International Journal of Research
International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 07, July 2015
Available at http://internationaljournalofresearch.org

Effective Resource Allocation in Cloud Systems

SharathBabuCheggoju[1]&P.Darshan[2]
[1]M.Tech Student, Dept of CSE, ChilkurBalaji Institute OfTechnology,Rangareddy, T.S, India
[2]Associate Professor& HOD, Dept of CSE, ChilkurBalaji Institute Of Technology,Rangareddy, T.S, India

Abstract: *As the virtual machine (VM) technology being developed rapidly and services being used by multiple users, resources in cloud systems needs to be shared more effectively and allocated on demand. Cloud computingenvironment involves high cost infrastructure on one hand andneed high scale computational resources on the other hand.These resources need to be provisioned (allocation andscheduling) to the end users in most efficient manner so that thetremendous capabilities of cloud are utilized effectively andefficiently.But there are some constraints on cloud computing like payment budget and on demand resources.This paper formulates a deadline-driven resource allocation approach based on VM resource isolation technology, and also propose a novel solution with polynomial time, which could minimize users' payment in terms of their expected deadlines. By analyzing the upper bound of task execution length based on the possibly inaccurate workload prediction, further propose an error-tolerant method to guarantee task's completion within its deadline. Validate its effectiveness over a real VM-facilitated cluster environment under different levels of competition*

Keywords: virtual machine; Optimal Resource Allocation; Earliest Deadline First

1. Introduction

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the common use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation. Cloud computing consists of hardware and software resources made available on the Internet as managed third-party services. These services typically provide access to advanced software applications and high-end networks of server computers. Cloud computing platforms are rapidly growing as the most preferred option for hosting applications in many business contexts. Small and Medium sized companies are relying on public cloud infrastructures to implement their applications, to reducethe initial cost. Larger companies alsohosting their applications using either public clouds for expanding their existing infrastructures or rapid deployment of test environments, or private clouds for dynamic on-demand provisioning of virtual resources among their internal divisions. Traditional job scheduling is often formulated as a kind of combinatorial optimization problem or queue-based multiprocessor scheduling problem, due to the nonguaranteed performance isolation for multiple tasks running on the same machines. That is, most of the existing deadline-driven task scheduling solutions from single cluster environment confined in LAN to the Grid computing environment suitable for WAN are also strictly subject to the queuing model under which a single machine's multiple resources cannot be further split to smaller fractions at will. This will eventually cause the raw-grained resource allocation, relatively low resource utilization and suboptimal task execution efficiency. But all the resources provisioned by cloud system are supposed to be under a payment model. Each task's workload is likely of multiple dimensions. First, the compute resources in need may be multi-attribute (such as CPU, disk-reading speed, network bandwidth, etc.), resulting in multidimensional execution in nature. Second, even though a task just depends on one resource type like CPU, it may also be split to multiple sequential execution phases, each calling for a different computing ability and various price on demand, also leading to a potentially high-dimensional execution scenario..

International Journal of Research
International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 07, July 2015
Available at http://internationaljournalofresearch.org

2. Literature Survey

There are many solutions proposedby different people to achieve efficient resourceallocation scheme. Resource allocation and provisioning in cloud computingenvironment is done with the main aim of achieving load balancing.Based on various factors like spatial distribution of cloud nodes,algorithm complexity, storage/replication, point of failure etc.different techniques have evolved to provision the resources inbalanced manner. The provisioning is done taking into account whether the environment is static or dynamic.In the elasticity of a utility matches the need of businesses providing services directly to customers over the Internet, as workloads can grow (and shrink) far faster than 20 years ago.. From the cloud provider's view, the construction of very large datacenters at low cost sites using commodity computing, storage, and networking uncovered the possibility of selling those resources on a pay-as-you-go model below the costs of many medium-sized datacenters, while making a profit by statistically multiplexing among a large group of customers. From the cloud user's view, it would be as startling for a new software startup to build its own datacenter as it would for a hardware startup to build its own fabrication line Cloud Computing users, relieved of dealing with the twin dangers of over-provisioning and under-provisioning our internal datacenters.

