Virtual Machine Migration Policies in Clouds

Amritpal Singh[1], Supriya Kinger[2]

[1]M.Tech Research Scholar, Department of Computer Science Engineering,
Sri Guru Granth Sahib World University, Fatehgarh Sahib, India

[2]Assistant Professor, Department of Computer Science Engineering,
Sri Guru Granth Sahib World University, Fatehgarh Sahib, India

Abstract: ***The virtual machine (VM) migration brings multiple benefits such as higher performance, improved manageability and fault tolerance. Failures are often common in systems comprising thousands of processors and often we encounter these failures using a well defined Fault Tolerance Mechanism (FTM).This technique is enabled using a Virtual Machine (VM) migration policy. Managing the placement and migration of VMs in a data centre is a significant challenge. This paper presents an overview of Cloud and its Services, then Fault Tolerance is discussed in detail, giving review of VM Migration policies.***

Keywords: Cloud Computing, Virtualization, VM Migration, Fault Tolerance, Migration Policies.

1. Introduction

The term *Cloud Computing* implies that you receive IT processing as a service rather than as a product or software. The simple method to visualize this is to compare to electricity: local computing is comparable to everyone owing a mechanical generator to produce their own electricity [1]. Cloud computing is about centralizing the computing activity, similar to producing electricity in power plants and distributing it via grids. Cloud computing is broken down into three segments: "Applications"," Platforms", and "Infrastructure". Each segment serves a different product for businesses and individuals around the world [1].A central server administers the system, monitoring traffic and client demands to ensure everything runs smoothly. Cloud computing is an emerging computing technology that uses the internet and central remote servers to maintain data and applications [1]. This technology allows for much more efficient computing by centralizing storage, memory, processing and bandwidth. Cloud computing is an outgrowth of the ease of access to remote computing sites provided by the internet. Cloud computing relies on sharing of resources to achieve coherence and economies of scale similar to a utility (like the electricity grid) over a network [1], [2].

1.1 Cloud Services

1.1.1Software as a service (SaaS): SaaS is a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network, typically the internet [1][2].

1.1.2Platform as service (PaaS): Platform as a service encapsulates a layer of software and provides it as service that can be used to build higher level services. There are at least two perspectives on PaaS depending on the perspective of the producer or consumer of the services.

1.1.3Infrastructure as a service (IaaS): IaaS delivers basic storage and compute capabilities as standardized services over the network. Servers, storage systems, switches, routers and other systems are pooled and made available to handle workloads that range from application components to high performance computing applications.

1.2 Cloud computing Models

1.2.1Public Clouds: Public clouds are run by third parties, and applications from different customers are likely to be mixed together on the cloud's servers, storage systems, and networks. Public clouds are most often hosted away from customer premises.

1.2.2Private Clouds: Private clouds are built for the exclusive use of one client, providing the utmost control over data, security, and quality of service.

1.2.3Hybrid Clouds: Hybrid clouds combine both public and private cloud models. They can help to provide on-demand, externally provisional scale.

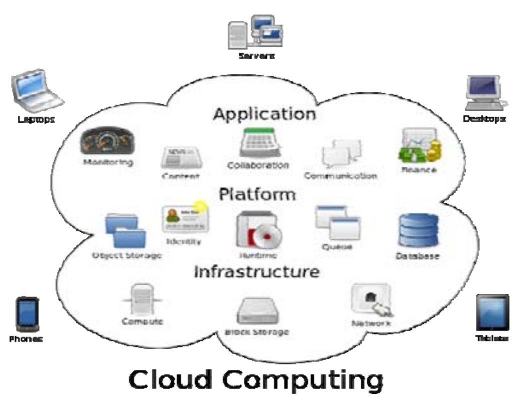

Figure 1: Cloud Computing at a Glance

2. Virtualization

Virtualization addresses IT's most pressing challenge: the infrastructure sprawl that compels IT departments to channel 70% of their budget into maintenance, leaving scant resources for business-building innovation [3]. The difficulty stems from the architecture of today's X86 computers: they're designed to run just one operating system and application at a time. As a result, even small datacenters have to deploy many servers, each operating at just 5% to 15% of capacity—highly inefficient by any standard.

Virtualization software solves the problem by enabling several operating systems and applications to run on one physical server or “host.” Each self-contained “virtual machine” is isolated from the others, and uses as much of the host’s computing resources as it requires [3].

