A Review on Load Balancing of Virtual Machine Resources in Cloud Computing

Pradeep Kumar Tiwari and Sandeep Joshi

Abstract An effective load balance (LB) management achieves high performance computing (HPC) and green computing. Users can run their jobs on virtual machines (VMs). Virtual machine (VM) has own resources (CPU and memory). VM migrates from host to another host during fail of VM, hot spot and high resource demand. Effective LB management is based on scheduling policy and management Strategies. In this paper it is discussed the available scheduling mechanisms, goals and strategies of load balancing techniques. The aim of this work to elaborate the key analysis of research works on LB.

Keywords Virtual machine · Physical machine · Load balancing

1 Introduction

Cluster, grid and cloud computing using the fundamental concept of distributed system to achieve HPC (High Performance Computing). Distributed paradigms depend on distributed application. Applications and operating system can run separately on VMs. The core concept of resource pool and management is virtualization. Hardware virtualization (CPU partitioning and memory) can be achieved by Hypervisor. It is divided into Type 1 (hosted hypervisor) and Type 2 (bare-metal hypervisor). Type 1 as explore in Fig. 1, hypervisor is directly installed on the ×86 based hardware and provide direct access to the hardware resources. Type 2 explored in Fig. 2, hypervisor is installed and run as an application on top of an operating system and it is run on an operating system. Type 1 hypervisor more efficient rather than type 2. VM has own operating system and applications and they do not interfere with each others. Resource

P.K. Tiwari (✉) · S. Joshi
Department of Computer Science and Engineering, Manipal University Jaipur,
Jaipur, India
e-mail: pradeeptiwari.mca@gmail.com

S. Joshi
e-mail: sandeep.joshi@jaipur.manipal.edu

S.C. Satapathy and S. Das (eds.), *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2*,
Smart Innovation, Systems and Technologies 51, DOI 10.1007/978-3-319-30927-9_36

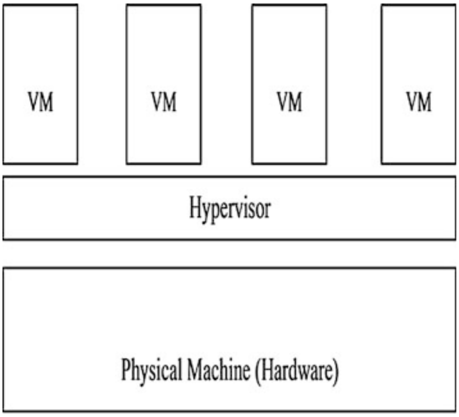

Fig. 1 Type 1 hypervisor

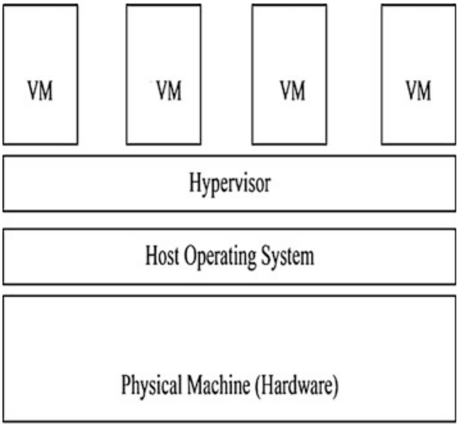

Fig. 2 Type 2 hypervisor

distribution among VMs not affected by VM failure and VM will be migrated without downtime. The LB policy should manage quality of service (QoS) and service level agreement (SLA). Week management of resources, load imbalance, hot spot and scattered information among heterogeneous servers are the main cause of SLA violation. The load imbalance problem occurs during frequently changing demand in heterogeneous environments. Load balance between low load to high load machine can be manages load imbalance. LB is not easy to manage on frequently changing high resource demand. Transfer the load, selection of VM, Location to migrate VM and information policies are responsible to manage load balance [1].

