Understanding the Concept of Migration for Datacenter on Cloud Computing

Kruti Shah[1] Manthan Shah[2]
[1,2]Masters in Information Technology (Pursuing)
[1,2]Kalol Institute of Technology & Research Center, Gujarat, India

Abstract— Virtual Machines (VMs) refers to the software implementation of a computer that runs its own OS and applications as if it was a physical machine. Live Virtual Machine(VM) migration is the process of moving VM from one physical host to another without stopping the services or applications running on VM. Live migration is used in the case of proactive failure, load balancing, resource scheduling, and server consolidation. In this paper, we have tried to explain goals of migration and different phases of live migration. There are 3 main steps in live migration process and we have explained its importance in detail.

Key words: Cloud Computing, Virtualization, Live Migration.

I. INTRODUCTION

Cloud computing can be considered as a new computing paradigm with many exciting features like greater flexibility and lower expenses for installation and maintenance. Cloud computing is defined by NIST[1] as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. Because of this, cloud computing has been receiving a good deal of attention lately. Compared with general network service, cloud computing is easy to extend, and has a simple management style. Cloud is not only a collection the computer resources, but also provides a management mechanism and can provide services for millions of users simultaneously. The concept of virtualization is now applied to cloud data centers. When the storage and computing capacity of the server cluster are surplus, we need not purchase servers, all we need to do is to add a virtual machine running on the server. If the cluster is large enough, the request of adding server will have marginal effect, and then we can save the money that should be used for purchasing new servers.

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 to be considered while performing the live VM-migration are downtime and migration time. Migration time refers to the total amount of time required to transfer a virtual machine at source node to destination node without affecting its availability. Down time refers to the time during which the service of the VM is not available.

In this Paper Section 2 describes related work for Migration. Sections 3 describes Live-Migration with Datacenter model. Section 4 and 5 describe Conclusion and References.

II. RELATED WORK FOR MIGRATION

Elmroth et al. [2] have formulated technology neutral interfaces and architectural additions for handling placement, migration, and monitoring of VMs in federated cloud environments, the latter as an extension of current monitoring architectures used in grid computing. The interfaces presented adhere to the general requirements of scalability, efficiency, and security in addition to specific requirements related to the particular issues of interoperability and business relationships between competing cloud computing infrastructure providers. In addition, they may be used equally well locally and remotely, creating a layer of abstraction that simplifies management of virtualized service components.

Beloglazov et al. [3] have proposed a method for dynamic consolidation of VMs based on adaptive utilization thresholds, which ensures a high level of reaching the SLA (Service Level Agreements).They also validate the high efficiency of the proposed technique across different kinds of workloads using workload traces from more than a thousand Planet Lab servers. Dynamic consolidation of virtual machines (VMs) and switching idle nodes off allow Cloud providers to optimize resource usage and reduce energy consumption.

Beloglazov et al. [4] have invented an efficient resource management policy for virtualized Cloud data centers. The objective is to continuously consolidate VMs leveraging live migration and switch off idle nodes to minimize power consumption, while providing required Quality of Service. We present evaluation results showing that dynamic reallocation of VMs brings substantial energy savings, thus justifying further development of the proposed policy. Moreover, modern Cloud computing environments have to provide high Quality of Service (QoS) for their customers resulting in the necessity to deal with power performance.

Voorsluys et al. [5] have formulated performance evaluation on the effects of live migration of virtual machines on the performance of applications running inside Xen VMs. Results show that in most cases, migration overhead is acceptable but cannot be disregarded, especially in systems where service availability and responsiveness are governed by strict Service Level Agreements (SLAs). Despite that, there is a high potential for live migration applicability in data centers serving enterprise-class Internet applications. In particular, the capability of virtual machine (VM) migration brings multiple benefits such as higher performance, improved manageability and fault tolerance. Moreover, live migration of VMs often allows work-load movement with a short service downtime.

Sonnek et al. [6] have formulated a decentralized affinity-aware migration technique that incorporates heterogeneity and dynamism in network topology and job communication patterns to allocate virtual machines on the available physical resources. Our technique monitors network affinity between pairs of VMs and uses a distributed bartering algorithm, coupled with migration, to dynamically adjust VM placement such that communication overhead is minimized. Besides, their technique is able to adjust to dynamic variations in communication patterns and provides both good performance and low network contention with minimal overhead.

III. LIVE MIGRATION

Virtualization technology is the one that hides the details of physical hardware and provides virtualized resources for high-level applications. An essential characteristic of a virtual machine is that the software running inside it is limited to the resources and abstractions provided by the VM. The software layer that provides the virtualization is called a Virtual Machine Monitor (VMM) or hypervisor. It virtualizes all of the resources of a physical machine, thereby defining and supporting the execution of multiple virtual machines. Virtualization can provide significant benefits in cloud computing by enabling virtual machine migration to balance load across the data centres.