2.1 Virtualization Benefits

Virtualization is the single most effective way to reduce IT expenses while boosting efficiency and agility [3]. The benefits include:

2.1.1 Run multiple operating systems and applications on a single computer.
2.1.2 Consolidate hardware to get vastly higher productivity from fewer servers.
2.1.3 Save 50% or more on overall IT costs.
2.1.4 Speed and simplify IT management, maintenance, and the deployment of new applications [3].

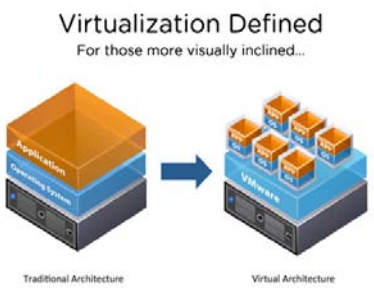

Figure 2: Virtualization

3. The Need for Fault Tolerance Strategy

Failures are likely to be more frequent in systems with thousands of processors. Therefore, schemes for dealing with faults become increasingly important [4]. Cloud computing has emerged as a global – infrastructure for applications by providing large scale services through cloud servers. The services can be either storage service or computation service. These services can be configured dynamically by making use of virtualization. Any application in cloud computing environment can be represented by a workflow. However, this computing environment still cannot deliver the quality, robustness and reliability that are needed for the execution of various workflows because of different failures like link failure, failure of server providing the service, malicious code in the executing node, datasets required by the task may be locked by other tasks etc. The scheduling system in a cloud should overcome such failures which can be provided by fault tolerant scheduling algorithms.

Fault tolerance is achieved by compromising task replication and task resubmission methods. Tasks are replicated based on heuristic metric and priority of the task. Heuristic metric is calculated by finding the tradeoff between replication and resubmission factor which gives the replication number based on impact of the resubmission. The heuristic metric is considered because replication alone may lead to resource wastage and resubmission alone may increase make span. Tasks are prioritized based on out degree, earliest deadline and high resubmission impact. Tasks are scheduled to meet deadline by considering priority and thereby to reduce wastage of resources by restricting unnecessary replication of tasks [5]. Fault tolerant scheduling algorithms can be categorized based on check pointing, traces of data, replication of tasks and resubmission of tasks. Each category has its own advantages and disadvantages [5].

3.1 Benefits of Fault Tolerance

In order to provide an effective control scheme with parameter guidance for cloud resource services, failure detection is essential to meet users' service expectations. It can resolve possible performance bottlenecks in providing the virtual service for the cloud computing networks. Fault-tolerance is one means to attain reliability, and is typically implemented by using some form of diversity.

Load balancing or Fault tolerance can be got with the migration of virtual machines, which can improve the performance of the datacenter. The service fault-tolerance scheme is based on deployment of virtual machines and service fault-tolerance scheme based on virtual machine migration. The effectiveness of the fault-tolerance scheme is proven with theoretical analysis, and the stability of the scheme is confirmed through experimental simulation.

4. Related Work on VM migration Policy

4.1 Elmroth et al. [6] 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.

4.2 Beloglazov et al. [7] 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.

4.3 Beloglazov et al. [8] 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 trade-off.

4.4 Voorsluys et al. [9] have formulated we present a 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.

4.5 Sonnek et al. [10] 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.

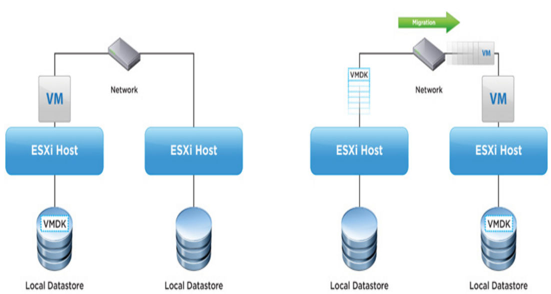

Figure 3: Live Migration of Virtual Machines

5. Fault Tolerance Based VM Migration Policy & Consolidation of VMs

In this section we propose several heuristics for dynamic consolidation of VMs based on an analysis of historical data of the resource usage by VMs.

5.1 Dynamic Consolidation Issue:

We split the problem of dynamic VM consolidation into four parts:

5.1.1Determining when a host is considered as being overloaded requiring migration of one or more VMs from this host.

