International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Framework for Parallel Data Processing in Cloud Systems

Shripad S. Lokhande[1], Manoj Limchand Bangare[2]

[1]ME-IT, Department of Information Technology, Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

[2]Assistant Professor, Department of I.T. Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Abstract: ***In recent years ad hoc parallel data processing has emerged to be one of the effective applications for Infrastructure-as-a-Service (IaaS) clouds. Major Cloud computing companies have started to integrate frameworks for parallel data processing in their product portfolio, making it easy for customers to access these services and to deploy their programs. However, the processing frameworks which are currently used have been designed for static, homogeneous cluster setups and disregard the particular nature of a cloud. Consequently, the allocated compute resources may be inadequate for big parts of the submitted job and unnecessarily increase processing time and cost. In this project, we discuss the opportunities and challenges for efficient parallel data processing in clouds and present our research project Nephele. Nephele is the first data processing framework to explicitly exploit the dynamic resource allocation offered by today's IaaS clouds for both, task scheduling and execution. Particular tasks of a processing job can be assigned to different types of virtual machines which are automatically instantiated and terminated during the job execution.***

Keywords: Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud

1. Introduction

A Model which has been Computing is being transformed with the services that are produce and delivered in a manner like the traditional utilities such as gas, water, and Electricity and so on.. In similar way cloud models, users can access services based on their needs without any regard where the services hosted or how they delivered and satisfy the user requirements.

At Now, it is frequently to access the content across the Internet independently without reference to any hosting infrastructure underlying. Infrastructure consists of large data centres which are maintained and monitored around the clock by information providers. Cloud computing is recent technology of this paradigm. In the potential of business or enterprise applications which are revealed as highly developed advanced services which access to a public network. Service providers are collaborative consumption by the earnings to be made by charging consumers for accessing these services. Consumers are attracted by the opportunity for reducing prices. However, cloud applications may be a vital and crucial impact for the core level business operations of the user it is important and essential factor that the user have guarantees from service providers on the service delivery.

As with user requirements cloud service providers has to ensure that service delivery should be in flexible and efficient while keeping the users isolated from the underlying infrastructure. In software and microprocessor technology many drastic changes which led to the increasing ability of commodity hardware to run applications within Virtual Machines efficiently. Virtual Machines allow both the isolation of applications from the underlying the situations of hardware and other Virtual Machines, and the customization of the platform to suit the needs of the end-user or the clients.

2. Problem Statement

A server which deploys many applications and shares their resources which leads the clients to initiate the communication and share resources from server, decreasing the performance when clients are increased.

3. System Analysis

In this project the Sun java language is used for developing the piracy protection software, java is the product from sun Microsystems and it provided for free of cost. So, it is bought without spending money. It can download from the sun website directly from the Internet itself. So, it is not needed to look from any third party or in the market. It is freely downloaded it from the Internet. This Exact Knowledge Hiding through Database Extension Software will be cost effective.

- **Web service**
  Web service is a way of communication over a network between two devices. These were intended to solve three main problems such as Complexity, and Interoperability.
- **Database**
  Database represents the client's database which consists of the details about the transactions on the client's side.
- **JDBC**
  JDBC is a Java-based data access layer technology (Java Standard Edition platform) from Sun Microsystems. JDBC is an acronym as it is unofficially referred to as Java Database Connectivity, with DB being universally recognized as an abbreviation for database. An ODBC-to-JDBC bridge enables connections to any ODBC-accessible data source in the JVM host environment. A JDBC driver is a software component enabling a Java application to interact with a database. The JDBC driver gives out the connection to the database and implements the protocol for transferring the query and result between client and database.

- **XML**
  Extended Markup Language is used to set the rules for exchange of information in the proposed model. It is a Markup language that is used to define set of rules for encoding documents in a format that is both understood by human and as well as machines. The design goals of Extended Markup Language include simplicity, generality, and usability over the Internet. It can also be defined as textual data format with strong support through Unicode for the languages of the world. Even though the design of Extended Markup Language focuses on documents, it is use for the representation of arbitrary data structures. The proposed model can be used to provide any web service, but in this thesis, I am concerned about only Amazon Web services. The web service provides security web services and as well as storage system.

