International Journal of Engineering Inventions
ISSN: 2278-7461, www.ijeijournal.com
Volume 1, Issue 6 (October2012) PP: 60-63

Comparative Study on Parallel Data Processing for Resource Allocation in Cloud Computing

Abstract—Parallel data processing in cloud has emerged to be one killer application for infrastructure as service to integrate framework for products like portfolio, access these services and deploys the program. Scheduling job process in cloud computing for parallel data processing framework is Nephele. Our analysis presents expected performance of parallel job processing. Nephele is the processing framework to explicitly exploit the dynamic resource allocation IaaS cloud for task execution is assigned to different virtual machines which are automatically instantiated or terminated during the job execution is efficient in cloud never applied in existing systems and proposed comparison shows the task scheduling in resource allocation, parallel processing in cloud computing. In future extends our work by implementing novel parallel data processing advance to Nephele framework.

Keywords—Cloud Computing, Virtual Machine, Job Scheduling, Resources.

I. INTRODUCTION I

Cloud computing is a information technology extend their hands in order to improve their financial ability is done by improving the various quality of services parameters such as performance throughput reliability scalability load balancing. The services such as disk storage virtual servers applications design development testing environment are added benefit of cloud computing. Cloud computing technology makes the resource as a single point of access to the client and is implemented as pay usage and abstracted infrastructures completely virtualized environment quipped with dynamic free of software and hardware installations. Growing number of companies have to process huge amounts of data in a cost efficient manner operators such as internet search engines like Google Yahoo or Microsofts.

The development of distributed applications on top such architecture many of these companies have also built customized data processing framework Google Map Reduce Merge can be classified by terms like high throughput computing or many task computing depending on the amount of data and the number of tasks involved in the computation. Although these systems differ in design their programming models share similar objectives namely hiding the hassle of parallel programming fault tolerance and execution optimizations form the developer. The processing framework then takes care of distributing the program among the available nodes and executes each instance of the program on the appropriate fragment of data. To process large amounts of data occasionally running their own data center is obviously not an option. Cloud computing has emerged as a promising approach to a large IT infrastructure on a short term pay per usage basis which includes Amazon EC2 customer allocate access and control a set of virtual machines which run inside their data centers and only charge them for the period of time the machines are allocated.The virtual machine abstraction of clouds fits the architecture assumed by the data processing a popular open source implementation of Google Map Reduces framework already have begin to promote using their framework.

II. SECTION

2. Relatedwork:

MapReduce is a programming framework that supports the model is hide details of parallel execution and allow users to focus only on data processing and consists of two primitives functions: Map and Reduce input for MapReduce is a list of (key1, value1) pairs and Map() is applied to each pair to compute intermediate key value a pairs are grouped together on the key equality basis (key2,list(value2))). For each key2 Reduce() works on the list of all values then produce zero or more aggregated results Map Reduce utilizes. Current data processing frameworks like Google Map Reduce designed for cluster environment. Today's processing framework typically assume the sources mange consist of a static set of homogeneous compute nodes. One of an IaaS cloud key features is provisioning of compute resources on demand allocated at any time through a well-defined interface available in a seconds. Moreover cloud operators like Amazon let their customers rent VMs of different types with different computational power in sizes of main memory and storage hence the compute resource available in a cloud are dynamic and heterogeneous. Parallel processing is flexible leads to a variety of new possibilities for scheduling data processing jobs.

The clouds virtualized nature to enable use-case for efficient parallel data processing it is imposes new challenges compared to classic cluster setups.In a cluster the compute nodes are typically interconnected through a physical high-performance network. The topology of the network, i.e. the way the compute nodes are physically wired to each other, is usually well-known and, what is more important, does not change over time. Current data processing frameworks offer to leverage this knowledge about the network hierarchy and attempt to schedule tasks on compute nodes so that data sent from one node to the other has to traverse as few network switches as possible [9]. That way network bottlenecks can be avoided and the overall throughput of the cluster can be improved.

It was possible to determine network hierarchy in a cloud and use it for topology aware scheduling the obtained information will not necessarily remain valid for the entire processing time. VMs may be migrated for administrative purposes between different locations inside the data centre without any notification, rendering any previous knowledge of the relevant network infrastructure obsolete. As a result, the only way to ensure locality between tasks of a processing job is currently to execute these tasks on the same VM in the cloud. This may involve allocating fewer, but more powerful VMs with multiple CPU cores. E.g., consider an aggregation task receiving data from seven generator tasks. Data locality can be ensured by scheduling these tasks to run on a VM with eight cores instead of eight distinct single-core machines, no data processing framework includes such strategies in scheduling algorithms.

III. SECTION

3.1. Problem definition:

To integrate the cloud computing task such as portfolio access these services and to deploy their programs for efficient parallel processing. Each vertex in the graph represents process flow edges define the communication between these tasks aslo decided to use Directed Acyclic Graph is revelant to Nephele. The user must write the program code for external task must be assigned to a vertex and must be connected by edges to define the communication paths of the job.