The Clouds do not have a clear and complete definition in the literature yet, which is an important task that will help to determine the areas of research and explore new application domains for the usage of the Clouds. To tackle this problem, the main available definitions extracted from the literature have been analyzed to provide both an integrative and an essential Cloud definition. Although their encompassing definition is overlapped with many grid concepts Virtualization is the key enabler technology of Clouds, as it is the basis for features such as, on demand sharing of resources, security by isolation, etc. Usability is also an important property of Clouds. Also, security enhancements are needed so that enterprises could rely sensitive data on the Cloud infrastructure. Finally, QoS and SLA enforcement will also be essential before ICT companies reach high levels of confidence in the Cloud. NGG and OGF efforts are highly devoted to this task, enforcing standardization to enable a Cloud federation that can then deal with the required massive scalability.

There has been much prior work on task scheduling that considers resource requirements that address how much resource tasks consume. By addresses the performance impact of task placement constraints. Task placement constraints impact which resources tasks consume. Task placement constraints, such as characteristics specified by the Condor Class Ads mechanism, provide a way to deal with machine heterogeneity and diverse software requirements in compute clusters. Experience at Google suggests that task placement constraints can have a large impact on task scheduling delays. Although our data is obtained from Google compute clusters, the methodology that we develop is general. In particular, the UM metric applies to any compute cluster that employs a Class Ads style of constraint mechanism.

3. Optimal Resource Allocation

Traditional job scheduling is often formulated as a kind of combinatorial optimization problem or queue-based multiprocessor scheduling problem. That is, most of the existing deadline-driven task scheduling solutions from single cluster environment confined in LAN to the Grid computing environment suitable for WAN are also strictly subject to the queuing model under which a single machine's multiple resources cannot be further split to smaller fractions at will. Some successful platforms or cloud management tools leveraging VM resource isolation technology include Amazon EC2 and Open Nebula. Traditional optimization problems are often subject to the precise prediction of task's characteristic (or execution property), which is nontrivial to realize in practice. Accordingly, as the state of the art, further analyze our algorithm's optimality approximation ratio given the possibly wrong predictions of tasks' execution properties.

By formulate demand for computing power and other resources as a resource allocation problem with multiplicity, where computations that have to be performed concurrently are represented as tasks and a later task can reuse resources released by an earlier task. By present an algorithm (Optimal

International Journal of Research
International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 07, July 2015
Available at http://internationaljournalofresearch.org

Resource Allocation) with a proof of its approximation bound that can yield close to optimum solutions in polynomial time. Enterprise users can exploit the solution to reduce the leasing cost and amortize the administration overhead.

3.1 Distribute Load

The Process distributed incoming task to available system resources and achieving good load balance in a fully decentralized and heterogeneous cloud environment. Allocate resource for task with its resource requirements that can minimize a task's execution time. Each user needs to precisely predict the execution property (i.e., workload ratio) for task, before constructing the resource allocation with minimized payment for its execution under a user-specified deadline.

3.2 Minimize Payment

Each task's execution may involve multidimensional resources, such as CPU and disk I/O. Upon receiving the request from user, the scheduler checks the pre collected availability states of all candidate nodes, and estimates the minimal payment of running the task within its deadline on each of them. The host that requires the lowest payment will run the task via a customized VM instance with isolated resources. Specifically, the VM will be customized with such a CPU rate and disk I/O rate that the task can be finished within its deadline and its user payment can also be minimized meanwhile. Finally its computation results will be returned to users.

3.3 Fault Tolerance

Cloud systems usually do not provision physical hosts directly to users. If the resources provisioned are relatively sufficient, by guarantee task's execution time always within its deadline even under the wrong prediction about task's workload. Each task can be guaranteed to be finished within its original deadline even though task properties cannot be predicted accurately.

4. New Proposed Scheme

The optimal solution to the problem with unbounded capacities. Focus on such a question: what is the final upper bound of task execution length as compared to its predefined deadline, when running it using the resource vector allocated with inaccurately predicted workload information .It implement a web service-based prototype that can compute a set of combined matrix operations. Each matrix operation is called by some user task through a web service API and each task is executed in a VM container. Algorithm is evaluated on such a real cluster environment. Users can submit their computation request by editing their mathematical formulas. By make use of Parallel Colt to perform math computations, each consisting of a partially ordered set of operations. Parallel Colt is such a library that can effectively calculate complex matrix-operations like matrix multiply, in parallel via multiple threads.

4.1 System Architecture

The user that requires the lowest payment will run the task via a customized VM instance with isolated resources. Specifically, the VM will be customized with such a CPU rate and disk I/O rate that the task can be finished within its deadline and its user payment.

