ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Priority Based Memory Channel Architecture for Virtualization in Cloud Computing

[I]Kulwinder Kaur, [II]Dr. Vinay Gautam
[I]M.Tech Student, [II]Asst. Professor, CSE
[I,II]Desh Bhagat University, Punjab, India

Abstract
Live migration in cloud is key aspect of cloud computing. In this work, we propose to overcome the performance issues of Single Memory channel with Multi channel memory channel. Single memory channel affect the performance of CPU, communication and memory. Here, we present a priority memory channel model to overcome the issues which affecting the performance of system. Priority Algorithm is used to fulfil the commitments of SLA agreement through giving priority to highest paying customers and medium level of priority to less paying customers. The Starvation problem is also solved by applying the Round Robin algorithm. The cores of the CPU are continuously increasingly but in the other way it's an overhead. It's necessary to utilize the cores of CPU properly. The rapid increase in the cloud and its infrastructure has lead to increase in the various aspects of the cloud computing. This has resulted in significant increase in the virtualization technology and lays more emphasis on virtual machines, live migration and so on. So these issues will be dealt with high attention to get solutions.

Keywords
Cloud computing, SLA, Live migration, Memory channel

I. Introduction

Cloud Computing is the need of hour. For our purposes, the Cloud is a large group of interconnected computers. These computers can be personal computers or network servers; they can be public or private. It will allow the users to share resources rather than having their own personal devices and local servers to handle application. The word cloud is phrased as "the cloud" which is used as a metaphor for the internet. It is the type of internet based computing in which the user gets various types of services such as applications, servers and storage. With cloud computing, the software programs you use aren't run from your personal computer, but are rather stored on servers accessed via the Internet. If your computer crashes, the software is still available for others to use. Same goes for the documents you create; they're stored on a collection of servers accessed via the Internet. Anyone with permission can not only access the documents, but can also edit and collaborate on those documents in real time.

Web based Cloud Computing: Under this the companies rather than building full application needed by them they use the functionality offered to them by the web services.

Infrastructure as a Service (IaaS): It works on the storage and by using the IaaS, organizations do not have to worry about the dedicated servers on site. They can shrink and grow their storage as per their need. The consumer does not control the cloud infrastructure but it has control over operating systems, storage, and deployed applications.

Software as a Service (SaaS): Under SaaS the users can use an application simultaneously without worrying about the storage capacity or other issues. It can be used by many personas for a variety of functions. The common example is the Google docs. Various applications are accessible in from client in cloud computing either by using thin client network (web browser) or a program interface.

Platform as a Service (PaaS): Under these services the organizations does not have to worry about maintaining servers and hard drives. They can run their own applications on the cloud service's platform.

Utility Services: In this type of cloud computing, companies can store all of their data remotely and they can create their own virtual data centre. This is suitable for the companies those who have a large database to store.

Managed Services: These are the applications that are used by the cloud providers rather than the members of an organization. A common example of managed services is an anti-spam service.

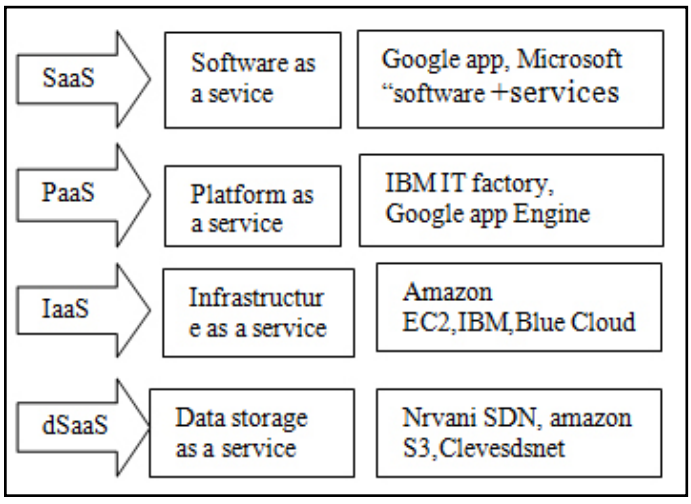

Fig. 1: Types of Services in Cloud Computing

Cloud computing is an advancement in distributed system, where large number of computers are connected through a real-time communication network such as internet, intranet and extranet. Generally it is referred to network-based services, which are provided by real server hardware, and are served by virtual hardware, simulated by software running on one or more real machines. A virtual machine is a software implementation of computer in which the programs are executed just like a physical machine. Virtual machines can be classified into two groups based on their use and degree of correspondence to any real machine.

1. A system virtual machine provides a complete system platform which supports the execution of a complete operating system. [6]
2. A process virtual machine is designed to run a single program, which means that a single program is supported by it. [6]

Currently cloud computing hosts various heterogeneous applications with different performance requirements including high performance computing, web applications. How-ever there are many challenges in providing reliable and guaranteed performance service in such consolidation environment. Researchers are always trying to find new solutions of these challenges and enhance virtual machine scheduling algorithm, resource allocation and migration strategies. In cloud computing problem of some dirty pages also occur, which are to be cleaned so that the pages of the application can be prevented from over writing. A dirty page is a page that has been modified in main memory (physical memory) but yet not written in the disk. [1] The concept of dirty pages can be made clearer by knowing about virtual RAM. Virtual memory is software representation of RAM. New operating systems do not allow the direct access to RAM instead they create virtual RAM. Virtual memory is memory taken from the HDD and converted to RAM [6].