A. Migration Goals[14]

Migration is mainly done for dynamic resource management. Its main goals are as follows:

– *Load Balancing:* This reduces the inequality of resource usage levels across all the PMs in the cluster. This prevents some machines from getting overloaded in the presence of lightly loaded -machines with sufficient spare capacity. Live migration can be used to balance the system. The overall system load can be balanced by migrating VMs from overloaded PMs to under-loaded PMs.

– *Server Consolidation:* In order to reduce server sprawl in data centres, server consolidation algorithms are required. These algorithms are VM packing heuristics which try to pack as many VMs as possible on a PM so that resource usage is improved and unused or under-utilized machines can be turned off. Consolidation will result in reduced power consumption and thus reducing overall operational costs for data centre administrators. Live migration of VMs could achieve this. Based on the load conditions, under-utilized machines having resource usage below a threshold and overloaded machines having resource usage above a certain threshold are identified, and migrations are triggered to tightly pack VMs to increase overall resource usage on all PMs and free up resources/PMs if possible.

– *Hotspot & Coldspot Mitigation*: The detection of hotspots and coldspots are always based on thresholds which are set by the data center owner or based on the Service Level Agreements specified by the clients. Usually, a higher resource usage value close to maximum is set as the upper threshold and a very low resource usage value is set as the lower threshold. PMs having resource usage values beyond the upper threshold are said to have formed hotspots, and whose usage values below the lower threshold are said to have formed coldspots. The former implies over-utilization and the latter implies under-utilization, applicable across any resource dimension. These conditions are inherently taken care of in the above mentioned consolidation and load balancing algorithms.

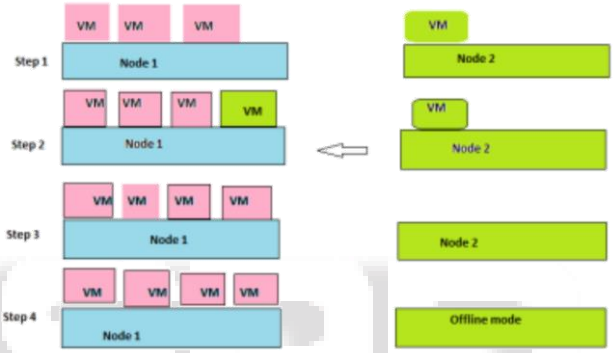


Fig. 1: Live VM migration

The above diagram shows the live virtual machine migration. Initially three VM working on node 1 and one VM working on node2.Migrations allow the movement of VM from node 1 to node 2. Live virtual machine migration having mainly two performance metrics.

– Total migration time: It is defined as the total time taken to migrate a virtual machine from its host machine to the target machine.

– Down time: Down time is defined as the duration of time at which services are not available to the users.

The key challenges in the live virtual machine migrations are minimal service downtime and total migration time [4]. Migration techniques are trying to reduce both migration time and down time.

For each of the three goals — consolidation, hotspot mitigation and load balancing — VM migration-based heuristics need to address three important questions:

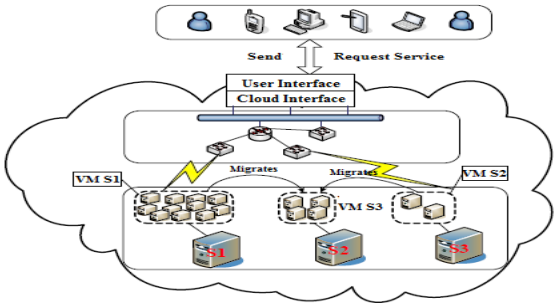
– When to migrate

– Which VMs to migrate

– The set of destination host machines for migration.

Migration heuristics address each of these questions based on several constraints: overhead of the migration process, impact on applications during migration, degree of improvement in intended goals of performance of resource utilization, and so on.

B. Data Centre Model



– When to migrate

There are many situations when migration of VMs becomes necessary to maintain the overall efficiency of the data center.

– Periodi

The migrations in a data center can be triggered periodically. For example, data centers in one part of world may be heavily used in daytime (9 a.m. to 9 p.m.), whereas they may be underloaded during the night. Such "time of day" migration of VMs ensures that VMs are "near" clients, and the communication delays and overheads are minimized. Migrations can also be done periodically to consolidate the reduced loads.

