International Journal of Conceptions on Computing and Information Technology
Vol.2, Issue 1, Jan' 2014; ISSN: 2345 - 9808

Issues in adapting cluster, grid and cloud computing for HPC applications

D A Prathibha
Assistant Professor, Dept. of IT,
Sri Sai Ram Engineering College,
Chennai, India.
Somaprathi25@gmail.com

Dr. B Latha
Professor & Head, Dept. of CSE,
Sri Sai Ram Engineering College,
Chennai, India.

Dr. G Sumathi
Professor & Head,
Dept. of Information Technology,
Sri Venkateswara College of Engineering,
Chennai, India.

Abstract— **Currently computing industry is facing the problem of storing and analysis of big data that cannot be processed by traditional data processing frameworks. In addition to this there are lot of applications which require the use of streaming of data either offline or online. For the applications which are adhoc in nature use of dedicated supercomputers makes it unprofit for the organization. In this paper we are investigating the different alternatives for running these adhoc data intensive applications .We are considering different options like usage of cluster computing, Grid computing ,cloud computing. In case of cloud platform different approaches like Public cloud usage or deploying at IaaS level or PaaS Level is compared. In cloud computing resource allocation plays major role and the issues of resource allocation specific to data intensive application are identified.**

Keywords- Big Data, Cloud Computing, Grid Computing, IaaS, PaaS, Cluster Computing.

I. INTRODUCTION

As the cost of the information storage devices is reduced drastically in this decade there is also shift in the way the data is generated and used. To days challenge for the corporate is to manage the information which is not in terms of tera bytes but it is very large in terms of peta bytes, exabytes or it may be Zetta bytes and this huge volume of data is referred as big data. Big data refers to large data sets that are challenging to store ,save and analyze unlike relational databases.

There are lot of commodity devices which are producing live stream of data that is been produced by different sensors attached to these devices. One of the important use of these live streams is television channels which are giving the option to their audience to report the news stories 24x7 days. In this scenario the application has to handle massive data which may be more than terabytes of data. One of the interesting live streaming application the prestigious MARS rover[1] called curiosity which was shown live on the TV watched by millions of people worldwide was actually been tested with cloud computing platform to check whether the live telecast will be able to perform well with that many requests from worldwide. Even the London 2012 Olympics was also been tested and deployed on cloud computing platform. Scientific community also requires access to high end data processing frameworks as they need to simulate High Performance Computing(HPC) application on large data sets for example scientists working on climate analysis have to process the data in terms of peta bytes. Many of these climate data analysis programs are parallelizable in a data driven manner and these programs sometimes contain software dependencies. Traditionally scientists run these programs on massively parallel supercomputers. Running the programs on supercomputer has limitations like supercomputing facility is machine architecture dependent, migrating to new supercomputing facility has significant overhead, Hence alternatives to supercomputers like cluster, grid computing and cloud computing are explored in this paper.

So the challenge for the today's data processing frame works is to handle this huge amount of data which cannot be handled by traditional data processing frame works by using fixed set of hard disk arrays. The important characteristics of stream based application are that data arrives at continuously and the processing speed must match with the arrival rate. One challenge is that data intensive applications may be built upon conventional frameworks, such as shared-nothing Database Management Systems (DBMSs), or new frameworks, such as Map Reduce , and so have very different resource requirements. A second challenge is that the parallel nature of large- scale data-intensive applications requires that scheduling and resource allocation be done to avoid data transfer bottlenecks. A third challenge is the need to support effective scaling of resources when large amounts of data are involved.

The rest of the paper is organized as follows section II gives the characteristics of data intensive applications. In section III, IV and V we present issues of using Cluster, Grid and Cloud platform for HPC applications. Followed by conclusion and future work is presented.

II. CHARACTERISTICS OF DATA INTENSIVE APPLICATIONS

Massive parallel processing applications can be classified as data intensive if the application processes data in terms of tera bytes , peta bytes or zetta bytes. These applications require to work with large data sets and data transfer hence these applications are highly network bound. A subset of data intensive application is Data streaming applications are becoming more popular recently, such as astronomical observations, large-scale simulation and sensor networks, which brings new challenges to resource management. Characteristics of such applications are that (1) they are continuous and long running in nature; (2) they require efficient data transfers from distributed data sources (3) it is often not feasible to store all the data in buffer for later processing because of limited storage and high volumes of data to be processed; (4) they need to make efficient use of high performance computing (HPC) resources to carry out computation-intensive tasks in a timely manner.

III. CLUSTER COMPUTING FOR HPC APPLICATIONS

A computing cluster is a group of servers dedicated to sharing an application's workload. Servers in the cluster run on a homogeneous environment that includes both an up-to-date version of the runtime environment (application and binary dependencies) and shared access to I/O files. Having a dedicated computing environment such as a computing cluster eliminates the need for virtualization and offers effective central administration of the computing cluster.