The user that requires the lowest payment will run the task via a customized VM instance with isolated resources. Specifically, the VM will be customized with such a CPU rate and disk I/O rate that the task can be finished within its deadline and its user payment. It can also be minimized meanwhile .Finally its computation results will be returned to users, running two VMs that are allocated with half of the total physical resources, so its availability vector. If there are no workloads being executed simultaneously for a particular task, its total execution time will be the sum of the individual processing times on different dimensions

The user that requires the lowest payment will run the task via a customized VM instance with isolated resources. Specifically, the VM will be customized with such a CPU rate and disk I/O rate that the task can be finished within its deadline and its user payment. It can also be minimized meanwhile .Finally its computation results will be returned to users, running two VMs that are allocated with half of the total physical resources, so its availability vector. If there are no workloads being executed simultaneously for a particular task, its total execution time will be the sum of the

individual processing times on different dimensions. If the execution of the workloads overlaps, however, the task's completion time would be shorter. Accordingly, it's final execution .For simplicity, denote task execution time as where denotes a constant coefficient .Load distributor which distribute the resource which is under the control of dead line recovery and monitor.

5. Conclusion

This paper proposed a approach to optimize the shared resource allocation among multiple users with low level analysis on upper bound of task execution length under predictionerrors. It also proposed a dynamic approach that adapts to the loaddynamics over task execution progress.An efficient algorithm (Optimal Resource Allocation) to determine the optimal resource, aiming to minimize user's payment on task and also Endeavour to guarantee its execution deadline meanwhile in addition analyze the approximation ratio for the expanded execution time generated by the algorithm to the user-expected deadline, under the possibly inaccurate task property prediction. When the resources provisioned are relatively sufficient, guarantee task's execution time always within its deadline even under the wrong prediction about task's workload characteristic. The approximation ratio for the expanded execution time generated by our algorithm to the user-expected deadline, under the possibly inaccurate task property prediction. When the resources provisioned are relatively sufficient, By guarantee task's execution time always within its deadline even under the wrong prediction about task's workload characteristic.

References

[1] D. Milojicic, I.M. Llorente, and R.S. Montero, "Opennebula: ACloud Management Tool," IEEE Internet Computing, vol. 15, no. 2, pp. 11-14, Mar./Apr. 2011..

[2] Y. Wu, K. Hwang, Y. Yuan, and W. Zheng, "Adaptive Workload Prediction of Grid Performance in Confidence Windows," IEEE Trans. Parallel and Distributed Systems, vol. 21, no. 7, pp. 925-938, July 2010..

[3] P. Crescenzi and V. Kann, A Compendium of NP Optimization Problems, ftp://ftp.nada.kth.se/Theory/Viggo-Kann/compendium.pdf, 2012

[4] S. Chaisiri, R. Kaewpuang, B.-S. Lee, and D. Niyato, "CostMinimization for Provisioning Virtual Servers in Amazon ElasticCompute Cloud," Proc. 19th Ann. IEEE/ACM Int'l Symp. Modeling,Analysis and Simulation of Computer and Telecomm. Systems(MASCOTS '11), pp. 85-95, 2011.

[5] L.M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A Break in the Clouds: Towards a Cloud Definition," SIGCOMMComputer Comm. Rev., vol. 39, no. 1, pp. 50-55, 2009.

[6] D. Gupta, L. Cherkasova, R. Gardner, andA.Vahdat, "Enforcing Performance Isolation across Virtual Machines in Xen," Proc.ACM/IFIP/USENIX Int'l Conf. Middleware (Middleware '06), pp. 342-362, 2006.

[7] M.C. McElvany and P.D. Stotts, "Guaranteed Task Deadlines for Fault-Tolerant Workloads with Conditional Branches," Real-Time Systems, vol. 3, no. 3, pp. 275-305, 1991.

[8] S. Boyd and L. Vandenberghe, Convex Optimization. CambridgeUniv. Press, 2009.

[9] S. Di, D. Kondo, and W. Cirne, "Characterization and Comparison of Cloud versus Grid Workloads," Proc. 14th Int'l Conf. Cluster Computing, pp. 230-238, 2012.

[10] O. Sinnen, Task Scheduling for Parallel Systems. Wiley-Interscience, May 2007.

[11] K. Ramamritham, J.A. Stankovic, and W. Zhao, "Distributed Scheduling of Tasks with Deadlines and Resource Requirements," IEEE Trans. Computers, vol. 38, no. 8, pp. 1110-1123, Aug. 1989.

[12] Amazon Elastic Compute Cloud, http://aws.amazon.com/ec2/, 2012.

[13] I. Foster and C. Kesselman, The Grid 2: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, Nov. 2003.