ISSN:0975-9646

A Survey on Resource Management in Cloud Computing

Sanchari Saha
(Assistant Professor)
Department of CSE, MVJCE, Bangalore

Abhilash K.V
(PG-Student)
Department of CSE, MVJCE, Bangalore

Abstract-A cloud computing infrastructure is a complex system with a large number of shared resources. These are subject to unpredictable requests and can be affected by external events beyond your control. Cloud resource management requires complex policies and decisions for multi-objective optimization. It is extremely challenging because of the complexity of the system, which makes it impossible to have accurate global state information. It is also subject to incessant and unpredictable interactions with the environment Resource management is a core function required of any man-made system. It affects the three basic criteria for system evaluation are performance, functionality and cost. Inefficient resource management has a direct negative effect on performance and cost. It can also indirectly affect system functionality. Some functions the system provides might become too expensive or ineffective due to poor performance. This paper survey on resource managements in cloud computing.

Keywords-**cloud computing, resource management**

I.CLOUD COMPUTING

Cloud computing, or the cloud, is a colloquial expression used to describe a variety of different types of computing concepts that involve a large number of computers connected through a real-time communication network such as the Internet. Cloud computing is a term without a commonly accepted unequivocal scientific or technical definition. In science, cloud computing is a synonym for distributed computing over a network and means the ability to run a program on many connected computers at the same time. The phrase is also, more commonly used to refer to network-based services which appear to be provided by real server hardware, which in fact are served up by virtual hardware, simulated by software running on one or more real machines. Such virtual servers do not physically exist and can therefore be moved around and scaled up (or down) on the fly without affecting the end user arguably, rather like a cloud. The popularity of the term can be attributed to its use in marketing to sell hosted services in the sense of application service provisioning that run client server software on a remote location.

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. At the foundation of cloud computing is the broader concept of converged infrastructure and shared services. The cloud also focuses on maximizing the effectiveness of the shared resources. Cloud resources are usually not only shared by multiple users but are also dynamically re-allocated per demand. This can work for allocating resources to users. This approach should maximize the use of computing powers thus reducing environmental damage as well since less power, air conditioning, rackspace, etc. is required for a variety of functions. Proponents claim that cloud computing allows companies to avoid upfront infrastructure costs, and focus on projects that differentiate their businesses instead of infrastructure. Proponents also claim that cloud computing allows enterprises to get their applications up and running faster, with improved manageability and less maintenance, and enables IT to more rapidly adjust resources to meet fluctuating and unpredictable business demand.

Resource Management is an important issue in cloud environment. Cloud computing is the delivery of computing and storage capacity as a service to a community of end-recipients. The name comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts services with a user's data, software and computation over a network.

II. RESOURCE MANAGEMENT IN LARGE CLOUD ENVIRONMENTS

In the context of large-scale distributed computing, a key problem in resource management is that of mapping a set of applications onto a system of machines that execute those applications and, for each such machine, assigning local resources for those applications that run on it. The quality of the allocation process is often measured through a utility function, which is an aggregation function computed from local state variables. An optimal allocation maximizes such a system utility. The machine resources that are allocated to applications include CPU, memory, storage, network bandwidth, access to special hardware/software, etc. The resource demand of an application can change over time. In response to such changes, the resource allocation process needs to be repeated many times over; in other words, it has to be dynamic, in order to ensure that the system utility is maximized at all times. Optimal resource allocation in the sense of utility maximization is often computationally expensive. In the context of grid computing, for instance, the problem of scheduling jobs onto machines such that the total execution time is minimized can be formulated as the minimum make span scheduling problem, which is known to be NP-hard. Second, in the context of cloud computing, the problem of placing applications onto machines is often modeled as a variant of the knapsack problem, which is also known to be NP-hard. In this work, we model the managed system as a dynamic set of nodes that represents the machines of a cloud environment. Over time, nodes may join or leave the system, or they may fail.

III.PROCESS DELAY IN RESOURCE ALLOCATION

Process delay is the time between allocating resources and accurate measuring the effect of the resource allocation on application QoS. In ne-grained cloud management, VM resource allocation relies on precise operations that set resources to desired values assuming the observation of instant reconfiguration effect or process delays would affect the effectiveness of the management. By setting the management interval to 30 seconds, the authors in observed that under sustained resource demands, a VM needs minutes to get its performance stabilize after memory reconguration. Similar delayed can also be observed in CPU reconguration, partially due to the backlog of requests in prior intervals. The difficulty in evaluating the immediate output of resource allocations makes the modeling of application performance even harder.

Issues in Large Scale Cloud Management

In a cloud, hosted applications such as multi-tier websites and parallel computingprograms may run on a group of VMs that span multiple physical hosts. These VMs form a resource pool. The resource management of these applications requires that the resource pool should obtain sufficient resources and these resources are properly distributed to each VM. These multi-VM applications usually involve synchronous multi-stage execution in which one stage is blocked until the completion of previous stages. As a result, the performance of the application needs the coordination of all the physical machines that host the virtual cluster. Due to initial placement and load balancing, the actual deployment of these VMs can show an arbitrary topology on physical nodes. As the numbers of physical hosts and VMs increase, the cloud infrastructure is divided into several sub-clusters, each of which is responsible for the resource allocation of one application. These sub-clusters may or may not overlap with each other and the topology can change over time. In such a large scale cloud environment, no centralized management is practical.