– Due to Hot Spot

A hot spot is the overloaded condition of a PM. It can also be defined as the state when performance of a system falls below the minimum acceptance level. Detection of a hot spot can be done both proactively and reactively. Proactive hot spot detection techniques predict the occurrence of a hot spot by analyzing the trends in resource utilizations of the VM. If the resource utilization shows an increase for some time window, it is likely that it may result in a hot spot in the future. Such time series analysis-based techniques help avoid hot spots even before they occur. One such technique to predict CPU utilization is discussed in [9]. More sophisticated proactive techniques analyze the request arrival rates. Increase in request arrivals suggests that the VM will require more resources to fulfill them, thus causing a potential hot spot. Reactive hot spot detection techniques use more direct techniques like observing the page thrashing rate, CPU and memory utilization levels, and so on. Hot spots can be locally mitigated if enough capacity is available at the host PM. Extra resources can be allocated to the VM showing signs of overload. When extra capacity is not available locally, migration is the only option available.

– Excess Spare Capacity

Low utilization of PMs results in resource wastage. An optimum level of utilization is required to be maintained for the efficient working of a data center. Physical machines that have excess spare capacity (i.e., low resource utilization) cause overall inefficiency in the data center. At the level of a PM, the hypervisors have monitoring tools, similar to normal operating systems, which can provide the utilization information of different resources for that machine. Resource utilization levels of PMs across a data center are continuously monitored, and whenever the utilization levels fall below a certain threshold, migrations can be triggered. When a number of PMs are underutilized, VMs are migrated from such machines to make them completely free. Such "freed" PMs can then be shut down to save power, which results in consolidation.

– Load Imbalance

Virtual machines change their resource requirements dynamically. This dynamism leads to imbalances in the resource utilization levels of different PMs. Some PMs can get heavily loaded while others may be lightly loaded. In a data center, resource utilization levels of PMs are monitored continuously. If there is large discrepancy in the utilization levels of different PMs, load balancing is triggered. Load balancing involves migration of VMs from highly loaded PMs to low loaded ones. An overloaded PM is undesirable as it causes delays in service of user requests. Similarly, the PMs that are lightly loaded cause inefficient resource utilization.

– Addition/Removal Of Virtual Machines And Physical Machines

Virtual machines and PMs can be added and removed in a virtualization- based data center. Addition/removal of VMs and PMs affects the availability of the resources and may require a change in the placement plan of VMs. A new PM can be used to offset the load of an overloaded PM by migrating VMs from the latter to the former. Similarly, hosting new VMs may result in future overloads of some PMs, which again require migrations to be triggered.

– Which virtual machine to migrate

Selecting one or more VMs for migration is a crucial decision of the resource management heuristic. The migration process not only makes the VM unavailable for a certain amount of time but also consumes resources like network and CPU on source and destination PMs. Performance of other VMs that are hosted on source and destination PMs are also affected due to increased resource requirements during migration. Some VM selection approaches are straightforward and only consider the VM that is resource constrained (e.g., in a hot spot); other approaches employ a more holistic approach where all the VMs on a PM are considered before selecting the candidate VM. Generally, the aim of VM selection is to minimize the migration effort.

– *Resource Constrained Virtual Machine*

This is the easiest way to select the candidate VM for migration. The VM whose resource requirements cannot be locally fulfilled is selected for migration. During hot spots, it is easy to find the most loaded VMs; hence, this simple selection can work. However, in operations like consolidation and load balancing, where the cause is not a single VM, the choice is not straightforward.

– *Affinity based*

These heuristics also incorporate other objectives instead of considering the resource requirement only. For example, some *affinity-aware* migration techniques consider communication costs among VMs while performing migration. For instance, if two VMs are communicating with each other, it is better to host them on the same PM. This will reduce the overall communication cost among the VMs by reducing network usage. Similarly, memory sharing between VMs can also affect the VM selection for migration. Migrating a VM to a PM where it can share memory with other hosted VMs can result in effective memory usage [9]. Virtual machines, which share memory, can be migrated together with less effort as similar-content memory pages are required to be transferred only once. Such a scheme is known as *gang scheduling* of VMs. The approach is to proactively track the identical contents of

collocated VMs and transfer those contents only once while migrating all those VMs simultaneously to another PM. This method optimizes both memory and network overhead of migration. Such mechanisms can be fruitful when an entire rack of servers have to be evacuated and all the collocated VMs running on them have to be shifted to a different location. [10]

– Where to migrate

During migration, the destination PM should have enough resources so that it can support the incoming migrating VM. Here we discuss factors for selecting a PM as a destination for a migrating VM.

– Depending on Available Resource Capacity

Only considering the availability of resources at the destination is not enough. Some other factors also need to be taken into consideration, such as whether the destination is a *best fit* (leaving minimum remaining resources) for the migrating VM, how will the performance of VMs that are already hosted on destination PM get affected. The destination selection to minimize waste of resources is a field of research in itself. The schemes proposed use heuristics of bin packing and vector packing problems [11, 12] for destination selection since the optimal placement solution is intractable. For example, in vector-based destination selection, vector arithmetic like dot product is performed on resource vectors to find the best fit [13]. Virtual machines and PMs are sorted in some order based on their resource requirements and then the First Fit or Best Fit scheme is applied to select most suitable PM.