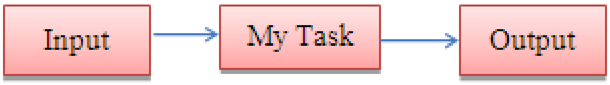

Figure 1 shows the problem domain

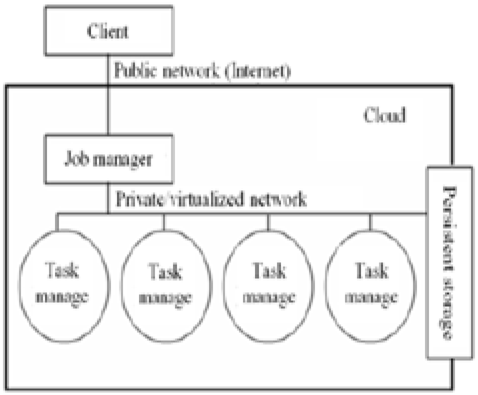

Figure 2 shows the Proposed Architecture

To avoid our problem scheduling Nephele task is considerable interest has focused on preventing identity inference in location based services proposes spatial cloaking techniques which describe existing techniques for query processing at the end alternative location privacy. Processing is based on theorem uses search operations thus the NAP query evaluation methodology is readily deployable on existing systems and can be easily adapted to different network storage schemes. Network-based anonymization and processing framework, the first system for K- anonymous query processing in road networks. NAP relies on a global user ordering and bucketization that satisfies reciprocity and guarantees K-anonymity identifiesthe ordering characteristics that affect subsequent processing, and qualitatively compare alternatives.

3.2. Parallel strategies:

Constructing an execution graph from a user submitted job graph may leave different degrees to Nephele to most efficient execution graph. Unless the user provides any job annotation which contains more specific instructions,a simple strategy each vertex of the task graph is transformed into one execution vertex the default channel types assigned to its own execution instance unless the user annotations or other scheduling such as memory channels is available in IaaS cloud. One idea is to refine the scheduling for recurring jobs is to use backend data continuously running tasks and the underlying instance based on java capable of breaking down what percentage of processing time a task thread actually spends processing user code

3.3. Scheduling a Job:

once received a valid job graph from the user Nephele manager transforms it into a execution graph, it is primary data structure for scheduling and monitoring the execution of a Nephele job. Unlike the job graph, execution contains all the concern information required to schedule and executes the received job on the cloud explicitly. A new data processing for cloud takes many ideas of existing frameworks but refines them to match the dynamic nature of a cloud architecture follows a master work pattern. Before submitting a Nephele compute a job user must with the virtual machine in the cloud which runs the job manager. The Job Manager receives the client’s job is responsible to scheduling the task and coordinates their execution. It is capable of communicating with the interface. The cloud operator provides to control the instantiation of svirtual machine interface cloud controller. Controller is the Job Manager can be allocate or de-allocate virtual machine according to the current job execution phase. These comply with common Cloud computing terminology and refer to these VMs as instances. The term instance type will be used to differentiate between VMs with different hardware characteristics.

The actual execution of tasks which a Nephele job consists of is carried out by a set of instances. Each instance runs called Task Manager (TM).

IV. SECTION

4. Performance Analysis:

The first performanceof Nephele to the data processing framework have chosen Hadoopas competitor, because it is an open source softwareand currently enjoys high popularity in the dataprocessing community. Hadoop hasbeen designed to run on a very large number of nodes(i.e. several thousand nodes). According to observations, the software is typically used with significantlyfewer instances in current IaaS clouds. In fact,Amazon itself limits the number of available instancesfor their Map Reduce service to 20 unless the respectivecustomer passes an extended registration process [2]. The second task subsequently is to calculatethe average of these k smallest numbers. The job isa classic representative for a variety of data analysisjobs whose particular tasks vary in their complexityand hardware demands. While the first task has to sortthe entire data set and therefore large amounts of main memory and parallel execution,the second aggregation task requires almost no mainmemory and, at least eventually, cannot be parallelized. For the first experiment,implemented the task as a sequence of MapReduceprograms and executed it using Hadoop on a fixed setof instances. For the second experiment, reused thesame Map Reduce programs as in the first experimentbut devised a special Map Reduce wrapper to makethese programs run on top of Nephele. The goal ofthis experiment was to illustrate the benefits of dynamicresource allocation/de-allocation while still maintainingthe Map Reduce processing pattern. Finally, as the thirdexperiment, we discarded the Map Reduce pattern andimplemented the task based on a DAG to also highlightthe advantages of using heterogeneous instances.For all three experiments, we chose the data set size tobe 100 GB. Each integer number had the size of 100 bytes.As a result, the data set contained about 109 distinctinteger numbers. The cut-off variable k has been set to2 _ 108, so the smallest 20 % of all numbers had to bedetermined and aggregated.