Researchers proposed numerous mechanisms to manage LB but still have many recommendations. Network, Compute and memory management in multitenant environments with scattered servers not easy to manage. This review is based on process (CPU) and memory management policies and its effect on load management. Available approaches are static, dynamic and dynamic consolidation. The modern load balancing mechanism can be automated the load. Recent mechanism provides high performance, energy saving, dynamic load requirement and minimize the cost. Effective dynamic load management during frequently changing environment and high resource demand does the vital role for resource management. VMs can have only available resources on PM. Resource demands from VMs grater then available resource on PM is called a hot spot. The opposite situation, when resources are underutilized of PM is termed as cold spot. Hotspot and cold spot can be managed by moving VMs. Hop spot mitigation selected VM and moved to less loaded PM [2].

2 Live Migration of Virtual Machines

Management of live migration technique must follow these two policies (a) Energy efficient migration technique and (b) Load balanced with fault tolerance technique.

2.1 Energy Efficient Migration Technique

Servers usually consume 70 % of total access energy. Live migration technique must follow the energy efficient mechanism. If VMs loads in an ideal situation and no need to use of others VMs in this situation remain VMs should go into sleep (energy saving) mode [1].

2.2 Load Balance with Fault Tolerance Technique

VMs loads should be scalable and monitored during load transfer and when resources are being utilized. VMs must be able to take loads of crashed VMs or not working VMs. Fault tolerance should gives assurance to clients, if any machine or system will be crashed then it must manage the clients' jobs according to SLA. Fault tolerance improves efficiency of VMs and makes an effective, robust mechanism during crashes of VMs [1, 2].

3 Resource Management Goals and Management Strategies

Resource management is a semantic relationship between resource availability and resource distribution. Network, compute and storage are a main resource component. Resource manager manages the resources according the availability of resources and provide to end user as per of SLA agreement. Resource management problems include allocation, provisioning, requirement mapping, adaptation,

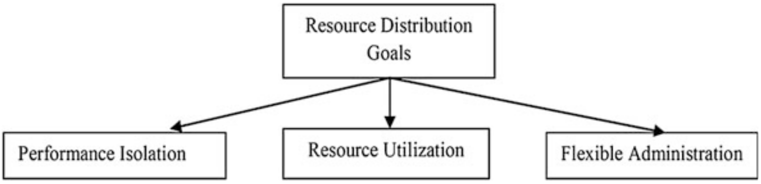

Fig. 3 Goals of resource management

discovery, brokering, estimation, and modeling. The main goals as shown in the Fig. 3 of resources distribution are (a) Performance Isolation, (b) Resource Utilization and (c) Flexible Administration.

Performance Isolation: VMs are isolated from each other and do not affect to another VM capacity. Failure of VM does not affect the performance and load. The load will migrate to another VM. Hyper-V provide the facility of quick migration and VMware provide the facility of live migration. The Load should migrate VM to VM or VM migrate to another physical server due to occurrences of high resource demand or failure of VM [3, 4].

Resource Utilization: Dynamic resource allocation management does the effect on resource utilization and provides resources from PM. Resource utilization is based on a minimum consumption of available resources without down time. Effective resource management managers maximize the resource availability and minimize the energy consumption. Resource manager must give the attention to SLA based on demand resource requirements. Load Manager mapped and measured load of individual VM. VMs resource utilization can be map with PM resources and manager can utilize unused resources of PM [5].

Flexible Administration: Resource availability administrator must able to manage high load resource demands and utilize in synchronized manner. VMware uses distributed resource scheduler (DRS) to manage VMs capacity (resource reservations, priorities and limit) and VM migration. VMware, distributed power management (DPM) system manages power on or off management of used and not used VM to achieve energy efficient resource management [6]. Microsoft Hyper-V uses system center virtual machine manager (SCVMM) support system to manage virtual machine manager 2008 R2. SCVMM increases the flexibility in storage management and migration management of VMs among hosts.

4 Load Balancing Approaches

4.1 Static Consolidation

Defines pre-reserved dedicated resource allocation to VM according the need of end users. VMs resource allocation based on total capacity of PM and migration does not perform till all demands are changing. Optimal and sub optimal scheduling belongs to static scheduling. Optimal scheduling has the information about the job and resources. Job scheduling and resource allocation decision can be taken on feasible time. If any problem occurs during feasible job scheduling then suboptimal manage [7].