In this dissertation, we aim to design, implement and evaluate a resource management mechanism that delivers stable and adaptive control over cloud resources. In the face of dynamic application demands, the management scheme should leverage cloud elasticity, transparently adding or removing virtualized resources at one grain to match the workloads. Resource allocations are in response to the observation of changes in application level metrics. These metrics include but not limited to performance metrics (e.g. response time and throughput), expenditure metrics (e.g. dollars per hour) and energy consumption metrics (e.g. Joule per hour). Overall, there are two main requirements in the design of automatic resource management are transparency and assurance. Transparency requires the management to perform automated resource adjustments during the life time of cloud applications without cloud users' intervention. Transparency implies that the resource management should be able to translate application SLOs to resource requirements. It also requires the management to be adaptive to workload and cloud dynamics. Assurance refers to the guarantee of SLOs in the presence of background management operations. It requires that the management should be responsive to SLO violations and be stable during oscillations. The design of automatic resource management should address the following challenges:

(1) It should dene a metric that synthesizes multiple application-level metrics and measures a VM's capacity. (2) It should be able to deal with multiple resources. (3) It should employ modelfree approaches to handle complex resource to performance relationship. (4) It should be able to provide accurate resource allocation in the presence of process delays. (5) It should scale well assuming no information on actual VM deployment.

IV.CLUSTER WIDE CLOUD RESOURCE MANAGEMENT

We present a distributed reinforcement learning approach to the cluster wide cloud resource management. We decompose the cluster wide resource allocation problem into sub problems concerning individual VM resource configurations. The cluster wide allocation is optimized if individual VMs meet their SLA with a high resource utilization. For scalability, we develop an efficient reinforcement learning approach with continuous state space. For adaptability, we use VM low level runtime statistics to accommodate workload dynamics.

V.SINGLE RESOURCE MANAGEMENT

This approach assumes non-work-conserving CPU mode and no interference between co-hosted VMs, which can lead to resource under provisioning. Recent work enhanced traditional control theory with Kalman filters for stability and adaptability. But the work remains under the assumption of CPU allocation. The authors in applied domain knowledge guided regression analysis for CPU allocation in database servers. The method is hardly applicable to other applications in which domain knowledge is not available. The allocation of memory is more challenging. The work in dynamically controlled the VM's memory allocation based on memory utilization. Their approach is application specific, in which the Apache web server optimizes its memory usage by freeing unused http processes. For other applications like MySQL database, the program tends to cache data aggressively. The calculation of the memory utilization for VMs hosting these applications is much more difficult. Xen employs Self-Ballooning to do dynamic memory allocation. It estimates the VM's memory requirement based on OS-reported metric: Commited_AS. It is effective expanding a VM under memory pressures, but not being able to shrink the memory appropriately. More accurate estimation of the actively used memory (i.e. the working set size) can be obtained by either monitoring the disk I/O or tracking the memory miss curves. However, these event driven updates of memory information cannot promptly shrink the memory size during memory idleness.

VI.MULTIPLE RESOURCES MANAGEMENT

Automatic allocation of multiple resources or for multiple objectives poses challenges in the design of the management scheme. Complicated relationship between resource and performance and often contradicted objectives prevent many work from being automatic but heuristic. They used a MIMO controller to automatically allocate CPU share and I/O bandwidth to multiple VMs. However, the ARMA model may not be effective under steady workload because the recursive least square (RLS) method is effective only when there is enough steepness between two consecutive measurements. The authors also rely on the assumption that drastic variations in workloads that cause significant changes in the model parameters are rare, which limits the applicability of this approach to wider range of platforms. Most importantly, the cost function which directs the resource allocations does not emphasize on the release of unused resources.

VI.CONCLUSION

This paper suruvey on resource managements in cloud computing.it is concludes that if the demand for the resource is below its capacity, then it is under-utilized and if the demand is above its capacity it is overutilized or unable to meet the demand.the problem can overcome using the resource can be integrated with a reactive cloud architecture capable of automatically scaling it horizontally or vertically in response to fluctuating demand. An elastic provisioning system is established to dynamically allocate and reclaim cpus and ram for a virtual server in response to the fluctuating processing requirements of its hosted resources

REFERENCES

[1]. The Basics of Cloud Computing Alexa Huth and James Cebula , © 2011 Carnegie Mellon University. Produced for US-CERT, a government organization.

[2]. Dynamic Resource Allocation using Virtual Machines for Cloud Computing Environment , Zhen Xiao, Senior Member, IEEE, Weijia Song, and Qi Chen

[3]. M. Armbrustet al., “Above the clouds: A berkeley view of cloud computing,” University of California, Berkeley, Tech. Rep., Feb 2009.

[4]. M. Nelson, B.-H. Lim, and G. Hutchins, “Fast transparent migration for virtual machines,” in Proc. of the USENIX Annual Technical Conference, 2005.