5.1.2 Determining when a host is considered as being under loaded leading to a decision to migrate all VMs from this host and switch the host to the sleep mode.

5.1.3 Selection of VMs that should be migrated from an overloaded host.

5.1.4 Finding a new placement of the VMs selected for migration from the overloaded and under loaded hosts [11].

5.2 Brief Analysis of Policies for VM migration.

5.2.1 The Minimum Migration Time Policy: The Minimum Migration Time (MMT) policy migrates a VM v that requires the minimum time to complete a migration relatively to the other VMs allocated to the host. The migration time is estimated as the amount of RAM utilized by the VM divided by the spare network bandwidth available for the host j.

5.2.2 The Random Choice Policy: The Random Choice (RC) policy selects a VM to be migrated according to a uniformly distributed discrete random variable whose values index a set of VMs allocated to a host.

5.2.3 The Maximum Correlation Policy: The Maximum Correlation (MC) policy is based on the idea that the higher the correlation between the resource usage by applications running on an oversubscribed server, the higher the probability of the server overloading. According to this idea, we select those VMs to be migrated that have the highest correlation of the CPU utilization with other VMs. To estimate the correlation between CPU utilizations by VMs, we apply the multiple correlation coefficient. It is used in multiple regression analysis to assess the quality of the prediction of the dependent variable. The multiple correlation coefficient corresponds to the squared correlation between the predicted and the actual values of the dependent variable. It can also be interpreted as the proportion of the variance of the dependent variable explained by the independent variables [11]. In general; figure 3 shows the live migration of virtual machines [12].

6. Conclusion

The Provisioning of host to the virtual machines in a datacenter, needs a careful thought process, first it must address how the host must be reserved, and then, If those reservations comes under challenge under over-subscription, over utilization and performance degradation. How VM allocation policy must change to address these issues to that datacenters take right decisions to re-allocate or migrate VM machines. We can mitigate such issues by designing algorithms that can detect over-utilization of resources and predict any adversity in datacenters. Once this is done, We can either build a fall back VM allocation policy or in simple words, Fault Tolerance mechanism which may be active/proactive in nature.

Therefore, there is an urgent need to build VM allocation policy that works as fault tolerance algorithm to migrate and consolidate the VM dynamically and efficiently based on the non-linear nature of workflow jobs or tasks. We have conducted comparative analysis of the fault tolerance techniques and VM migration policies. As such, we believe that implementation of fault tolerance using VM migration policies in Cloud Computing is an area full of challenges.

References

[1] P. Mathur and N. Nischal, "Cloud Computing: New challenge to the entire computer industry," 1st International Conference on Parallel, Distributed and Grid Computing, pp. 223-228, 2010.

[2] K. Goyal and Supriya, "Security Concerns In the World of Cloud Computing", IJARCS International Journal of Advanced Research in Computer Science, Volume 4, No. 4, pp. 230-234, 2013.

[3] "Virtualization Basics and its Benefits," vmware.com. [Online]. Available: http://www.vmware.com/in/virtualization/virtualization-basics/what-is virtualization.html.[Accessed: April 21,2013].

[4] S. Chakravorty, C. Mendes and L. Kale, "Proactive Fault Tolerance in MPI Applications via Task Migration," IEEE International Conference on High Performance Computing (HiPC), pp. 1-12, 2006.

[5] J. K, J. Nirmala and M. Bhanu S, "Fault Tolerance workflow scheduling based on replication and resubmission of tasks in Cloud Computing," International Journal on Computer Science & Engineering, Jun2012, Vol. 4 Issue 6, pp. 996-1006, 2012.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Beloglazov and R. Buyya," Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centres", Wiley InterScience 2012, pp. 1-24, 2010.

[12] VMware vsphere "Live Migration of Virtual Machines," vmware.com. [Online]. Available: http://www.vmware.com/products/datacenter-virtualization/vsphere/vmotion.html.[Accessed: April 25, 2013].

Author Profile

Amritpal Singh received his bachelor's degree in Information Technology from Institute of Engineering & Technology (IET), Bhaddal, India in 2012.He started his master study in the Computer Science Engineering in 2011 from Sri Guru Granth Sahib World University, Fatehgarh Sahib, India. Currently, he is at the last semester and completing his research work.

Supriya Kinger is pursuing her doctoral program from Thapar University, Patiala, India. Currently, She is an Assistant professor in department of computer science engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, India.