www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 5 Issue 09 September 2016 Page No.18029-18034

Virtual Machine Migration in Cloud Data Centers for Resource management

Chandramouli Reddy[1], Suchithra R[2]

Research Scholar, Jain University ,J C Road Campus, Bangalore Karnataka – 560027, India
email4mouli@gmail.com
Centre for Emerging Technologies Jain Global Campus, Jain University 45th KM, NH 209, Jakkasandra P.O. Kanakapura Taluk, Ramanagara, Dist. Karnataka – 562112, India
r.suchithra@jainnuniveristy.ac.in

Abstract: Cloud computing delivers the computing services over the internet. Cloud services help individuals and organization to use data that are managed by third parties or another person at remote locations. Virtual Machine (VM) is an emulation of a particular computer system. In cloud computing, Virtual machine migration is a useful tool for migrating Operating System instances across multiple physical machines. It is used to load balancing, fault management, low-level system maintenance and reduce energy consumption. There are various techniques and parameters available for VM migration. This paper presents the various virtual machine migration techniques. Many cloud data centers have problems in understanding and implementing the techniques to manage, allocate and migrate the resources in their premises. The consequences of improper resource management may result into underutilized and wastage of resources which may also result into poor service delivery in these data centers. Resources like; CPU, memory, Hard disk and servers need to be well identified and managed.

Keywords: *Cloud Computing, Virtualization, VM Live migration, Data Centers, Resource Management, Server Sprawling, Server Consolidation*

Introduction:

Cloud computing distributes the computing tasks to the resource pool made from a large number of computers. VMs refer to one instance of an operating system along with one or more applications running in an isolated partition within the computer. There will be multiple virtual machines running on top of a single physical machine. When one physical host gets overloaded, it may be required to dynamically transfer certain amount of its load to another machine with minimal interruption to the users. This process of moving a virtual machine from one physical host to another is termed as migration. In the past, to move a VM between two physical hosts, it was necessary to shut down the VM, allocate the needed resources to the new physical host, move the VM files and start the VM in the new host. Live migration makes possible for VMs to be migrated without considerable downtime. The transfer of a VM actually refers to the transfer of its state. This includes its memory, internal state of the devices and that of the virtual CPU. Among these, the most time-consuming one is the memory transfer. Two parameters are considered while performing the live VM-migration are:

1) **Downtime -** Down time refers to the time during which the service of the VM is not available.

2) **Migration Time -** Migration time refers to the total amount of time required to transfer a virtual machine at source to destination node without affecting its availability. Virtualization is major concept of cloud computing. It is becoming popular in cloud computing environments.

2. Related Work

Virtualization help in partitioning of one physical machine into number of virtual machines that runs concurrently and it also shares the same physical resources. Virtual machine migration is done from one physical machine to another machine. It is used for load balancing and physical machine fault tolerant. It can also be used to reduce power consumption in cloud data centers.

Virtual Machine Migration methods are divided into two types: 1) Hot (live) migration- Virtual machine keeps running while migrating and does not lose its status.

2) Cold (non-live) migration- The status of the VM loses and user can notice the service interruption.

User doesn't feel any interruption in service in hot (live) migration. In live migration process, the state of a virtual machine to migrate is transferred. The state consists of its memory contents and local file system. Local file system need not be transferred. In cold migration, first, VM is suspended, then its state is transferred, at last VM is resumed at destination host.

a) Live VM Migration- Live migration is a technology used for load balancing and optimization of VM deployment in data centers. With the help of live migration, VMs can be transferred to another node without shutting down. Live migration is classified into two steps – (i) Control is switched to the destination. (ii) Data Transferring (memory/disk) to the destination.

Pre-copy- In this, first Memory is transferred and after this execution is transferred. The pre-copy method is used to transfer the memory to the destination node over a number of iterations.

Post-copy- In this, First execution is transferred and after this, memory is transferred. Unlike pre-copy, in post copy the Virtual CPU and devices on the destination node is transfer in the first step and starts the execution in second step. Following metrics are used to measure the performance of live migration.

i) Preparation- In this, resources are reserved on the destination which performed various operations.

ii) Downtime- Time during which the VM on the source host is suspended.

iii) Resume- It does the instantiation of VM on the destination but with the same state as suspended source.

iv) Total time- The total time taken in completion of all these phases is called Total Migration time.