II. Litereture Survey

In the literature review many aspects of live migration are and all other issues which are to be handled during the study are seen. There are many ways that shows how the performance of the system can be upgraded and the cost can be minimized to increase the speed and to save both energy and time. The problems which are faced due to the single memory channel are discussed. Though before going deep into the architecture, the concept of live migration and what the single memory channel is explained below:

Ibrahim Takouna et.al (2012) [1] In this research paper live migration is the one of the fascinating feature of virtualization technologies. It is used as a solution for hosts load balancing and maintenance. Live migration consists of two phases: pre-copying and stop-and-copying. In this paper rather than taking the CPU utilization only they have considered the other overheads such as memory bus, network. Here the performance of an application is being discussed.

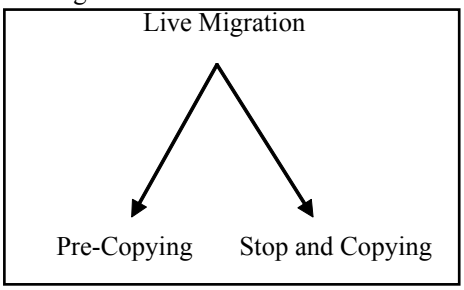

Fig. 2: Phases of live migration

There are different communication techniques of VMs that are simulated including shared memory for multi-threaded applications and network for multi-processes applications. Here the influence of memory bus is shown on multi-threaded applications and multi-processes applications. Next thing which they have given is the NBP benchmark analysis.

Sai Prashanth Muralidhara et.al (2011) [2] The second thing in research is to know how to partition the single memory channel. So here it is discussed about what is channel partitioning. So, first of all it is said that the performance benefits of mapping the pages of applications with largely different memory intensities to separate channels.

Conventional page mapping - in conventional page mapping the requests have to wait until the earlier requests are processed. Though the requests coming from the second source are creating a disturbance to the other but still it continues to run the processes which have arrived earlier.

Channel Partitioning - in this approach the requests that are coming do not wait and are processed at the time they arrive. So there latency of all the requests are eliminated and thus the there is increase in the processing.

James H.Anderson et.al (2005) [3] Next is about the scheduling algorithm which is to be used for our model. The algorithm chosen is the earliest deadline first. This algorithm is scheduled in the multiprocessors. It is a real time operating system algorithm, whish picks the dynamic task priorities, and the job with the nearest absolute deadline gets highest priority. And most of all it is an online algorithm. It is a dynamic algorithm used in real time OS to place processes in a priority queue. Whenever a task finishes the queue will search a task which is closest to its deadlines. It guarantees that the total CPU utilization is not more than 100%. The new algorithm proposed to restrict the task migrations. There is no total utilization constraint. The new algorithm proposed restricts the task migrations. There no total utilization constraint. Scheduling algorithms can be categorised into (1) static (2) dynamic but fixed (3) fully dynamic.

Chongmin Li et.al (2010) [4] this paper tells the concept, which says that memory scheduling algorithms should be designed to handle the memory requests from different threads. This can provide better system throughput and the fairness in the working of the system. A new algorithm known as "priority based fair scheduling" is discussed in which it is said that in it classifies threads memory access behaviour by dynamically updated priorities. Here it says that the threads those are latency sensitive they have top priority for giving the throughput to the system.

They propose a memory scheduling algorithm, Priority- Based Fair Scheduling (PBFS), which classifies threads memory access behaviour by dynamically updated priorities. Latency sensitive threads have top-priority to guarantee system throughput, and starvation of memory-sensitive threads can be avoided simultaneously. Simulation results show that compared with a FCFS scheduler, PBFS improves the system throughput and Fairness metric by 7.4% and 7.7% respectively. The implementation of PBFS is easy and the hardware overhead is small. Starvation of memory-sensitive threads is avoided as there is no latency sensitive thread can consecutively issue a number of memory request. The implementation of PBFS is easy and the hardware overhead is small.

Jaidev P. Patwardhan et.al (2004) [5] In this paper research has been done about the different workloads of different machines. It says that to improve the performance of the web servers it is likely to have good understanding of workloads of different machines. Here they discussed about five different types of workloads in which two serves the static web content and the rest three serves the dynamic web content. From the study of this paper it can be known that the dynamic web content is very small(less of CPU cycles). And for the static web loads there is networking overhead of (upto 25% of CPU cycles).

Karthik Kumar et.al (2011) [6] According to this paper they purposed the method to allocate the resources for real- time tasks. They use the infrastructure as a service model. There is a condition; the real time task has to be completed in the particular time period and also before the deadline. For this problem they purpose a scheme that is EDF- greedy scheme. According to

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

this scheme they consider the temporal overlapping to allocate resources efficiently.

