International Journal of Pure and Applied Mathematics
Volume 119 No. 14 2018, 1341-1346
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

TRANSPORT-SUPPORT WORKFLOW COMPOSITION AND OPTIMIZATION FOR BIG DATA WITH VIRTUALIZATION

N. Balaji[1], K. Maheswari[2], S. Sumitha[3], S. Sujitha[4] and S. Gurulingam[5]
[1]Associate Professor, [2, 3, 4]Assistant Professor, [5] Principal, Department of CSE, SVCE & Technology, Pondicherry University, India.

ABSTRACT: Big data is a large amount of information collection, which surpasses the capabilities of the traditional algorithm and the technologies to produce useful value. The real power is not just a massive data but it is the innate that you can learn from this data to make good and fast decisions. The load imbalance that occurs when processing large data is a common problem, with this problem processing QOS can affect the application. A new model classification algorithm in this paper has been proposed to address these issues. The proposed technique will tackle this situation by virtualization. The client can change the change the resource unit depending upon the performance.

Keywords: Big data, Scheduling Algorithms, Virtualization, Virtual Machine.

INTRODUCTION

Today, the system and the people uses a remarkable generation web of large size of data.[1] The size of the data on the web is measured in Exabyte and Petabytes. By 2025, the Internet will surpass everyone's brain sizeLive in the whole world. This company's growth of data is due to the development of large amounts of digital sensors, calculations, communications and reservesData meetings.

Gartner's company claims that information or data will be 21st century oil. In the last 25 years, data has grown in various fields in different sectors. According to the International Data Corporation (IDC) statistics report, in 2011, the amount of data created in the world was 1.8 fever. It has expanded nearly nine times over five years. Now with the inclusion of marketing, smart city, disease control and prevention and business intelligence applications, it is important to note that large data can play a significant role everywhere in the universe.

[2] Social networking sites that collect extensive data are resources for grinding from the facebook, Twitter, where the billions of Ayurveda are posted daily data. Share Markets are large data with the algorithmic stack processAlso by collecting data through exchange and transaction. E-commerce sites can be used for service providers. Other goods and resources of large data are the items that are satisfying the fuel pockets and telecom companies and so on. The three parameter features of large data are Velocity, Volume Accuracy. The data is growing in size and the data doubles every two years in a veritable way. Big data is sometimes structured data and SometimeMusic Redistricted data. Structure or data that is organized in rows and columns, and thenData is called structured data.

I. RELATED WORKS

Big data Computation [16-20] new to develop and need more researchto make it more reliable and to have a better user experience interm of task computation.[3] Computational science sometimes requires large data files to be transferred over high bandwidth-delay-product (BDP) wide-area networks (WANs). Experimental data (e.g., LHC, SKA), analytics logs, and file system backups are regularly transferred between research centres and between private-public clouds. Fortunately, a variety of tools (e.g., Grid FTP, UDT, PDS) have been developed to transfer bulk data across WANs with high performance. However, using large-data transfer tools could adversely affect other network applications on shared networks. Many of the tools explicitly ignore TCP fairness to achieve high performance. Users have experienced high-latency and low-bandwidth situations when a large-data transfer is underway. But there have been few empirical studies that quantify the impact of the tools. As an extension of our previous work using synthetic background traffic, itperform an empirical analysis of how the bulk-data transfer tools perform when competing with a non-synthetic, application-based workload (e.g., Network File System). That characterized and showed that.

[4] That document documents the consensus reached by the Multiprotocol Label Switching (MPLS) Working Group within the IETF to focus its efforts on "Resource Reservation Protocol (RSVP)-TE: Extensions to RSVP for Label-Switched Paths (LSP) Tunnels" (RFC 3209)

as the MPLS signalling protocol for traffic engineering applications and to undertake no new efforts relating to "Constraint-Based LSP Setup using Label distribution Protocol (LDP)" (RFC 3212). The recommendations of section 6 have been accepted by the IESG.

