Efficient Techniques for Load Balancing in Cloud

Prof. Krishnajali J. Shinde
Computer Engineering Department
Atharva College Of Engineering, Malad
Mumbai -400095

Prof. Satish Ranbhise
Computer Engineering Department
Atharva College Of Engineering, Malad
Mumbai -400095

Abstract— **Cloud computing is the new technology, which is totally dependent on the internet to maintain large applications, where data is shared over one platform to provide better services to the clients belonging to a different organization. Load balancing is one of the main challenges in cloud computing. It is a technique which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overloaded. So that Load balancing techniques help in optimal utilization of resources and hence in enhancing the overall performance of the system. The goal of load balancing is to minimize the resource consumption which will further reduce energy consumption .Its main motive is to optimize the usage of resources, boost turnout, fault tolerance, scalability, increase throughput, response time, etc [1]. It becomes a severe problem with the increase in list of users and types of applications on cloud. The main highlights of this paper is on the load balancing techniques in cloud computing**.

Keywords :- Cloud computing, Load Balancing Virtualization, Scheduling, Load Balancing Algorithms, Virtual Machine.

I. INTRODUCTION

Cloud computing is new technology which is based on internet in which internet can represented as a cloud. Cloud is a platform that provides resources like services, applications and storage network to a computers and devices based on pay-per- model [1.2].Many Cloud computing providers have setup a few data centers at various geographical places over the web as a way to serve wants of their buyers around world. Today's this technology rising at a quick rate. The fundamental purpose of using this technology is to growth the performance and efficiency and reduces the cost. The increases amount of information storage quickly in cloud computing environment. The increase data storage very fast so the load balancing is a primary concern in cloud computing. When a numbers of jobs occur equal time then load balancing is predominant issue. Load balancing helps to work distribute between all to be had nodes to be certain that no node is overloaded and no want is free. Cloud computing is an on demand service in which shared resources, information, software and other devices are provided according to the clients' requirement at specific time. It's a term which is generally used in case of Internet. The whole Internet can be viewed as a cloud. Capital and operational costs can be cut using cloud computing.

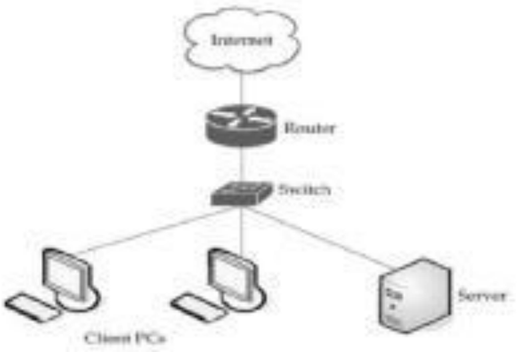


Figure 1: A cloud is used in network diagrams to depict the Internet

Load balancing is the new idea that facilitates networks and resources by a maximum throughput with minimum response time. Proper load balancing can help in utilizing the available resources optimally, thereby minimizing the resource consumption. Load balancing schemes depending on whether the system dynamics are important can be either static or dynamic. Static load balancing scheme divide the traffic equivalently between the services. Dynamic load balancing scheme chooses the lightest server preferred to balance the traffic and selecting an appropriate server needed real time communication with network.

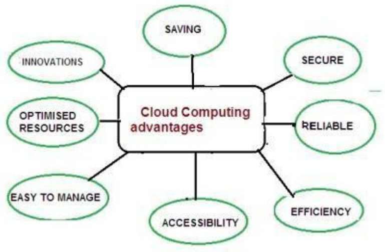


Figure2. Characteristics of Cloud Computing

1.1 Cloud Service Models: The three main services provided by the cloud are IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service). The basic and a short description of these three services are as follows:

IaaS: Infrastructure as a Service (IaaS) is the delivery of computer hardware (servers, networking technology, and storage and data centre space) as a service. It also includes the delivery of various operating systems and virtualization technologies to manage the resources. The IaaS customers rent computing resources instead of buying and installing them in their own data centre. The service is typically paid for on a usage basis (Pay and Use).

PaaS: Platform as a Service (PaaS) is a category of cloud computing services that provide a platform allowing customers to develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an application.

SaaS: Software as a Service (SaaS) is a software licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted. Consumers purchase the ability to access and use an application or service that is hosted in the cloud.

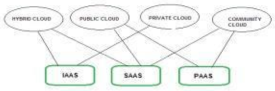


Figure3. Cloud Services Models

1.2 Cloud Deployment Models

The cloud group characterizes four cloud organization models:

1) *Public Cloud:* **-** This type of cloud is utilized by the general public users and the cloud service provider has the full responsibility for public cloud with its own qualities, policy, costing, profit, and charging model. Many popular cloud services are Google App Engine, Amazon EC2and salesforce.com.It is used for pay-as-you-go scalability, ideal for heavy/unpredictable traffic

2) *Private Cloud:* **-**Private cloud will be cloud bases worked for a solitary association and give security to its resources.