V. SECTION V

5. Comparative Study:

Cloud computing can build by following deployment methods. Public cloud is cloud computing in mainstream sense where the resources are dynamically provisioned to the general on pay-per-use model over the internet via web-based applications. Hybrid cloud is composition of more clouds that remain unique entities but are bound together the benefits of multiple methods. Private cloud is infrastructure built by a single organization which monitors internally or third-party behind a firewall. Resource allocation of cloud in static allocation is mostly for homogenous compute nodes and dynamic resource allocation either they consider private cloud or hybrid cloud never provide priority based resource allocation to re-evaluate the tasks that are in job scheduling. Task scheduling is a activity uses a set of inputs to produce a set of outputs process in fixed set are statically assigned to processors at runtime or compile time to avoid overhead of balancing task. In cloud computing, each application of users willrun on a virtual operating systems, the cloud systemsdistributed resources among these virtual systems. Everyapplication is completely different and is independentand has no link between each other whatsoever, Forexample, some require more CPU time to computecomplex task and some others may need more memory tostore data. Resources are sacrificed on activities performedon each individual unit of service.

VI. CONCLUSION

In this paper we have discussed the challenges andopportunities forefficient Nepheleparallel data processing in cloud, first dataprocessing framework to exploit the dynamic resourceprovisioning offered by today's IaaS clouds which describes the Nephele basic architecture and presented aperformance and comparison to the well-established dataprocessing framework Hadoop. Performance analysis gives impression on how the ability to assign specific virtual machine types to specific tasks of processing task as well automatically allocate or de-allocate virtual machine toimprove the overall resource utilization and, consequently,reduce the processing cost. With a framework like Nepheleat hand, there are a variety of open research issues, whichwe plan to address for future work. In particular, we areinterested in improving Nephele ability to adapt toresource overload or underutilization during the jobexecution automatically. Our current profiling approachbuilds a valuable basis for this, however, at the moment thesystem still requires a reasonable amount of userannotations.

REFERENCES

1. Amazon Web Services LLC. Amazon Elastic Compute Cloud(Amazon EC2). http://aws.amazon.com/ec2/, 2009.

2. Amazon Web Services LLC. Amazon Elastic MapReduce.http://aws.amazon.com/elasticmapreduce/, 2009.

3. AmazonWeb Services LLC. Amazon Simple Storage Service.http://aws.amazon.com/s3/, 2009.

4. D. Battr´e, S. Ewen, F. Hueske, O. Kao, V. Markl, and D. Warneke.Nephele/PACTs: A Programming Model and Execution Frameworkfor Web-Scale Analytical Processing. In SoCC '10: Proceedingsof the ACM Symposium on Cloud Computing 2010, pages 119–130, New York, NY, USA, 2010. ACM.

5. Daniel Warneke, Member, IEEE, and Odej Kao, Member,IEEE," Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 22, NO. 6, JUNE 2011.

6. R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey, D. Shakib, S.Weaver, and J. Zhou, "SCOPE: Easy and Efficient Parallel Processing of Massive Data Sets," Proc. Very Large Database Endowment, vol. 1, no. 2, pp. 1265-1276, 2008.

7. Saurabh Singh, GauravAgarwal, Invertis University Bareilly, India," Integration of Sound Signature in Graphical Password Authentication System" International Journal of Computer Applications (0975 – 8887)ss Volume 12– No.9, January 2011.

8. Amazon Web Services LLC, “Amazon Elastic Compute Cloud (Amazon EC2),” htt://aws.amazon.com/ec2/, 2009.

9. Dornemann, T., E. Juhnke and B. Freisleben, 2009. On-demand resource provisioning for BPEL workflows using amazon's elastic compute cloud. Proceedings of the 9th IEEE/ACM International Symposium Cluster Computing and the Grid, May 18-21, IEEE Xplore Press, Shanghai, pp: 140-147.

10. R. Chaiken, B. Jenkins, P.-A.Larson, B. Ramsey, D. Shakib,S. Weaver, and J. Zhou. SCOPE: Easy and Efficient ParallelProcessing of Massive Data Sets. Proc. VLDB Endow., 1(2):1265–1276, 2008.

11. H. chih Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker. Map-Reduce-Merge: Simplified Relational Data Processing on LargeClusters. In SIGMOD ’07: Proceedings of the 2007 ACM SIGMODinternational conference on Management of data, pages 1029–1040,New York, NY, USA, 2007. ACM.

12. M. Coates, R. Castro, R. Nowak, M. Gadhiok, R. King, andY. Tsang. Maximum Likelihood Network Topology Identificationfrom Edge-Based Unicast Measurements. SIGMETRICS Perform.Eval. Rev., 30(1):11–20, 2002.

13. R. Davoli. VDE: Virtual Distributed Ethernet. Testbeds and ResearchInfrastructures for the Development of Networks &Communities,International Conference on, 0:213–220, 2005.

K.C. Ravi Kumar M.Tech inComputer Science and Engi -neering from JNTU Hderabad ,Currently he has been working as Assoc.profesor in CSE department in SriDevi Women’s Engg College, having 17 years of- Academic Experience. His areas of research include Data Warehousing and Data mining ,Information Retrival Systems, Information Security.

B.Viplava Reddy M.Tech in Computer Science and Engineering in Sridevi Women’s Engineering College B.Tech Information Technology from Jaya PrakashNarayana College of Engineering. Her research areas t include Data Warehousing and Data mining.