[5] Distributed computing applications maintain more specialized protocols, demanding greater control over the communication surface. Here, high-level protocols, such as Atomic Multicast and Group RPC, are described in the network subsystem that supports modular, fine-grained construction. Micro-protocol objects depending on the extension of the x-kernel's standard hierarchical model with the protocol composed in a standard runtime framework. Each micro-protocol identifies a particular semantic property that leads to a highly modular and configurable implementation. In contrast to such systems, this approach provides a subtle granularity and more convenient inter-communications communication. The model and model implementation in Mac is described. Performance Outcomes the group for micro-protocol suite involves implementing variations of the RPC.These papers contribute the performance of the networks and our proposed system will change the performance of the server so that both server and the network performance has been optimized.

II. PROPOSED ARCHITECTURE

The process of scheduling the tasks into the cluster resources in a manner that minimizes task completion time and resource utilization is known as Task Scheduling. The main functional requirements of task scheduling are scalability, dynamism, time and cost efficiency, handling different types of processing models, data and jobs, and more. The other major objectives of task scheduling are reducing the number of task migrations and allocating the number of dependent and independent tasks in a near optimal manner which decreases the overall computation time of a job and improves the utilization of cluster resources effectively. In addition to that, the big data platforms construct the task graph per job which can be generated either in a static or dynamic mode. Hence, the task scheduling should have the ability to handle and schedule both static and dynamic task graphs. In summary, the main contributions of this research paper are:

1. Investigated various static and dynamic task scheduling algorithms and task scheduling considerations for processing of batch and streaming jobs.
2. Explored the task scheduling algorithms available in the popular open-source big data platforms such as Hadoop, Mesos, Spark, Flink, Heron and Storm for batch and stream job processing.
3. Proposed a task scheduling model/system which considers both static and dynamic task graphs and provides the ability to schedule batch, streaming,MPI, and micro services job types.

In static task scheduling,[6] the jobs are allocated to the nodes before the execution of a job and the processing nodes are known at compile time. Once the tasks are assigned to the appropriate resources, the execution continues to run until task completion. The main objective of the static task scheduling strategy is to reduce the scheduling overhead that occurs during the runtime and minimize the number of nodes/processors.

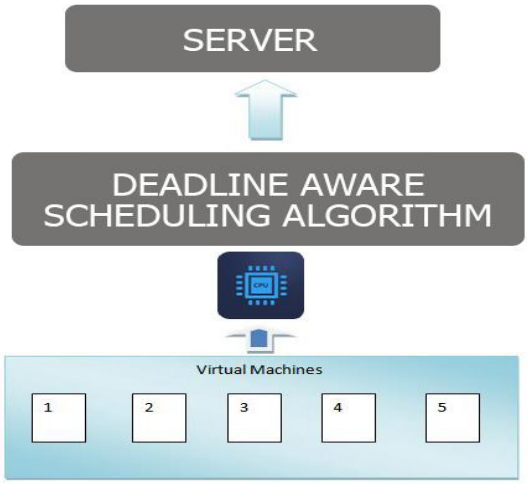

Fig.1 Proposed Architecture

Deadline Aware Scheduling Algorithm

1. Find fittest VM (Reqtype)
2. {
3. Compute utilization for each VM;
4. Compute load for each VM;
5. Compute Ti for each VM;
6. Sort descending;
7. Find VM with high trust value;
8. Return VM ;
9. }

Ti = 1* Initiali + 3*(1/Faulti)+4*(1/Utilization i)+ 5*(1/loadi).

Initiali : initialization time of VM.

Faulti : fault rate over VM.

Utilization i : utilization oh VM at current point of time.

Loadi : current load of VM

However, the major disadvantages of the static task scheduling algorithms are that they don't consider the workload of the resources and resource requirements of an application that obviously leads to overutilization or underutilization of the resources which may pave the way for job execution failure. Capacity Scheduling, Data Locality-Aware Scheduling, Round Robin Scheduling, Delay Scheduling, FIFO Scheduling, First Fit Scheduling, Fair Scheduling, Matchmaking Scheduling, and so on are some of the examples of static task scheduling algorithms. It is impossible to discuss all the static task scheduling related research papers within this paper. Hence, we briefly discussed some of the closely related static task scheduling works.

In Big data, batch processing is an efficient way of processing a large volume of data collected and stored over a period of time whereas streaming refers to the processing of real-time data in an interactive manner. Batch and streaming processing have their own advantages and disadvantages. It is important to decide the processing system based on the nature of job type, the source of input data, and processing time. The Big data stream should continue to process the data streams of online data. The Big data batch processing requires high performance computing cycles whereas the Big data stream processing requires low latency for efficient processing.