4. Scope of Project

Our System is Useful for Cloud user and cloud service provider.

5. Problem with Existing System

The information pertained to Web access and inflation in order to alter the request adapted to content access patterns are introduced here. Even then the key role of Web services in information infrastructure services will not be altered. The expansion of the scope of Web services using large-scale resources such as cloud computing will become more important in the future, thus technological enhancement and research oriented work are focused on the same . We plan to continue this research as part of larger research goals such as Cloud computing is generating enormous amounts of discussion and excitement in the world of corporate IT.

6. Related Work

For cloud computing the related author work is as given below

- D. Battr´e, S. Ewen, F. Hueske, O. Kao, V. Mark, and D. Warneke, *"A Programming Model and Execution Framework for Web-Scale Analytical Processing"*, ACM 2010.[1]This Study proposes business model for cloud computing based on the concept of presented a generic system for web scale data processing. Tasks executed in a flexible that can able to execute arbitrary acyclic data flows.
- Daniel Warneke and Odej Kao " *Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud*," in IEEE Transactions On Parallel And Distributed Systems, VOL. 22, NO. 6, JUNE 2011.[2] The author describes the parallel data processing with work scheduling, resource management and communication.
- D.H. chi Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker, *" Map Reduce Merge: Simplified Relational Data Processing on Large Clusters"*, international conference on Management of data, vol. 1, pp. 1029-1040, October 2007.[3] In this paper, the author studied Map-Reduce is a programming model that enables easy development of scalable parallel applications to process vast amounts of data on large clusters of commodity machines
- M. Coates, R. Castro, R. Nowak, M. Gadhiok, R. King, and Y. Tsang, *"Maximum Likelihood Network Topology Identification from Edge-Based Unicast Measurements,"* SIGMETRICS, vol.30, pp. 11-20, 2002.[4] The author studied the problem of discovering network topology solely from host-based, unicast measurements, without internal network cooperation.
- R. Davoli. VDE: Virtual Distributed Ethernet., *"Testbeds and Research Infrastructures for the Development of Networks & Communities, International Conference*, pp. 231-220, September 2005.[5] In this paper author discusses on service architecture that merges the paradigms and technologies of the Internet with the cellular and fixed telecommunication worlds
- J. Dean and S. Ghemawat., *"MapReduce"* Simplified Data Processing on Large Clusters, vol. 6, no. 1, Springer, pp. 01-10. 2005[6] in this paper the author describe the Map Reduce is a programming model and associated implementation for processing and generating large data sets.
- E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, and D. S. Katz, *"A Framework for Mapping Complex Scientific Workflows onto Distributed Systems,"* Future Generations Computer Systems, vol. 1, issue 6, pp. 219-237, 2005.[7] in this paper the concept of cloud as well as how to use workflow framework for distributed system.
- T. Dornemann, E. Juhnke, and B. Freisleben, *"On-Demand ResourceProvisioning for BPEL Workflows Using Amazon's Elastic Compute Cloud,"* ACM International Symposium on Cluster Computing and the Grid, vol. 3, no. 1, pp. 140-147, 2009. [8] this paper author discussing the concept of On-Demand Resource Provisioning for BPEL Workflows Using Amazon's Elastic Cloud Computing.
- M.Malathi, *"Cloud Computing Concepts"*, in IEEE conference,2011.[09]In this paper the concepts of Cloud Computing like cloud deployment models, delivery models, advantages of using cloud and risks involved in it analysed. This paper tries to bring awareness among managers and computing professionals to use Cloud computing as an alternative to large in-house data centers.
- YashpalsinhJadeja, KiritModi, *"Cloud Computing - Concepts, Architecture and Challenges"*, in International Conference on Computing, Electronics and Electrical Technologies [ICCEET],2012.[10] In this paper the author focusing on architecture of cloud computing, refining of the concept of cloud computing as well as challenges. In this paper discuss what makes all this possible, what is the architectural design of cloud computing and its applications.
- ShyamPatidar, DheerajRane, Pritesh Jain, *"A Survey Paper on Cloud Computing"*, in Second International Conference on Advanced Computing & Communication Technologies,2012.[11] In this paper, described the definition, styles, characters of cloud computing and cloud computing services. Though each cloud computing platform has its own strength, one thing should be noticed is that no matter what kind of platform there is lots unsolved issues. For example, continuously high availability, Performance, Data Confidentiality and Audit

ability, Synchronization in different clusters in cloud platform, interoperation and standardization, the security of cloud platform.