III. Problem Formulation

Since it is clear from the review that there are many problems in existing single memory channel which should be taken care off. The major problems which that are concluded are related to the utilization of CPU, memory subsystem and network. Again there is an issue of migration overhead on the memory subsystem, network. Here there is a problem of communication overhead on CPU and the scheduling of jobs. In order to handle these problems a new model is designed i.e. Priority based memory channel model in which work related to improve the system performance and to utilize the resources fully has been performed. The main task of this model is to stop the jobs from waiting. Once a job is arrived it should be processed at that particular period of time and also take care the SLA agreement. So scheduling should be applied so that the jobs do not have to wait for their processing. Channel portioning algorithm is applied so that the preferred channel is provided to high intensity memory H-high, M-Medium, A-Average and L-Low queues are there.

IV. Proposed Model

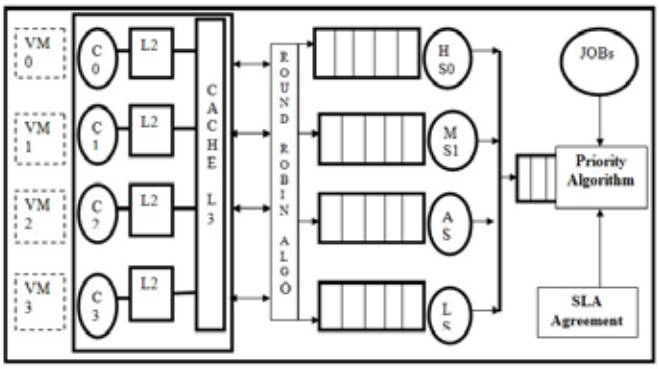

Fig. 3: Priority Based Memory Channel Model

The main drawback of the single memory channel was its single virtual queue. The reason behind it was the scheduling algorithm. So to overcome the problem the numbers of virtual queues were increased from one to multiple. A new scheduling algorithm is used in the research. The scheduling algorithm used is the research is Round Robin algorithm. In the research the priority of the processes is set and then the processors process the instruction given to it.

The working of the Priority based algorithm is stated below:

i. In Priority scheduling, the priority is involved, especially the overhead with small unit time and SLA Agreement.
ii. Service Queue are Label with H-high, M-Medium, A-Average and L-Low according to the priority of job.
iii. High Priority jobs are sent to H, M, A and L respectively depends upon the nature of SLA agreement so that QoS will met.
iii. There should be balanced throughput between FCFS and SJF, shorter jobs are completed faster and jobs having high priority complete on urgent basis.
iv. Good average response time, waiting time in round robin is dependent on number of processes and not the average process length.
v. Because of high waiting time deadlines are rarely met in pure round robin system, so round robin with priority is used.
vi. Starvation can never occur, like in FCFS.

V. Conclusion

The new proposed model will efficient in terms of performance, throughput and overall performance of the system. The algorithm will be implemented on real environment or authenticated simulation tools to define the effectiveness of the old and the new model. The model proposed is somehow capable of performing better in current scenarios and in the current environment, as technology is growing day by day to newer heights. Our future work is to implement this model into cloud computing.

VI. Acknowledgement

I wish to express sincere gratitude and indebtedness to my respected supervisor, Dr. Vinay Gautam and Er. Gurdeep Kaur (H.O..D) for helping me with his experienced and professional ideas to reach where I am today and still showing the light to excel.

I lack words to express my cordial thanks to the members of Departmental Research Committee Dr. Vinay Gautam ,Er Ashish Jalota for their useful comments and constructive suggestions during all the phases of the present study as well as critically going through the manuscript.

References

[1] Takouna, I., Dawoud, W., & Meinel, C. (2012, November). Analysis and Simulation of HPC Applications in Virtualized Data Centers. In Green Computing and Communications (GreenCom), 2012 IEEE International Conference on (pp. 498-507). IEEE.

[2] Muralidhara, S. P., Subramanian, L., Mutlu, O., Kandemir, M., & Moscibroda, T. (2011, December). Reducing memory interference in multicore systems via application-aware memory channel partitioning. In Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture (pp. 374-385). ACM.

[3] Anderson, J. H., Bud, V., & Devi, U. C. (2005, July). An EDF-based scheduling algorithm for multiprocessor soft real-time systems. In Real-Time Systems, 2005.(ECRTS 2005). Proceedings. 17th Euromicro Conference on (pp. 199-208). IEEE.

[4] Li, C., Wang, D., Wang, H., & Xue, Y. Priority Based Fair Scheduling: A Memory Scheduler Design for Chip-Multiprocessor Systems. Tsinghua National Laboratory for Information Science and Technology.

[5] Patwardhan, J. P., Lebeck, A. R., & Sorin, D. J. (2004). Communication breakdown: analyzing CPU usage in commercial web workloads. In Performance Analysis of Systems and Software, 2004 IEEE International Symposium on-ISPASS (pp. 12-19). IEEE.

[6] Kumar, K., Feng, J., Nimmagadda, Y., & Lu, Y. H. (2011, July). Resource allocation for real-time tasks using cloud computing. In Computer Communications and Networks (ICCCN), 2011Proceedings of 20th International Conference on (pp. 1-7). IEEE.