3. Related Work

3.1 Data Centers

Data centers have grown in popularity as the processing power required by businesses exceeds what they can maintain within their own corporate infrastructure. Data centers have sprung up to act as large server farms that use economy of scale to provide computation and storage resources to one or more businesses with greater efficiency. A business may own and operate its own data centers, or a data center may be operated by a service provider that in turn rents shares of its resources to other businesses.

Data center operators face challenges from the initial capacity planning stages of deploying and provisioning new applications, to efficiently allocating resources to meet the application performance guarantees of live systems. At the same time, they must deal with the problems of maintaining the hardware reliability, cooling, and energy needs of running thousands of servers. All of these issues are compounded by both the massive scale of modern data centers, and the fast paced workload dynamics of many data center applications.

Table 2.1 lists four key problem areas for data center operators. Administrators first must deal with infrastructure challenges such as determining data center architectures and providing sufficient cooling and power for large numbers of servers. A current trend in data center architecture is the use of large scale, modular data centers composed of shipping containers filled with servers [1], but more radical proposals range from micro data centers placed inside condominium closets [2] to floating barges filled with servers running off of power generated from ocean currents [3]. The increasing energy consumption of data centers is a growing concern, and work is being done in the “green computing” area to better manage the power and cooling systems in data centers [4, 5, 6].

Next, data center operators must deal with the deployment and planning problems related to estimating a data center’s capacity and initial provisioning for new applications [7]. This may require models of an application’s resource requirements [8, 9],and an understanding of how they are impacted by different hardware configurations [10].As data centers attempt to improve resource utilization through server consolidation, it also becomes necessary for data center operators to understand how the placement of applications impacts performance and resource consumption [11].

Efficient resource management is a key concern for data center operators looking to both meet application Service Level Agreements (SLAs) and reduce costs. Shared hosting platforms attempt to multiplex physical resources between multiple customer applications [12, 13]. However, without virtualization, it is difficult to provide strong isolation between applications, and operating systems must be modified to fairly allocate resources [14, 15].

This work focuses on three of these areas: deployment, resource management, and reliability, with an emphasis on how virtualization can provide improved solutions.

Area	Challenges
Infrastructure	Server & Network architecture, Cooling & power Management
Deployment	Capacity Planning, Service Placement, application modelling
Resource Mgmt.	Storage, server, & network provisioning, monitoring
Reliability	High availability, fault tolerance, security
Applications	Clustering frameworks, performance, configuration management

3.2 Server Virtualization

Virtualization is not a new technology, but it has regained popularity in recent years because of the promise of improved resource utilization through server consolidation. The virtualization of commodity operating systems in [16] has led to the wide range of both commercial and open source virtualization platforms available today [17, 18, 19, 20].Virtualization can be performed at the application level (e.g., Oracle's Virtual Box or VMware's Player software), within an operating system (e.g., KVM or Linux Vservers), or even below the operating system (e.g., VMware ESX). In this thesis we focus on the Xen and VMware full system virtualization platforms. VMware's ESX platform is a full virtualization technique that provides a "bare-metal" hypervisor that manages a set of virtual machines running unmodified operating systems [21]. Xen uses a Para virtualization technique that requires small changes to be applied to the operating systems within each VM and a host OS to run device drivers, but allows for a simpler hypervisor layer [22].Both systems support fine grain management of memory and CPU resources, as well as the ability to transparently migrate running virtual machines from one physical server to another [23, 24, 25]. In this thesis we make frequent use of the resource management and migration tools provided by each virtualization platform.

3.3 Virtualization in Data centers

Server virtualization has become popular in data centers since it provides an easy mechanism to cleanly partition physical resources, allowing multiple applications to run in isolation on a single server. Virtualization helps with server consolidation and provides flexible resource management mechanisms, but can introduce new challenges.
Determining where to run applications in a shared environment remains a challenge [26], and virtualization adds new difficulties due to the variable virtualization overheads seen by different applications and platforms [27]. Our work explores a new factor to consider when placing VMs, the potential for memory sharing, and helps build models that characterize VM overheads.