3) *Community Cloud:* **-** In community cloud, cloud infrastructure which can be used through several organizations in a private community. This cloud is shared amongst many associations that have comparative cloud prerequisites.

4) *Hybrid Cloud: (combination of both private and public clouds)* **-** This cloud it utilizes a combination of no under two clouds where the clouds incorporate a blend of private cloud, public cloud or community cloud.

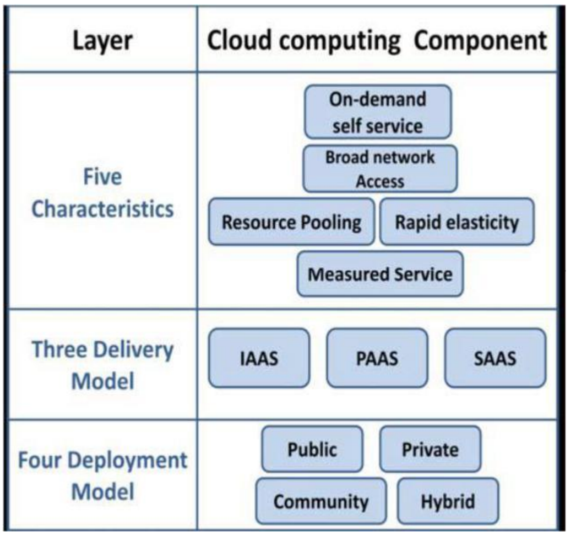
1.3 **Cloud Components**

A Cloud system consists of 3 major components such as clients, datacenter, and distributed servers.

1) ***Clients :*** End users interact with the clients to manage large data/ information related to the cloud
2) ***Datacenter:*** Datacenter is nothing but a collection of servers hosting different applications. An end user connects to the datacenter to subscribe different applications. Datacenter is set of hosts.

   This can be responsible regarding managing virtual models (VMs) (e.g., VM provisioning). It behaves similar to a IaaS provider from finding requests with regard to VMs via brokers

3) ***Distributed Servers***: Distributed servers are the parts of a cloud which are present throughout the Internet hosting different applications. But while using the application from the cloud, the user will feel that he is using this application from its own machine.



Layer	Cloud computing Component
Five Characteristics	On-demand self service Broad network Access Resource Pooling Rapid elasticity Measured Service
Three Delivery Model	IAAS PAAS SAAS
Four Deployment Model	Public Private Community Hybrid

Figure 4.Cloud Computing Model

II. VIRTUALIZATION

Virtualization means which are not exist in real, but it provides everything like real. In cloud computing, virtualization is very useful concept which means something which is not real and to create a virtual version of resource, such as a server, storage device, network or even an operating system. It is the software implementation of a computer which will execute different programs like a real machine. Even, something as simple as partitioning a hard drive is considered virtualization because a drive can be partitioned into more than one it uses for provisioning services to the client. On virtualization many running systems can be keep running on the single computer so resource utilization is increased. The hardware resources are combined for enhanced the productivity of server. Computer architecture is uses software for the proper resource utilization called Hypervisor. It is named the VMM (Virtual Machine Monitor) for running the more than one operating systems on the single host. There are two types of virtualization.

A. *Full Virtualization:* **-** In Full Virtualization, the entire installation of one machine is done on another machine. So all the real machine functionality may too be available in virtual machine.

B. *Para Virtualization:* - In Para Virtualization, on a solitary desktop different operating systems can be run. Here entire functionalities are not fully available; services are delivered in a partial manner.

III. LOAD BALANCING

Load balancing is the process of distributing the load among various resources in any system. Thus load need to be distributed over the resources in cloud-based architecture, so that each resources does approximately the equal amount of task at any point of time. Basic need is to provide some techniques to balance requests to provide the solution of the application faster. In cloud environment, load balancing is a method that circulates the dynamic nearby workload similarly across entire available nodes [2]. Load balancing is used for achieving an enhanced resource utilization and service provisioning, therefore enhancing the entire system performance. In load balancing, incoming tasks are coming from the different location are received by the load balancer and then tasks are circulated to the data for the appropriate load distribution. The important goal of a load balancing in cloud computing remains to growth the reply time of job with the aid of job by the whole load of procedure. Load balancers can operate in dual exclusive approaches: one is the cooperative and other is Non-cooperative. In non-cooperative mode, the duties run independently as a way to enhance the reply time of nearby tasks. In cooperative, the nodes work even as to be able to acquire the common purpose of optimizing the overall response time.