4.2 Dynamic Consolidation

This is periodically, current demand based VM migration approach. If required VMs resource demands are higher from physical available capacity than VM migrate to another PM. Dynamic scheduling is based on distributed and centralized policy. Distributed approach handle the scheduling and rescheduling of a job. The dynamic centralized approach takes the decision by centralized resource management manager [7].

4.3 Dynamic Consolidation with Migration Control

This approach gives the stability during high resource demand, hotspot and frequently change resource demands. This approach reduces the required number of physical servers and save the consumed energy. This approach based on heuristic and round robin mechanism.

5 Analysis of Load Balancing Mechanism

Researchers proposed a different mechanism to manage load balance. Many researchers approach ants, agent, genetic, heuristic honey bee, round robin, token routing, sender initiative, receiver initiative, max min, min min and many more static and dynamic approaches. Ants deal with to achieve a diversity of complicated tasks with consistency and reliability. In spite of the fact that this is generally self association as opposed to learning, ants need to adapt to a phenomenon that looks all that much like over preparing in learning methods [8].

5.1 Agent Based

Aarti et al. [9], proposed an autonomous agent based load balancing algorithm (A2LB) which provides dynamic load balancing for cloud environments. The proposed mechanism has been implemented and found to provide acceptable outcome.

Omar et al. [10], proposed the most ideal approaches to consequently adjust the load among machines in large scale circumstances. The proposed mechanism is based on the performance of two dissimilar applications with two diverse conveyance approaches, and experimental results demonstrates that few applications can consequently adjust the load among the machines and get alone a superior performance in large scale simulations with one distribution approach than the other.

5.2 Genetic Algorithm

Joseph et al. [11], proposed genetic algorithm based technique to allocate VMs using the Family Gene approach. Experimental results show that the proposed mechanism minimize energy consumption and the migrations. A user specifies the requirement of resources and service to a provider and makes a contract with the service provider. This user requirement is called the SLA. A cloud service provider ought to be to manage a superior level of SLA. The proposed mechanism is able to minimize energy consumption and the VM migrations. A proposed mechanism increases the SLA level, while keeping the number of active hosts at a minimal level.

Hu et al. [12] also proposed genetic algorithm based scheduling mechanism for load balancing among VMs. This mechanism selects the least loaded virtual machine for load transfer and optimizes the high migration cost. However, due to a large number of virtual machines and frequent service requests in the data center, there is chance of inefficient service scheduling.

5.3 Heuristic Approach

Ferreto et al. [13] proposed LP-Heuristic linear programming (LP) based heuristic worst fit decreasing (WFD), best fit decreasing (BFD), first fit decreasing (FFD) and AWFD almost worst fit decreasing (AWFD). The proposed mechanism is two ways resource management approach. First approach identifies VMs and maps the capacity from existing physical machine capacity and another approach short physical machine increasingly according to their capacities. LP goal is minimization of required physical machine and map VMs resource availability form hosted physical machine.

Beloglazov et al. [14] proposed energy aware data center location approach modified best fit decreasing (MBFD) and minimization of migration (MM) approach. MBFD optimize the current VM allocation and choose the most energy efficient nearest physical servers to migrate VM and MM approach minimize the VM migration needs.