Following are the limitations in adapting clusters for adhoc data intensive applications:

i. Cluster environment is static in nature hence it is not suitable for implementing adhoc data intensive applications.

ii. In cluster computing, a bunch of similar(or identical) computers are hooked up locally to operate as a single computer. The computers that make up the cluster cannot be operated independently as separate computers.

iii. Significant software development using cluster API is required for migrating existing applications to cluster platform.

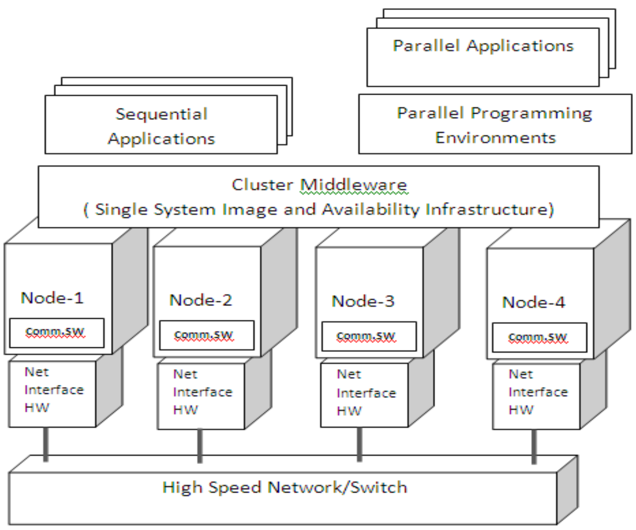

Fig. 1: General Architecture of a cluster

IV. GRID COMPUTING FOR HPC APPLICATIONS

Grid computing is similar to cluster computing in the sense that it involves a group of computers dedicated to solving a common problem, but differs from cluster computing by allowing a mixture of heterogeneous systems (different OSs and hardware) in the same grid. Working of grid computing is similar to the way electricity is distributed to the house hold from central power grid. Grid was basically developed for running HPC applications by sharing resources which are geographically distributed. Fig2 shows the architecture of grid computing. Here the goal is to access computers only when they are needed and to scale the problem so that even small computers can make a useful contribution. Currently there are lot of computational

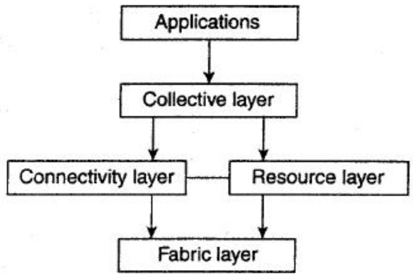

Fig. 2: Architecture of Grid computing

TABLE I. DATA STREAMING APPLICATIONS ON GRID ENVIRONMENT.

S. No	Grid based data stream application	Features
1	GATE (Grid-based Adaptive Execution on Streams)	Applications involving high-volume data Streams. A resource allocation algorithm based on minimal spanning tree (MST) to create a deployment configuration.
2	Streamline	It uses scheduling heuristicalgorithm which considers dynamic nature of grid and varying demands of streaming applications.
3	Pegasus	Creating job workflows for grid environment. It handles data transfers, job processing and data cleanups in a workflow manner, not in an integrated and cooperative way.
4	EnLIGHTened	It can work well with high end applications and reducing the bottlenecks posed by these applications. Collocate any type of resource on demand or in advance.
5	G-lambda	Uses Web service approach. This is a general framework for resource co-allocation, which does not pay enough attention to characteristics of data streaming applications, such as sustaining and controlled data provision

problems like Biomedical applications, Industrial research problems, Engineering research problems deployed on various grids like CERN ,TERA and GARUDA Grid systems can be commercial or open source.

The first three framework Gate, Streamline and Pegasus considers only computational resources and storage resources, it does not consider network resources and integrated resource management as it is considered in EnLIGHTened and G-lambda.

V. CLOUD COMPUTING FOR HPC APPLICATIONS

Cloud computing is the internet based computing in which everything (like hardware, software, Platform) are provided as services on demand. The main characteristic of the cloud computing is it is pay per use model .Cloud computing is on the top ten technologies in Gartner's list of the ten most look for technologies for the next year. The US National Institute of Standards and Technology has working definition which summarizes cloud computing[4] as "cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."Cloud computing concept can be compared to like a person instead of owing car taking call taxi when ever needed as for a common man it is not affordable to financially own a car and manage the expenses .So here the organizations do not invest in buying the resources instead they are obtained from the resource providers like Amazon on demand. Big players such as Amazon, Google, IBM, Microsoft and Sun Microsystems have begun to establish new datacenters for hosting Cloud computing applications in various locations around the world to provide redundancy and ensure reliability in case of site failures. In this model, the typical application comprises a thin client code local to the end user and a main server code local to the data-compute resource. The user application requests resource allocations from the cloud, and the provider generates a virtual system image that satisfies the resource request. The user pays only for resources consumed, not for operational and infrastructure costs.