Some commercial systems now exist for automating the management of VM resources [28, 29], and a variety of research projects have proposed schemes for management of processing and memory resources. Our work was some of the first to combine automated management of VM resources with dynamic migration to balance load within commercial data centers.

Reliability is an important feature for data center applications, and virtualization has been used to provide increased resiliency in the face of crash failures [30, 31]. Our work extends these ideas to provide disaster recovery services across data centers, allowing applications to fail over from one to another with no data loss. We propose a new replication approach based on the ideas of external synchrony [32], which uses speculative execution to combine the best aspects of synchronous and asynchronous approaches.

4. VIRTUAL MACHINE MIGRATION

Virtualization technologies promise great opportunities for reducing energy and hardware costs through server consolidation. However, to safely transition an application running natively on real hardware to a virtualized environment, one needs to estimate the additional resource requirements incurred by virtualization overheads.

4.1 Background and Motivation

Modern data centers employ server virtualization to slice larger, underutilized physical servers into smaller virtual ones. By allowing for greater resource multiplexing, virtualization can decrease energy utilization and hardware costs. While many businesses would like to lower costs by moving their applications from running on physical hardware to virtual machines, they must ensure that this transition will not disrupt application performance by incorrectly estimating the resource requirements of the virtualized application. A naive solution is to simply monitor the workload of native applications and attempt to provision the virtual servers based on the observed peak resource requirements. However, this does not account for the different types of overhead caused by the virtualization layer, and can lead to either over- or under-provisioning depending on the nature of the application.

4.2 Virtualization Overheads

Server consolidation is an approach to reduce the total number of servers in response to the problem of server sprawl, a situation in which multiple, under-utilized servers take up more space and consume more resources than can be justified by their workloads. A typical approach for evaluating which workloads can be efficiently consolidated together is based on multi-dimensional "binpacking" of resource usage traces. Under such an approach, each application is characterized by its CPU, I/0 and memory usage over time. Then a binpacking algorithm finds a combination of workloads with resource requirements which do not exceed the available server resources. After the initial workload placement, specialized workload management tools are used [33, 34] to dynamically adjust system resources to support the required application performance.

4.3 Implementation

The algorithm is choosing the best possible server for allocation, it prevents the wastage of free CPU available on the servers.

The Best Fit Resource Migration Algorithm (BRMA) is

Step 1: Calculate remaining memory for all Vm's
i.e., {Vm1, Vm2, Vmn}
Step 2: Calculate total memory of Vm's from the above calculation of Individual remaining resources.
Total memory = {Remaining memory Vm1+Vm2+……. +Vmn}
Step 3: If incoming memory > Total memory
Low memory condition
Exit;
Step 4: If incoming memory < Total memory
Calculate best fir for incoming memory allocation.

Step 5: Repeat step for all Vm's
Step 6: Obtain best fit Vm
Calculate best fit process (P1) to be migrated.
Step 7: If any Vm memory nearest to P1
-> Then assign P1 to Vm
-> Retabulate total memory remaining
-> Migrate incoming to freed Vm
Step 8: Else Exit;

4.4 Experimental evaluation

Before Migration: In this experiment, we consider three virtual machines that consists of process that may be allocated or migrated depending on the total memory allocation of virtual machines. Here, Best Fist Resource Migration algorithm can be used to migrate the virtual machines across the data centers

Memory	**VM_1**	**VM_2**	**VM_3**
Total Memory	25	30	26
Occupied Memory	23	15	20
Unoccupied Memory	2	15	6
Memory Occupancy by processes	{14,7,2}	{9,6}	{13,7}

Case 1:

Step 1: New Incoming process memory- 17

VM_1=|Incoming memory – Unoccupied memory| = |17-2| =15

VM_2=|Incoming memory – Unoccupied memory| = |17-15| = 2

VM_3=|Incoming memory – Unoccupied memory| = |17-6| =11

Least of {15, 2, 11} = 2 => VM_2

So, VM_2 is considered for migration

Step 2:

Now, Incoming process Memory = 17 and unused memory = 15

2 units to be free minimum in VM_2. Process nearest to 2 in {9, 6} is 6.