Table 1 Load balancing mechanism analysis

Author(s)	Technique	Strength	Scheme	Recommendations
Andreolini et al. [19]	Dynamic load management of virtual machines	Robust and selective reallocations	Performance based	Heterogeneous infrastructures and platforms
Beloglazov et al. [14]	Dynamic consolidation of virtual machines	SLA based load management	Energy efficient	Researchers can focus on multi-core CPU architectures
Ferreto et al. [13]	LP formulation and heuristic	Server consolidation with migration control	Energy efficient	Migration control without downtime
Forsman et al. [20]	Push and pull strategy	Rebalance the load when VMs added and removed	Load management	Downtime can be less
Jin et al.	Pre-copy model	Maximize the CPU utilization, Reduce the downtime from previous approaches	Live migration	Add network bandwidth controlling and memory writing pattern to optimize current pre-copy algorithm
Joseph et al. [11]	Genetic approach	SLA based resource management, maximize the hardware resource utilization with performance	Resource management	Modify the algorithm to decrease the calculation time in terms of prediction process to improve the genetic algorithm convergence speed
Lau et al. [15]	Guarantee reservation (GR) protocol and task batch composition (TBC) scheme	Maximize the processing speed of sender and receiver	Demand based load balanced	Performance on real time system
Li et al. [18]	Dynamic VM placement	Minimizing the total completion time	Performance based	Hybrid scheme of integrating off-line placement into online scenario
Zhang et al. [17]	Dynamically CP and heuristic allocation of cloud data center’s resources	Maximize the resource utilization from) first-fit and best-fit	Performance based	(i) Refining the model to account for the energy consumption for providing data center services (ii) Dynamically determine the reservation ratio and the duration threshold for long jobs

5.4 Dynamic Approach

Lau et al. [15] proposed adaptive load distribution algorithm. This research shows previous result analysis of sender and receiver initiative algorithm and manages the work in heterogeneous load. The proposed mechanism is able to handle adaptive load distribution algorithms for heterogeneous distributed systems.

Beloglazov et al. [16] proposed a novel technique for dynamic consolidation of VMs based on adaptive utilization thresholds. The proposed technique validates the high efficiency different kinds of workloads.

Zhang et al. [17] proposed an optimization based approach that manages long jobs to dynamically allocate a cloud data center's resources. This mechanism can achieve considerably better utilization by increasing the number of jobs. The authors use a constraint programming (CP) solution to schedule the long jobs, and use simple heuristics mechanism schedule the short jobs. The proposed mechanism is able to increase the number of jobs accommodated using dynamic scheduling by 18 %. It also compares the performance of CPU and memory.

Li et al. [18] proposed an off-line VM placement method through an emulated VM migration process, while the on-line VM placement is solved by a real VM migration process. The migration algorithm is a heuristic approach. This approach uses place the VM placement to its best PM directly, if this PM has enough capacity. Otherwise, it migrate another VM from this PM to accommodate the new VM. Outcomes results validate the high efficiency of algorithms.

Andreolini et al. [19] Authors proposed reallocations of VMs management algorithms in a large number of hosts. The novel algorithms identify the real critical instances and take decisions without recurring to typical thresholds. Experimental results show that proposed algorithms are truly selective and robust even in variable contexts, thus reducing system instability and limit migrations when really necessary. Table 1, show the analysis of researchers work area and his/her future

Table 2 Research Support analysis

Organization	Support system	Scheme	Support model
Eucalyptus	Linux based frame work	Open source	Execution control of VM in heterogeneous environment
GENI (Global Environment for Network Innovations)	Apache HTTP server	Free of charge for research and classroom use	Experimental heterogeneous network structure, distributed system and security
Google App Engine	Execution of web application, Java and Python	Freeware platform	Experimental heterogeneous network structure, distributed system and security
Grids Lab Aneka	.NET-based framework	On-demand	Multiple application models, persistence, and security solutions
Open Stack	Web-based dashboard	Open source	Large network of VM, storage system, resource management
Sun Network.com (Sun Grid)	C, C++ and FORTRAN based application	Open source	Job management

recommendation. Table 2, is based on a research support system and research model for load management.

6 Conclusion

In this paper, have discussed the Strategies, goals, and polices of load balance. Effective load management policies can work in heterogeneous workload with maximum utilization of the CPU. The resources are determined by, which VM will utilize and where the target host to migrate the VM. Energy efficient migration technique and load balance with fault tolerance technique make green computing. Performance isolation, resource utilization and flexible administration are mail goal to manage load balance. We have discussed many policies to manage resources and VM migration. Researchers can do the work on the CPU utilization and VM migration in minimum downtime. The recommendations for researchers are, they can do work on optimal resource management to improve performance, scalability of resources with the future prediction of resources need, and minimize the migration of VMs.