- ZHAN Ying, *"Research on Management of Data Flow in the Cloud Storage Node Based on Data Block"*, Third International Conference on Information and Computing, 2010.[12]. The proposed work is Storage management control optimized is effective method that will reduce the working time to large scale data storage management.
- R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey, D. Shakib S.Weaver, and J. Zhou, *"Easy and Efficient Parallel Processing of Massive Data Sets,"* VLDB Endow, vol. 1, no. 2, pp. 1265-1276, 2008.[14] The author describes the development of cost-efficient distributed storage solutions data and the development of distributed computing frameworks processing.

7. Frame Work

In this project, we develop the design and implementation of an automated resource management system for parallel data processing. Data processing regenerating a web index are split into several independent tasks, distributed among the available nodes, and computed in parallel. The processing data takes care of distributing the program among the available system and executes each instance of the program on the appropriate data.

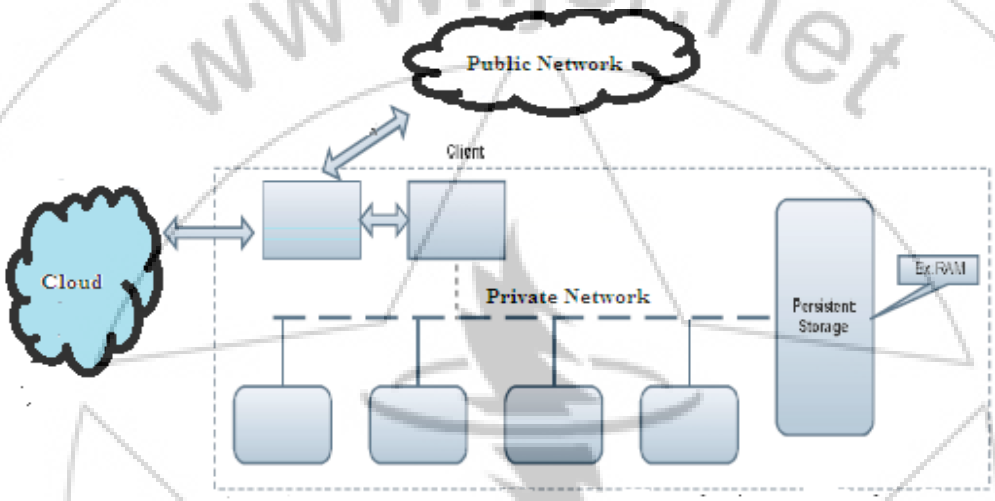

Figure 1: Structural overview of running in an Infrastructure-as-a-Service (IaaS) in cloud

8. Result

Performance results of our three experiment, respectively. All three plots illustrate the average instance utilization over time, i.e., the average utilization of all CPU cores in all instances allocated for the job at the given point in time. The utilization of each instance has been monitored with the Unix command "top" and is broken down into the amount of time the CPU cores spent running the respective data processing framework , the kernel and its processes , and the time waiting for I/O to complete (WAIT). In order to illustrate the impact of network communication, the plots additionally show the average amount of IP traffic flowing between the instances over time.

9. Conclusion

We have presented the design, implementation of a resource management system for cloud computing services. System multiplexes virtual to physical resources adaptively based on the changing demand. We use the combine VMs with different resource characteristics appropriately so that the capacities of servers are well utilized. Our project achieves both overload avoidance and clients can connect to virtual machines of the cloud and they can perform their business tasks in parallel by dynamic resource allocation.

10. Future Enhancement

Future work will be focused on Reducing Cost of data processing in the cloud. In particular, we are interested in improving Nephele's ability to adapt to resource over load or underutilization during the job execution automatically. IT technicians are spearheading the challenge. Several groups have recently been formed, such as the Open Cloud Consortium, and to establish a common language among different providers. In general, we think our work represents an important contribution to the growing field of Cloud computing services and points out exciting new opportunities in the field of parallel data processing.