This document uses different parameters to estimate the credible value for the host, the trial and the algorithmic algorithm proposed to increase the use of resources available on the trust value and VM allocation policy data center. We can proceed to extract information for each data center trust. The trust can be defined as an indirect reliability or a firm belief in a host based on its past performance parameters. Reliability depends: Startup Time: Enabling a virtual machine (VM) to the time taken by the host. Processing speed: Mips total number of a machine means processor number * MIPS number in each processor

III. MODULES

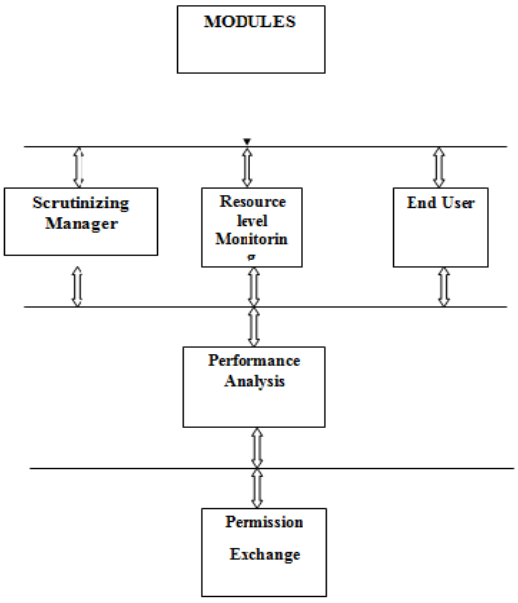

Fig.2. Modules

Fig.2. the modules are the main functional block. The Permission exchange will be done by these modules.

4.1 Scrutinizing Manager

The virtual machine incorporates SM to administer security. SM acts as a controller and scrutinizing agent to identify and mitigate adversary effects with lesser complexity. In order to prevent resource exhaustion due to intensive VM monitoring, SM relays information of the VM that are requested by the hypervisors. To protect user resources, two-level monitoring is initiated by the SM. The two-level monitoring overrules when monitoring is more vital; with identical events over different time intervals. SM possesses two-levels of monitoring that are briefed as follows:

4.2 Resource level Monitoring

SM verifies if any false data or irrelevant/compromising-kind of data is being injected to the resource after utilization. If the resource is found to be tampered then SM instructs the resource manager to liberate the resource such that a new resource is further allocated by the resource manager. Resource level monitoring improve efficient management and optimization at the time of allocation and reallocation. The main theme of RM is permission exchanges lower

level of RU to the users based on the cpu speed by Deadline aware scheduling algorithm.

4.3 End user

The end user first register to virtual machine for security purpose each of the user allocates specific secret key to the user by using of secret key end user must login the virtual machine. The task is allocated based on the cpu speed of the virtual machine.

IV. PERFOMANCE ANALYSIS

We have simulated proposed algorithm in NetBeans IDE as an Web Page. Proposedalgorithm is being tested over with Virtual Machine RU1,RU2, RU3,RU4, RU5. Trust value will bedepend on utilization of VM.

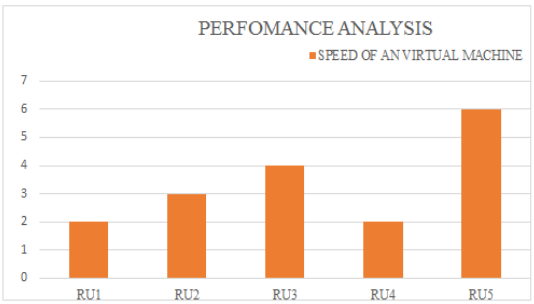

Fig.3. Performance analysis of VM

Fig.3. shows that the trust value of the RU5 is maximum, so that the speed of the machine is comparatively high. Hence, it is best for transferring large amount of data.If VM's reputation is higher, they will result in more destiny that VM. Many factors to count the reputation also reflect QoS of cloud computing. This paper also offered a way to provide reliability of a request, as well as providing reliable handling with the reputation of the VM factor that leads to reliability. The trust is scheduled according to a mathematical equation and schedule.