References

1. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. **1**(1), 7–18 (2010)
2. Vinothina, V., Sridaran, R., Ganapathi, P.: A survey on resource allocation strategies in cloud computing. Int. J. Adv. Comput. Sci. Appl. **3**(6) (2012)
3. Gupta, D., Cherkasova, L., Gardner, R., Vahdat, A.: Enforcing performance isolation across virtual machines in Xen. In: Middleware, pp. 342–362. Springer, Berlin (2006)
4. Nathan, S., Kulkarni, P., Bellur, U.: Resource availability based performance benchmarking of virtual machine migrations. In: Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering, pp. 387–398. ACM (2013)
5. Isci, C., Liu, J., Abali, B., Kephart, J.O., Kouloheris, J.: Improving server utilization using fast virtual machine migration. IBM J. Res. Dev. **55**(6), 1–4 (2011)
6. Resource management policy. Retrieved from https://pubs.vmware.com/vsphere-50/index.jsp#com.vmware.vsphere.vm_admin.doc_50/GUID-E19DA34B-B227-44EEB1AB-46B826459442.html, July 2015
7. Rathore, N., Chana, I.S.: Load balancing and job migration techniques in grid: a survey of recent trends. Wireless Pers. Commun. **79**(3), 2089–2125 (2014)
8. Mishra, R., Jaiswal, A.: Ant colony optimization: a solution of load balancing in cloud. Int. J. Web Semant. Technol. (IJWesT) **3**(2), 33–50 (2012)
9. Singh, A., Juneja, D., Malhotra, M.: Autonomous agent based load balancing algorithm in cloud computing. Proc. Comput. Sci. **45**, 832–841 (2015)
10. Rihawi, O., Secq, Y., Mathieu, P.: Load-balancing for large scale situated agent-based simulations. Proc. Comput. Sci. **51**, 90–99 (2015)
11. Joseph, C.T., Chandrasekaran, K., Cyriac, R.: A novel family genetic approach for virtual machine allocation. Proc. Comput. Sci. **46**, 558–565 (2015)

12. Hu, J., Gu, J., Sun, G., Zhao, T.: A scheduling strategy on load balancing of virtual machine resources in cloud computing environment. In: Proceedings. PAAP, pp. 89–96 (2010)
13. Ferreto, T.C., Netto, M.A.S., Calheiros, R.N., De Rose, C.A.: Server consolidation with migration control for virtualized data centers. Future Gener. Comput. Syst. **27**(8),1027–1034 (2011)
14. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Gener. Comput. Syst. **28**(5), 755–768 (2012)
15. Lau, S.M., Lu, Q., Leung, K.S.: Adaptive load distribution algorithms for heterogeneous distributed systems with multiple task classes. J. Parallel Distrib. Comput. **66**(2),163–180 (2006)
16. Beloglazov, A., Buyya, R.: Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers. In: Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science, vol. 4. ACM (2010)
17. Zhang, Y., Fu, X., Ramakrishnan, K.K.: Fine-grained multi-resource scheduling in cloud datacenters. In: 2014 IEEE 20th International Workshop on Local and Metropolitan Area Networks (LANMAN), pp. 1–6. IEEE (2014)
18. Li, K., Zheng, H., Wu, J., Du, X.: Virtual machine placement in cloud systems through migration process. Int. J. Parallel Emergent Distrib. Syst. 1–18 (ahead-of-print, 2014)
19. Andreolini, M., Casolari, S., Colajanni, M., Messori, M.: Dynamic load management of virtual machines in cloud architectures. In: Cloud Computing, pp. 201–214. Springer, Berlin (2010)
20. Forsman, M., Glad, A., Lundberg, L., Ilie, D.: Algorithms for automated live migration of virtual machines. J. Syst. Softw. **101**, 110–126 (2015)