Recently there has been interest in applying this computing technology to solve large scientific and industrial problems. The primary benefit of moving to Clouds is application scalability. Unlike Grids, scalability of Cloud resources allows real-time provisioning of resources to meet application requirements at run-time or prior to execution. The elastic nature of clouds facilitates changing of resource quantities and characteristics to vary at runtime, thus dynamically scaling up when there is a greater need for additional resources and scaling down when the demand is low. Table 2 lists the comparative features of cloud computing, grid computing and cluster computing.

TABLE II. COMPARISION OF CLUSTER ,GRID AND CLOUD COMPUTING FOR HPC APPLICATIONS

Cluster Computing	***Grid Computing***	***Cloud Computing***
In cluster ,static set of computing resources are used.	In grid computing the systems are heterogeneous (both in OS and Software)	Provides resources as per customer request on demand by using Virtualization techniques.
The advantage of cluster is that there is no overhead in resource provisioning as there is no virtualization	Not enough tools and technologies for marinating QOS parameters in SLA	In order to avoid under provisioning or over provisioning resources are either allocated or deallocated as per request hence there is a cost of overhead incurred due to virtualization .
Not Suitable for computing problems which are adhoc in nature	Grid computing systems becomes high cost solution for implementing parallel distributed solution.	Well suited for HPC applications which are adhoc in nature

A.IaaS service for HPC applications

IaaS service provides the users with the infrastructure resources like processor, memory on demand. Virtual machines are used to provide the users with the dedicated systems which share underlying hardware. Amazon EC2 is a big player in this field and provides resources in the sizes like small, medium and big. There has been lot of work done in deploying HPC applications on IaaS clouds. In this section we will present salient features o f the research work carried out in this area Warneke and Odej[5] have identified the opportunities and challenges present for adhoc parallel processing applications in IaaS clouds. They have designed parallel data processing framework called Nephele which considers dynamic and heterogeneous nature of IaaS cloud. The important feature of Nephele is it is able to split the parallel processing application into stages ,and for each stage depending on the requirement virtual machines are allocated/deallocated. The application is represented using DAG. There work also includes comparing the performance of a application on Nephele and Hadoop Map Reduce Frame work. The results have shown that assigning specific machine to a particular virtual machine and automatically allocate/deallocate the virtual machine will improve the resource utilization and reduce the cost.

Philip and Andrzej [6] have done performance study on running HPC applications on IaaS Clouds and on Cluster. The main focus of their work is analyzing the effect of virtualization and network communication on the performance of the HPC application. They have identified that the performance of the HPC application varies with the type of the hypervisor. There experimental results show that communication bound HPC application should be run on high speed inter connect clouds. One of the major issue in adapting public cloud for scientific application is the time taken for data transfer.

Bjorn Lohrmann et al [7] have carried out extensive work on analyzing the common design principles of massively parallel data processing application.There work mainly concentrates on running live data streaming applications with Nephele.The common design principles like master-slave pattern and the communication model is producer-consumer pattern are identified. Significance of large output buffers is shown as throughput will effect if individual data items are transferred. They have identified that the size of the output buffer and creation rate effects the latency of the data item. There work also includes automatic QOS optimization techniques with the techniques for adaptive output buffer sizing and dynamic task chaining to meet the user defined latency constraints for a given workflow. It has been shown that these strategies improve the workflow latency by a factor of 15 while preserving the required data throughput.

B.PaaS support for HPC applications

Lot of research is carried out in using IaaS service for deploying parallel processing applications on cloud platform. Recently efforts have been made to use platform as a service (PaaS) layer for these applications .In this section work carried out in this outset is presented.

Radu Prodan et al [8] have evaluated performance of executing HPC applications on one of the popular PaaS tool Google App Engine. GAE is a simple parallel computing framework that supports development of computationally intensive HPC algorithms and applications. The underlying Google infrastructure transparently schedules and executes the applications and produces detailed profiling information for performance and cost analysis. They have developed a java based generic Parallel Computing Framework to support the development of parallel applications with GAE. To implement an application it needs to be represented as Master Slave architecture with three abstract classes are required JobFactory, WorkJob and Result.The master application is a Java program that implements JobFactory on the user's local machine. JobFactory manages the algorithm's logic and parallelization in several WorkJobs. WorkJob is an abstract class implemented as part of each slave application.There work also includes cost analysis of resource provisioning overhead on some sample parallel processing applications on both GAE and Amazon EC2 shows that the computation costs were lower for GAE, owing mostly to the cycle-based payments as opposed to EC2's hourly billing intervals. The research concludes that Google App Engine offers relatively low resource-provisioning overhead and an inexpensive pricing model for jobs shorter than one hour. The authors are extending this work for running scientific workflows on Google App Engine.