6 is searched to fit across other VM's. VM_3 unoccupied memory is 6.

So, process from VM_2 is migrated to VM_3.

Memory	VM1	VM2	VM3
Total Memory	25	30	26
Occupied Memory	23	9	26
Unoccupied Memory	2	21	0
Memory Occupancy by processes	{14,7,2}	{9}	{13,7,6}

Step 3:

Now, Incoming process Memory = 17 and unused memory = 15

2 units to be free minimum in VM_2. Process nearest to 2 in {9, 6} is 6.

6 is searched to fit across other VM's. VM_3 unoccupied memory is 6.

So, process from VM_2 is migrated to VM_3.

Memory	VM1	VM2	VM3
Total Memory	25	30	26
Occupied Memory	23	9	26
Unoccupied Memory	2	21	0
Memory Occupancy by processes	{14,7,2}	{9}	{13,7,6}

5. SUMMARY AND FUTURE WORK

In this paper, we have implemented both migration and allocation algorithms for resource management in cloud datacenters using two resources, that is; CPU and memory. The purpose of this research was to minimize underutilized and avoid over utilized resources that may perhaps create unnecessary migrations due to unpredictable workloads. The results have shown how efficiently and effectively resources could be managed, allocated and migrated among the servers. The target was to have less number of tasks/jobs migrated with maximum resource allocation, less number of servers used and amount of resources saved with minimum costs. As migrations consume additional energy and have a negative impact on the performance, before initiating a migration, the reallocation controller had to ensure that the cost of migration does not exceed the benefit. Virtualization technologies can be used to improve resource management, simplify deployment, and increase the resilience of modern data centers.

Purpose of this research is to minimize underutilized and avoid over utilized resources that may perhaps create unnecessary migrations due to unpredictable workloads. Results have shown how efficiently and effectively resources could be managed, allocated and migrated among the servers. The target is to have less number of tasks/jobs migrated with maximum resource allocation, less number of servers used and amount of resources saved with minimum costs

6. References

[1] Katz, Randy. IEEE spectrum: Tech titans building boom. http://www.spectrum.ieee.org/green-tech/buildings/tech-titans-building-boom.

[2] Church, K., Hamilton, J., and Greenberg, A. On delivering embarrassingly distributed cloud services. Hotnets VII (2008).

[3] Clidaras, Jimmy, Stiver, David, and Hamburgen, William. Water-Based data center (patent application 20080209234).

[4] Chase, J., Anderson, D., Thakar, P., Vahdat, A., and Doyle, R. managing energy and server resources in hosting centers. In Proceedings of the Eighteenth ACM Symposium on Operating Systems Principles (SOSP) (Oct. 2001), p. 103116.

[5 Lim, K., Ranganathan, P., Chang, J., Patel, C., Mudge, T., and Reinhardt, S. Understanding and designing new server architectures for emerging warehouse-computing environments. In Computer Architecture, 2008. ISCA'08. 35th International Symposium on (2008), pp. 315–326.

[6] Raghavendra, R., Ranganathan, P., Talwar, V., Wang, Z., and Zhu, X. No power struggles: A unified multi-level power management architecture for the data center.ASPLOS, March (2008).

[7] Zhang, Qi, Cherkasova, Ludmila, Mathews, Guy, Greene, Wayne, and Smirni, Evgenia. R-capriccio: a capacity planning and anomaly detection tool for enterprise services with live workloads. In Middleware '07: Proceedings of the ACM/IFIP/USENIX 2007 International Conference on Middleware (New York, NY, USA, 2007), Springer-Verlag New York, Inc., pp. 244–265.

[8] Stewart, C., and Shen, K. Performance modeling and system management for multicomponent online services. In Symposium on Networked Systems Design and Implementation (2005).

[9] Zhang, Q., Cherkasova, L., and Smirni, E. A Regression-Based analytic model for Dynamic resource provisioning of Multi-Tier applications. In Proc. ICAC (2007).

[10] Stewart, Christopher, Kelly, Terence, Zhang, Alex, and Shen, Kai. A dollar from 15 cents: cross-platform management for internet services. In ATC'08: USENIX 2008 Annual Technical Conference on Annual Technical Conference (Boston, Massachusetts,2008), pp. 199–212.