V. CONCLUSION

In this document, the reliable value of a workflow for work scheduling, is received from users as well as from the service provider. Here they offer a mathematical equation, which strengthens reputation, which increases the reputation of VM in relation to faster execution and task of work by performance exchange. In this paper I have changed the resource unit system to change the performance of the movement of big data. And the future Enhancement should to be the server should automatically change the resource unit based upon the performance.

REFERENCES

[1] M.D. AntoPraveena., Dr. B. Bharathi, A Survey Paper on Big Data Analytics. International Conference on Information, Communication & Embedded Systems (ICICES 2017).

[2] Pradeep S., Jagadish S Kallimani., A Survey on Various Challenges and Aspectsin Handling Big Data. 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT).

[3] HamidrezaAnvari., Paul Lu, The Impact of Large-Data Transfers in Shared Wide-Area Networks: An Empirical Study. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

[4] L. Andersson., G. Swallow, The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols.

[5] Nina T. Bhatti., Richard D. Schlichting, A System for Constructing Configurable High-Level Protocols. SIGCOMM '95 Cambridge, MA USA 0 1995 ACM 0-89791 -711-1 /95/0008.

[6] Punit Gupta., SatyaPrakashGhrera., Trust and Deadline Aware Scheduling Algorithm for Cloud Infrastructure Using Ant Colony Optimization. 2016 1st International Conference on Innovation and Challenges in Cyber Security (ICICCS 2016).

[7] Sunaina Sharma., VeenuMangat.: 2015 Technology and Trends to Handle Big data: survey, Fifth International Conference on Advanced Computing & Communication Technologies.

[8] Muhammad HabiburRehman ., Chee Sun Liew., Assad Abbas., PremPrakashJayaraman., Teh Ying Wah., Samee U. Khan. Big Data Reduction Methods: A Survey. Data Sci. Eng. (2016) 1:265–284.

[9] NabeelZanoon., Abdullah Al-Haj., Sufian M Khwaldeh, Cloud Computing and Big Data is there a Relation between the Two: A Study. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 17 (2017) pp. 6970-6982

[10] ManikaManwal., Amit Gupta., Big Data and Hadoop -A Technological Survey

[11] Chun‐Wei Tsai., Chin‐Feng Lai., Han‐Chieh Chao., Athanasios V. Vasilakos, Big data

analytics: a survey Tsai et al. Journal of Big Data (2015) 2:21.

[12] VatsalJatakia., Sameer Korlahalli., KhushaliDeulkar., A Survey of Different Search Techniques for Big Data. 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).

[13] SunghaeJun, TECHNOLOGY ANALYSIS FOR INTERNET OF THINGS USING BIG DATA LEARNING. IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.

[14] Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A.S., Buyya, R.: Big data computing and clouds: trends and future directions. J. Parallel Distrib. Comput. 79–80, 3–15 (2015).

[15] AmoghPramodKulkarni, Mahesh Khandewal, ―Survey on Hadoop and Introduction to YARN‖, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).

[16] Padmapriya, V., Gowri, V., Lakshmipriya, K., PremKumar, K., Thiyagarajan, B., "Perspectives, motivations and implications of big data analytics", (2015) ACM International Conference Proceeding Series, 06-07-March-2015, art. no. 2743099,

[17] Rao, D.N., Sathian, D., Dhavachelvan, P., Raghav, R.S., Prem Kumar, K., "Big data scalability, methods and its implications: A survey of current practice", (2015) ACM International Conference Proceeding Series, 06-07-March-2015, art. no. 2743121,

[18] Karthikeyan, P., Sathian, D., Raghav, R.S., Abraham, A., Dhavachelvan, P., "A comprehensive survey on variants and its extensions of BIG DATA in cloud environment", (2015) ACM International Conference Proceeding Series, 06-07-March-2015, art. no. 2743097,

[19] Padmapriya, V., Gowri, V., Lakshmipriya, K., Vinothini, S., PremKumar, K., "Demystifying challenges, opportunities and issues of Big data frameworks", (2015) ACM International Conference Proceeding Series, 06-07-March-2015, art. no. 2743110,

[20] Bandi, R., Gouse, S., "A comparative analysis for big data challenges and big data issues using information security encryption techniques1, 2", (2017) International Journal of Pure and Applied Mathematics, 115 (8 Special Issue), pp. 245-251.