Thorsten Schuett and Guillaume Pierre[9] the authors have developed frame work called ConPaaS for writing scalable cloud applications without considering the complexity of the cloud. It is the platform as a service (PaaS) component of the Contrail FP7 project. It provides a runtime environment that facilitates deployment of end-user applications in the Cloud .It provide services for web hosting (PHP and Java), SQL and NoSQL databases (MySQL and Scalaris), data storage (XtreemFS) and for large scale data processing (Task Farming and MapReduce). ConPaaS contains two services specifically dedicated to BigData: MapReduce and TaskFarming. MapReduce provides users with the well-known parallel programming paradigm. TaskFarming allows the automatic execution of a large collection of independent tasks such as those issued by Monte-Carlo simulations. The ability of these services to dynamically vary the number of Cloud resources they use makes it well-suited to very large computations: one only needs to scale services up before a big computation, and scale them down afterwards.

An experimental application called Wikipedia clone is implemented on ConPaaS. It can load database dumps of the official Wikipedia and store their content in the Scalarix NoSQL database service. The business logic is written in Java and runs in the Web hosting service. Deploying Wikipedia in the Cloud takes about 10 minutes.

Increasing the processing capacity of the application requires two mouse clicks. Following features can be extended ConPaas i). user specifying the performance aspects instead of choosing the number of resources. To allow users to upload complex applications in a single operation.ii).Instead of starting and configuring multiple ConPaaS services one by one, a user will be able to upload a single manifest file describing the entire application organization. iii). To provide an SDK for external users to implement their own services as plug-in in the existing ConPaaS system.

VI. CONCLUSION

According to IDC, the size of the *digital universe* was about 0.18 zettabyte in 2006 and it is forecasting a tenfold growth by the end of 2011 to 1.8 zettabyte (a zattabyte is one billion terabytes).It is difficult to store and process such a huge amount of data with traditional data processing frameworks. In this paper, we have explored different alternatives for deploying HPC applications. More focus was given for adhoc data intensive and parallel processing applications. Cloud computing because of its elasticity characteristic as the advantage over grid and cluster for adhoc parallel processing applications as the resources allocated/deallocated as per demand. Related work done in adapting IaaS or PaaS service for HPC applications are also presented.

In future work we will work on investigating the different resource allocation strategies on the performance of the cloud application. Our work will also focus on developing a cost effective dynamic resource allocation technique for running a application in cloud environment.

REFERENCES

[1] http://Gigaom.Com/Cloud/How-Nasa-Battle-Tested-Its-Mars- Rover-Live-Stream.

[2] Ivona Brandic, Schahram Dustdar "Grid Vs Cloud – A Technology Comparison." It – Information Technology 53(4): 173-179 (2011).

[3] Wen Zhang, Junwei Cao, Yisheng Zhong "Grid Resource Management and Scheduling For Data Streaming Applications." In GRID 2008: pp.258 265.

[4] http://Www.Nist.Gov/Itl/Csd/Cloud-102511.Cfm

[5] Daniel Warneke And Odej Kao "Exploiting Dynamic Resource Allocation For Efficient Parallel Data Processing In The Cloud." IEEE Transactions On Parallel And Distributed Systems, Vol 22 No 6,pp. 985 997, June 2011.

[6] Philip C.Church And Andrzej Goscinski "IaaS Clouds Vs Cluster For HPC Applications."In The Second International Conference On Cloud Computing, Grids And Virtualization 2011,pp.39-45.

[7] Bjorn Lohrmann,Daniel Warneke,Odej Kao "Massively Parallel Stream Processing Under QOS Constraints With Nephele". *HPDC*ACM(2012), pp. 271-282.

[8] Radu Prodan,Michael Sperk And Simon Ostermann "Evaluating High-Performance Computing On Google App Engine." IEEE Software March/April 2012 pp. 52-57.

[9] Thorsten Schuett And Guillaume Pierre "Conpaas, An Integrated Cloud Environment For Big Data " ERCIM News 2012(89).

[10] D.Hasenkamp,A.Sim,M.Wehner And K.Wu "Finding Tropical Cyclones On A Cloud Computing Cluster: Using Parallel Virtualization For Large-Scale Scale Climate Simulation Analysis." In Second IEEE International Conference On Cloud Computing Technology And Science.2010 pp.201-208.

[11] http://embedded-computing.com/articles/accelerating- using-parallel-computing.