[11] Urgaonkar, B., Rosenberg, A., and Shenoy, P. Application placement on a cluster of servers. In Internernational Journal of Foundations of Computer Science (October 2007), vol. 18, pp. 1023–1041.

[12] Urgaonkar, B., Shenoy, P., Chandra, A., and Goyal, P. Dynamic provisioning for multi-tier internet applications. In Proceedings of the 2nd IEEE International Conference on Autonomic Computing (ICAC-05), Seattle, WA (June 2005).

[13] Chandra, Abhishek, Gong, Weibo, and Shenoy, Prashant. Dynamic resource allocation for shared data centers using online measurements. In Proceedings of the 2003 ACM SIGMETRICS international conference on Measurement and modeling of computer systems (San Diego, CA, USA, 2003), ACM, pp. 300–301.

[14] Chen, Xu, Mao, Z Morley, and Van der Merwe, Jacobus. ShadowNet: a platform for rapid and safe network evolution. In USENIX Annual Technical Conference (2009).

[15] Waldspurger, Carl A., and Waldspurger, Carl A. Lottery and stride scheduling: Flexible proportional-share resource management. In In Proc. First Symposium on Operating Systems Design and Implementation (1995), pp. 2–90.

[16] Bugnion, Edouard, Devine, Scott, and Rosenblum, Mendel. DISCO: Running Commodity Operating Systems on Scalable Multiprocessors. In SOSP (1997), pp. 143–156.

[17] Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., and Warfield, A. Xen and the art of virtualization. In Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP'03), Bolton Landing, NY (October 2003), pp. 164–177.

[18] Vmware esx bare-metal hypervisor. www.vmware.com/products/vi/esx.

[19] Kernel based virtual machine. http://www.linux-kvm.org/.

[20] Microsoft hyper-v server. www.microsoft.com/hyper-v-server.

[21] Vmware esx bare-metal hypervisor. www.vmware.com/products/vi/esx.

[22] Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R.,Pratt, I., and Warfield, A. Xen and the art of virtualization. In Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP'03), Bolton Landing, NY (October 2003), pp. 164–177.

[23] Clark, Christopher, Fraser, Keir, Hand, Steven, Hansen, Jacob Gorm, Jul, Eric,Limpach, Christian, Pratt, Ian, and Warfield, Andrew. Live Migration of Virtual Machines. In Proc. of NSDI (Boston, MA, May 2005).

[24] Nelson, Michael, Lim, Beng-Hong, and Hutchins, Greg. Fast Transparent Migration for Virtual Machines. In USENIX Annual Technical Conference (2005).

[25] Waldspurger, Carl A. Memory resource management in Vmware esx server. SIGOPS Oper. Syst. Rev. 36, SI (2002), 181–194.

[26] Urgaonkar, B., Rosenberg, A., and Shenoy, P. Application placement on a cluster of servers. In Internernational Journal of Foundations of Computer Science (October 2007), vol. 18, pp. 1023–1041.

[27] Cherkasova, L., and Gardner, R. Measuring CPU overhead for I/O processing in the xen virtual machine monitor. In USENIX Annual Technical Conference (Apr. 2005).

[28] Whitepaper, VMware. Drs Performance and best practices.

[29] Amazon ec2 elastic load balancing. http://aws.amazon.com/elasticloadbalancing/.

[30] Cully, Brendan, Lefebvre, Geoffrey, Meyer, Dutch, Feeley, Mike, Hutchinson, Norm, and Warfield, Andrew. Remus: High Availability via Asynchronous Virtual Machine Replication. In Proc. of NSDI (2008).

[31] Vmware high availability product page. www.vmware.com/products/vi/vc/ha.html.

[32] Nightingale, Edmund B., Veeraraghavan, Kaushik, Chen, Peter M., and Flinn, Jason.Rethink the Sync. In Proc. of OSDI (2006).

[33] IBM Enterprise Workload Manager. http://www.ibm.com/developerworks/autonomic/ewlm/.

[34] HP-UX Workload Manager. http://www.hp.com/products1/unix/operating/wlm.