– Depending on Affinity of Virtual Machines

Apart from selecting PMs solely on the basis of resource availability, some schemes try to leverage the relations (or affinity) between the VMs to identify a suitable host PM. For example, a scheme mentioned in [13] tries to achieve consolidation by collocating VMs that have high memory sharing potential. Periodically, based on memory fingerprints of VMs, best matches of hosts for VMs can be found and migrations can be triggered. This scheme is called *memory-aware* migration. The VM can be re-migrated if some other VM on some other PM becomes a better memory sharing partner. The overhead of migration is taken into consideration. The rationale behind this method is that VMs that can share part of their memory will require less overall memory than VMs that do not share memory. Similarly, if two VMs, hosted on different PMs, communicate heavily, one of the VMs can be migrated to the PM where its communicating partner is hosted.

IV. CONCLUSION

In this paper, we have listed so many advantages of Live Migration process. Live migration is the movement of a virtual machine from one physical host to another while continuously powered-up. Total migration time and downtime are two key performance metrics that the clients of a VM service care about the most, because they are concerned about service degradation and the duration that the service is completely unavailable. Migration enables several resource management goals like consolidation, load balancing, and hot spot mitigation. Researchers have leveraged live virtual machine migration to come up with efficient resource management mechanisms. The components — when to migrate, which VM to migrate, and where to migrate — and approaches followed by different heuristics to apply migration techniques for goals of consolidation and hot spot mitigation were discussed.

V. REFERENCES

[1] Cloud Computing for Dummies, Wiley Publishing, Inc.

[2] E. Elmroth and L. Larsson, "Interfaces for Placement, Migration, and Monitoring of Virtual Machines in Federated Clouds," 8th International Conference on Grid and Cooperative Computing, Aug 2009, pp. 253-260, 2009.

[3] Beloglazov and R. Buyya, "Adaptive Threshold-Based Approach for Energy- Efficient Consolidation of Virtual Machines in Cloud Data Centres," 8th International Workshop on Middleware for Grids, Clouds and e-Science, Dec 2010, pp. 1-6, 2010.

[4] Beloglazov and R. Buyya, "Energy Efficient Allocation of Virtual Machines in Cloud Data Centres," 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, May 2010, pp.577-578, 2010.

[5] W. Voorsluys, J. Broberg, S. Venugopal, and R. Buyya, "Cost of Virtual Machine Live Migration in Clouds: A Performance Evaluation," 1st International Conference on Cloud Computing, pp. 254-265, 2009.

[6] J. Sonnek, J. Greensky, R. Reutiman and A. Chandra, "Starling: Minimizing communication overhead in virtualized computing platforms using decentralized affinity-aware migration," 39th International Conference on Parallel Processing (ICPP) Sep 2010, pp. 228-237, 2010.

[7] Mayank Mishra, Anwesha Das, Purushottam Kulkarni, and Anirudha Sahoo, "Dynamic resource management using virtual machine migrations", IEEE Communications Magazine, vol. 50, pp. 34-40, September 2012.

[8] R. Nathuji and K. Schwan , "Virtual power: Coordinated power management in virtualized enterprise systems," in Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP'07), 2007, pp. 265–278.

[9] T. Wood et al., "Black-Box and Gray-Box Strategies forVirtual Machine Migration," Proc. 4th Conf. Symp.Networked Sys. Design & Implementation, 2007.

[10] U. Deshpande, X. Wang, and K. Gopalan, "Live Gang Migration of Virtual Machines," High-Performance Parallel and Distributed Computing, June 2011.

[11] M. Mishra and A. Sahoo, "On Theory of VM Placement: Anomalies in Existing Methodologies and Their Mitigation Using a Novel Vector Based Approach," Proc. 4thInt'l. Conf. Cloud Computing, 2011, pp. 275–82.

[12] R. M. Karp, M. Luby, and A. Marchetti-Spaccamela, "A Probabilistic Analysis of Multidimensional Bin Packing Problems," Proc.16th Annual ACM Symp. Theory of Computing, 1984, pp. 289–98.

[13] T. Wood et al., "Memory Buddies: Exploiting Page Sharing for Smart Colocation in Virtualized Data Centers," Proc. ACM SIGPLAN/SIGOPS Int'l. Conf. Virtual Execution Environments, VEE, 2009, pp. 31–40.

[14] T.Mahiba, Jayashree, "Live Virtual Machine Migration in Dynamic Resource Management of Virtualized Cloud Systems",International Journal of Latest Trends in Engineering and Technology,4th July 2013.