References

[1] Daniel Warneke , Odej Kao Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud , Einsteinufer 17,10587 Berlin, Germany,2011.
[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 Framework for Web-Scale Analytical Processing. In SoCC '10: Proceedings of the ACM

Symposium on Cloud Computing 2010, pages 119–130, New York, NY, USA, 2010. ACM.

[5] 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. VLDB Endow., 1(2):1265–1276, 2008.

[6] H. chih Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker. Map-Reduce-Merge: Simplified Relational Data Processing on Large Clusters. In SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pages 1029–1040, New York, NY, USA, 2007. ACM.

[7] M. Coates, R. Castro, R. Nowak, M. Gadhiok, R. King, and Y. Tsang. Maximum Likelihood Network Topology Identification from Edge-Based Unicast Measurements. SIGMETRICS Perform. Eval. Rev., 30(1):11–20, 2002.

[8] R. Davoli. VDE: Virtual Distributed Ethernet. Testbeds and Research Infrastructures for the Development of Networks & Communities, International Conference on, 0:213–220, 2005.

[9] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association.

[10] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, and D. S. Katz. Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems. Sci. Program., 13(3):219–237, 2005.

[11] T. Dornemann, E. Juhnke, and B. Freisleben. On-Demand Resource Provisioning for BPEL Workflows Using Amazon's Elastic Compute Cloud. In CCGRID '09: Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pages 140–147, Washington, DC, USA, 2009. IEEE Computer Society.

[12] I. Foster and C. Kesselman. Globus: A Metacomputing Infrastructure Toolkit. Intl. Journal of Supercomputer Applications, 11(2):115–128, 1997.

[13] J. Frey, T. Tannenbaum, M. Livny, I. Foster, and S. Tuecke. Condor- G: A Computation Management Agent for Multi-Institutional Grids. Cluster Computing, 5(3):237–246, 2002.

[14] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: Distributed Data-Parallel Programs from Sequential Building Blocks. In EuroSys '07: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007, pages 59–72, New York, NY, USA, 2007. ACM.

[15] A. Kivity. kvm: the Linux Virtual Machine Monitor. In OLS '07: The 2007 Ottawa Linux Symposium, pages 225–230, July 2007.

[16] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov. Eucalyptus: A Technical Report on an Elastic Utility Computing Architecture Linking Your Programs to Useful Systems. Technical report, University of California, Santa Barbara, 2008.

[17] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig Latin: A Not-So-Foreign Language for Data Processing. In SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1099–1110, New York, NY, USA, 2008. ACM.

[18] O. O'Malley and A. C. Murthy. Winning a 60 Second Dash with a Yellow Elephant. Technical report, Yahoo!, 2009.

[19] R. Pike, S. Dorward, R. Griesemer, and S. Quinlan. Interpreting the Data: Parallel Analysis with Sawzall. Sci. Program., 13(4):277–298, 2005.

[20] I. Raicu, I. Foster, and Y. Zhao. Many-Task Computing for Grids and Supercomputers. In Many-Task Computing on Grids and Supercomputers, 2008. MTAGS 2008. Workshop on, pages 1–11, Nov.2008.

[21] I. Raicu, Y. Zhao, C. Dumitrescu, I. Foster, and M. Wilde. Falkon: a Fast and Light-weight tasK executiON framework. In SC '07: Proceedings of the 2007 ACM/IEEE conference on Supercomputing, pages 1–12, New York, NY, USA, 2007. ACM.

[22] L. Ramakrishnan, C. Koelbel, Y.-S. Kee, R. Wolski, D. Nurmi, D. Gannon, G. Obertelli, A. YarKhan, A. Mandal, T. M. Huang, K. Thyagaraja, and D. Zagorodnov. VGrADS: Enabling e-Science Workflows on Grids and Clouds with Fault Tolerance. In SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, pages 1–12, New York, NY, USA, 2009. ACM.

[23] R. Russell. virtio: Towards a De-Facto Standard for Virtual I/O Devices. SIGOPS Oper. Syst. Rev., 42(5):95–103, 2008.