasks for reliable performance. All of these processes are then to be divided among all computational cluster nodes. After this, distributed and multi agents has to be verify the required knowledge from knowledge based cache and knowledge based replica. If our required information is to be found here in this step at computational cluster then it is no need for the inquiry of the data cluster here and if still our required output could not found here, then there is a need for the refinement of the query by the agents. Then after the refinement of the query it is to be forwarded to the data cluster. Then Data extraction from that cluster is to be performed and finally the obtained output is returned in Global Process area to the KRE. This KRE consists of very important data mining steps. These steps are transformation, matching of patterns, and extraction of knowledge modules (knowledgebase), Data mining query and cleaning. KRE synchronized and update the knowledgebase.

Our Cluster test bed consists of 8 Dual Core 128 bit 8.2-GHz Pentium processors, 4 GB RAM, 1TB HARD-DISK computers connected by a HP 2524 Fast-Ethernet switch. All the processors had a common memory and they all are executed through kernel level code. The module interconnection bus (MBus) is a 40 Mhz 64-bit bus that is capable of sustained bandwidth of 80 Mbytes per second and a peak of 320 Mbytes per second. All machines consist of processor and each processor comprises on 256KB second level cache and 8KB two way associative instruction caches. The data transfer rate in a single moved is recorded up to 128 bytes. The processors in our experiment environment are homogeneous i.e. all processors will have same architecture and speed. It uses Linux cluster operating system and cost approximately $20,000.

## 5 EXPERIMENTAL RESULTS

On the base of experimentation we can conclude that processing speed is directly proportional to computational power. The main input parameter for our experiment is a block of 1GB data.

The performance of the 4 processor clusters with a block is 2.24 Gflops(1 Gflop = 1 billion floating point operation per second) while performance of 60 Ghz mainframe system is 1.4 Gflop.

***Table 1: Processing speed***

| System | Processors speed(GHz) | Performance (Gflops) |
|---|---|---|
| Clusters | 65.6 | 2.24 |
| Mainframe | 60 | 1.4 |

Graph can be shown as:

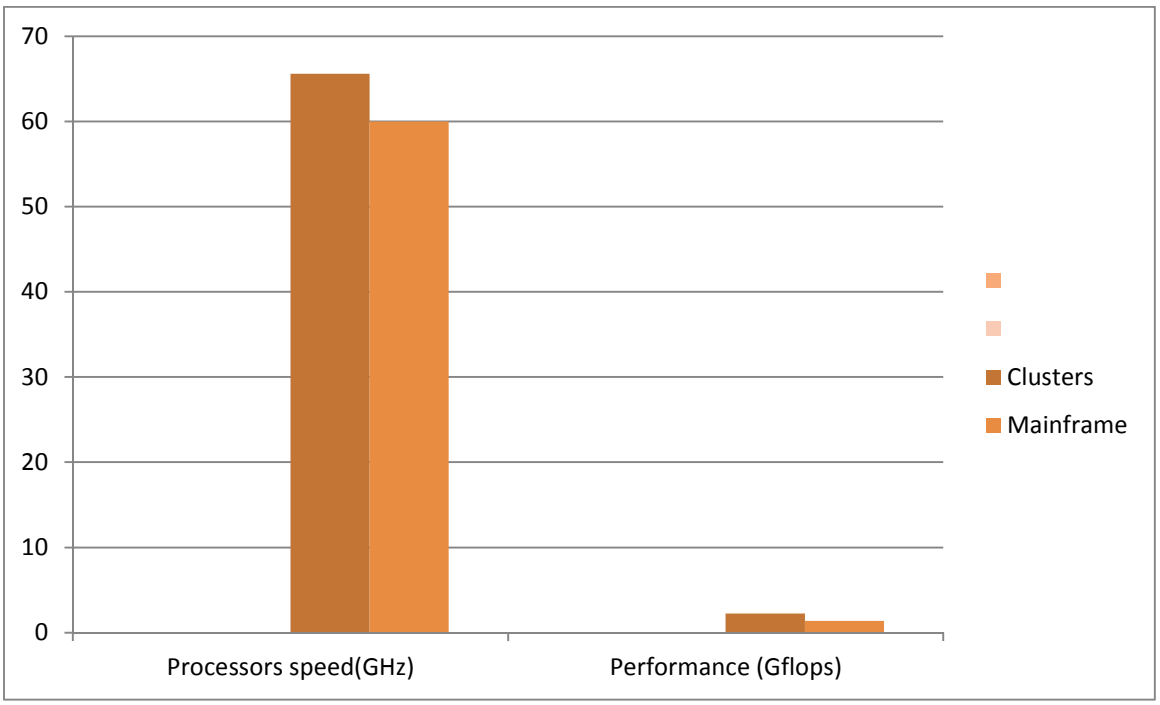


***Figure 2: The performance of 4 processors***

***Table 2: Cost***

| System | Cost |
|---|---|
| Cluster | $20,000 |
| Mainframe | $25,000 |

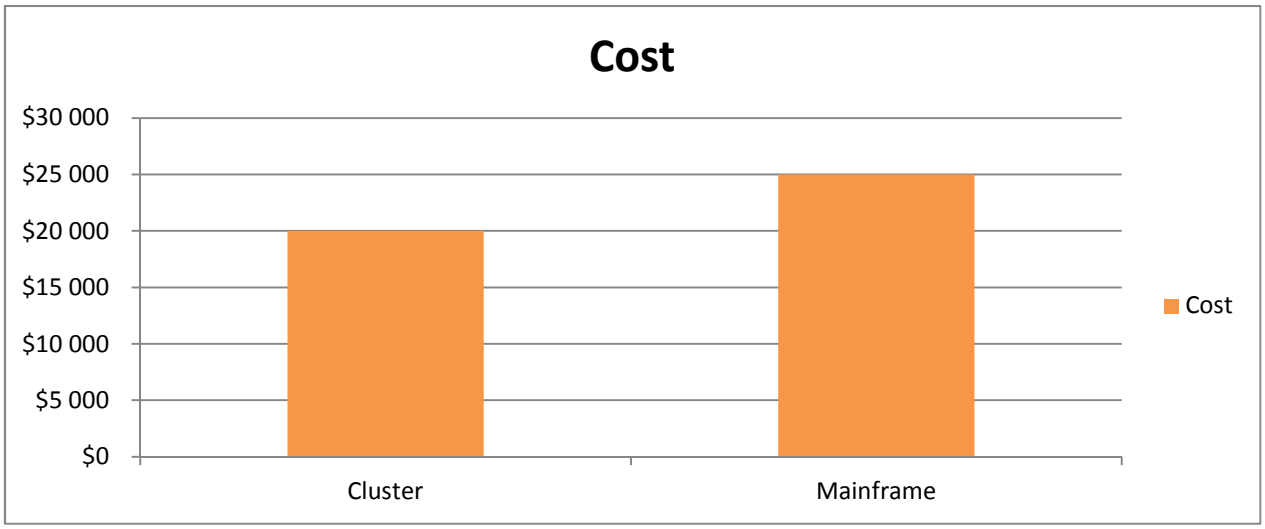


***Figure 3: Comparison of cost between Cluster and Mainframe***

## 6 CONCLUSIONS

In this research article we have discussed that how cluster computing is highly robust technique to process large data sets. Here we have proposed a framework for processing large data sets in medical advisory company. We have discussed here a sample of these application areas and how they benefit from the use of clusters.

We found that each node of cluster has very high computational capabilities but interconnecting switch is a bottleneck for speed. In future we'll conduct research to improve computational power by introducing some middleware.

## ACKNOWLEDGMENT

First, I am thankful to Allah Almighty, whose support and strength helped me in bringing this work to the end. After that I am thankful to my teacher Dr Ataul Aziz Ikram for his support and guidance.
Finally, thanks to my parents who endured this long process with me, always offering support and love. Without their support I would not be able to complete this work.

## REFERENCES

[1] Nodine et al., 1997; Duschka and Genesereth, 1997.
[2] Cluster Computing R&D in Australia, Asian Technology Information Program (ATIP)
[3] A Common Substrate for Cluster Computing Benjamin Hindman, Andy Konwinski, MateiZaharia, Ion Stoica, University of California, Berkeley.
[4] Liuping Wang, "Model Predictive Control System Design and Implementation Using MATLAB", Springer, Girona, Spain, 2009
[5] Integrating IS Curriculum Knowledge through a Cluster-Computing Project –A Successful Experiment Fred L. Kitchens, Sushil K. Sharma, and Thomas Harris Ball State University, Muncie, IN, USA.
[6] M.A. Baker, G.C. Fox, and H.W. Yau. Review of Cluster Management Soft-ware. NHSE Review, May 1996.
[7] J. Dongarra, G. Fagg, A. Geist, A. Gray, et al, HARNESS: a next generation distributed virtual machine, Journal of Future Generation Computer Systems, (15), pp. 571-582, Elsevier Science B.V., 1999.
[8] F. Mueller. On the Design and Implementation of DSM-Threads. In Proceed-ings of the PDPTA'97 Conference, Las Vegas, USA, 1997.
[9] H. Custer. Inside Windows NT. Microsoft Press, NY, 1993. [10] M.Saito and M. Yamakita, MPC for a Simplified Transmission Model with Backlash Using UKF, International Conference on Control Applications, Munich, Germany, October 4-6, 2006.
[10] K. Hwang and Z. Xu. Scalable Parallel Computing: Technology, Architecture, Programming. WCB/McGraw-Hill, NY, 1998
[11] C. Koelbel et al. The High Performance Fortran Handbook. The MIT Press, Massachusetts, 1994