International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 04 • Issue 2 pp. 274-277 September, 2013
www.csjournals.com

Allocation of Resources for Sharing Cloudlets

Peter Vinodchand. K. (Research Scholar)
Assistant Professor, Department of Computer Science
Government First Grade College. ALNAVAR.

Abstract: A cloudlet is known as a minimial-scale data center or a cluster of computers which are designed to swiftly provide cloud computing services to mobile devices, such as smartphones, tablets and wearable devices, within close geographical propinquity. These are specifically designed to support interactive and resource-intensive mobile applications like speech recognition, language processing, machine learning and virtual reality. The goal of a cloudlet is to augment the retort time of applications running on mobile devices by using low latency, high-bandwidth wireless connectivity and by hosting cloud computing resources, such as virtual machines, bodily closer to the mobile devices accessing them. They eliminate the wide area network (WAN) latency delays that can occur in traditional cloud computing models.

Introduction: A cloudlet is well thought-out as a form of cloud computing for the reason that it delivers hosted services to users over a network. However, it differs from a public cloud data axis, such as those operated by public cloud providers in a number of ways. At the outset it is self-managed by the businesses or users that utilize it, while a public cloud data center is managed around the clock by a cloud provider. It predominantly uses a local area network (LAN) for connectivity, in opposition to the public Internet. It is employed by a smaller quantity, more localized users than a major public cloud service and contains only "malleable affirm" copies of data, such as a cache copy, or code that is stored elsewhere.

Poles apart from cloud data centers that are optimized for launching existing VM images in their storage bands, cloudlets need to be much more agile in their provisioning. Their association with mobile devices is highly dynamic, with substantial whip due to client mobility. A user from distance may surprisingly show up at a cloudlet and it for an application such as a personalized language translator. For that user, the provisioning delay before he is able to use the application impacts usability. Many mobile services split the application into a front-end client program and a back-end server program following the traditional client-server model. The front-end mobile application offloads its functionality to the back-end servers for various reasons such as speeding up processing. With the advent of cloud computing, the back-end server is typically hosted at the cloud data center. Though the use of a cloud data center offers various benefits such as scalability and flexibility, its consolidation and centralization leads to a large separation between a mobile device and its associated data center. End-to-end communication subsequently involves many association hops and results in high latencies and low bandwidth. For the reasons of latency, some rising mobile applications require cloud off-load infrastructure to be close to the mobile device to

achieve low response time. In the ideal case the offload infrastructure could be located in a cellular base station or it could be LAN-connected to a set of Wi-Fi base stations.

Criterias for Resource Allocation:

a) Resource contention: A situation when two applications try to access the same resource at the same time.

b) Scarcity of resources: It arises when there are limited resources.

c) Resource fragmentation: It arises when the resources are isolated.

d) Over-provisioning of resources: Arises when the application gets surplus resources than the demanded one.

e) Under-provisioning of resources: Occurs when the application is assigned with a fewer numbers of resources than the demand.

Strategies:

Execution Time : Different kinds of resource allocation mechanisms are proposed in cloud. The actual task execution time and pre scheduling is considered for resource allocation. It overcomes the problem of resource contention and increases resource utilization by using different modes of renting computing capacities. But estimating the execution time for a job is a hard task for a user and errors are made very often . However the VM model considered is heterogeneous and proposed for IaaS. The matchmaking strategy in is based on schedulability criteria for assigning jobs to opaque resources in heterogeneous environment subjected to AR's (Advance Reservation).

Policy: Since centralized user and resource management lacks in scalable management of users, resources and organization level security policy it is proposed to incorporate a decentralized user and virtualized resource management for IaaS by adding a new layer called domain in between the user and the virtualized resources. Based on role based access control (RBAC), virtualized resources are allocated to users through domain layer. It requires a complex searching process, involving simulated allocation activities, to determine the target cluster. The clusters are assumed to be homogeneous and geographically distributed with the number of processors in each cluster being binary compatible.

Virtual Machine (VM): A system which can automatically scale its infrastructure resources is designed in composition of a virtual network, of virtual machines capable of live migration across multi- domain physical infrastructure. By using dynamic availability of infrastructure resources and dynamic application demand, a virtual computation environment is able to automatically relocate itself across the infrastructure and scale its resources. Users can set up and boot the required resources. It is implemented by enabling the users to dynamically add and/or delete one or more instances of the resources on the basis of VM load and the

conditions specified by the user. The RAS on IaaS differs from RAS on SaaS in cloud because SaaS delivers only the application to the cloud user over the internet.

Hardware Resource Dependency: To improve the hardware utilization, Multiple Job Optimization (MJO) scheduler is proposed. Jobs could be classified by hardware-resource dependency such as CPU bound, Network I/O-bound, Disk I/O bound and memory bound. MJO scheduler can detect the type of jobs and parallel jobs of different categories. Based on the categories, resources are allocated. This system focuses only on CPU and I/O resource. The common feature of these frameworks is to allocate virtual resources based on the available physical resources, expecting to form a virtualization resource pool decoupled with physical infrastructure. Because of the complexity of virtualization technology, all these frameworks cannot support all the application modes. Cloud infrastructure refers to the physical and organizational structure needed for the operation of cloud.

The stepwise resource co-allocation is done in three phases. The first phase determines the co-allocation scheme by considering the CPU consumption amount for each physical machine (PM). The second phase determines whether to put applications on PM or not by using simulated annealing algorithm which tries to perturb the configuration solution by randomly changing one element. During phase 3, the exact CPU share that each VM occupies is determined and it is optimized by the gradient climbing approach. This system mainly focuses on CPU and memory resources for co-allocation and does not considered the dynamic nature of resource request.

Advantages:

1) The biggest benefit of resource allocation is that user neither has to install software nor hardware to access the applications, to develop the application and to host the application over the internet.

2) No limitation of place and medium. The applications and data anywhere in the world, on any system can be reached.

3) The user need not need expend on hardware and software systems.

4) Cloud providers can share their resources over the internet during resource scarcity.

Limitations:

1) Since users rent resources from remote servers for their purpose, they don't have grip over their resources.

2) Migration problem occurs, when the users wants to switch to some other provider for the better storage of their data. It isnt easy to transfer huge data from one provider to the other.

3) In public cloud, the clients' data can be susceptible to hacking or phishing attacks. As the servers on cloud are interconnected, it is easy for malware to spread.

4) Peripheral devices like printers or scanners might not work with cloud. Many of them require software to be installed locally.

5) Networked peripherals have lesser problems.

6) More and deeper knowledge is required for allocating and managing resources in cloud, since all knowledge about the working of the cloud mainly depends upon the cloud service provider.

CONCLUSION: Cloud computing technology is increasingly being used in enterprises and business markets. An effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. Some of the strategies discussed above mainly focus on CPU, memory resource factors. Cloud users and cloud providers are the two troupe in a cloud computing environment, have different goals to pursue. Cloud providers promote high resource utilization to maximize profits while users pursue an opposite goal.

References:

https://www.section.io/engineering-education/resource-allocation-algorithms-cloud-environments

https://docs.pexip.com/server_design/resource_allocation_scenarios

https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-018-0105-8

https://ieeexplore.ieee.org/document/9463205