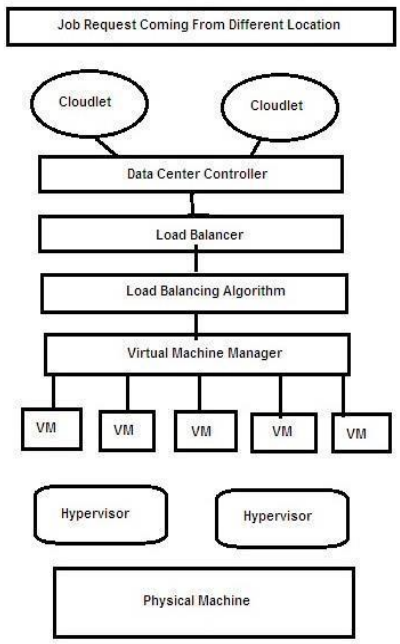


Figure 5. Cloud Load Balancing

In the cloud load balancing algorithms, is mostly divided into two different groups: static and dynamic load balancing algorithm:

A. Static Approach: This type of static approach is commonly described in implementation or design of system. This algorithm divisions the traffic alike between all of the users. This algorithm requires a previous knowledge of approach resources, so that no longer rely upon the existing state of system for decision of shifting of the load. These are much simpler and ignore the existing state or the load on the node within system.

B. Dynamic Approach: In this dynamic approach considered only the present condition of the system during load balancing decision. This

Dynamic methodology is used for broadly distributed system such a cloud computing. Dynamic approach has two parts:

1) Centralized Approach: **--** In this centralized approach Simplest as only one node is in charge for distribution and managing within the whole system.

2) Distributed Approach: **--** in this dynamic methodology each node freely constructs own load vector. Vector accumulating the load know-how of other nodes. In this all selections are made locally utilizing local load vectors. This procedure is more compatible for generally allotted systems comparable to cloud computing

IV. LOAD BALANCING ALGORITHMS IN CLOUD COMPUTING

In a cloud computing there are different load balancing algorithms. There are various issues while dealing with load balancing in a cloud computing environment. Each load balancing algorithm must be such as to achieve the desired goal. Load balancing in the cloud computing environment has an important impact on the performance. Good load balancing makes cloud computing more efficient and improves user satisfaction. Load balancing with cloud computing provides a good efficient strategy to several inquiries residing inside cloud computing environment set. A load balancing algorithm has five major components

A. *Transfer Policy:* In this policy it is responsible to determine when a task should be transferred from one node to the other node.
B. *Selection Policy:* This type of policy focuses on choosing the processor for load transfer so that the overall response time and throughput may be improved.
C. *Location Policy:* In this policy it determines the availability of essential resources for providing services and makes a selection based on location of resources.
D. *Information Policy:* In this policy it acquires workload related information about the system such as nature of workload and the average load on each node. It is also responsible for exchanging the information from one node to another, along with method of exchange and the amount of the information to be exchanged. For exchanging load information of a node, three methods can be used which are Broadcast approach, Global System Load, Polling approach.
E. *Load Estimation Policy:* In this policy it determines the total workload of a node in a system.

Some algorithms aim to achieving higher throughput, minimum response time, and maximum resource utilization. Some load balancing algorithms are listed below.

A. Task Scheduling founded on load balancing:--This algorithm consists two level task scheduling device of load balancing to meet element requirements of users. This algorithm obtains high resource utilization, and this algorithm attains load balancing via first mapping jobs to virtual machines after which all virtual machines towards host resources .This improving the task response time, also provide well useful resource utilization .

B. Opportunistic Load Balancing Algorithm (OLB):--This algorithm does not meet the existing workload of the Virtual machine (VM). It makes every node to be occupied. On this algorithm each and every unexecuted challenge can be finished in random order so that each task can be allotted to the node randomly. These processes slow manner because it will not calculate the existing operation time of the node.

C. Round-Robin Load Balancer:--Round-robin load balancer its' a static load balancing algorithm. The process allocation order is maintained locally independent of the allocations from remote processors. In Round Robin, it send the requests to the node with the least number of connections, so at any point of time some node may be heavily loaded and other remain idle [5], this problem is reduced by CLBDM This algorithm uses for allocating the jobs. In this algorithm it chooses the nodes randomly and after which jobs allot to all nodes in round robin manner.

D. Min-Min algorithm:--Min-Min Job Scheduling

Algorithm it's a static scheduling algorithm. Min-Min algorithm begins with a suite of un-scheduling jobs. On this algorithm the jobs having minimum execution time first identifies and these jobs are scheduled first in this algorithm. Then it will calculate the expected completion time intended for each tasks according to available virtual Machines then the resource that has the least completion time for selected task is scheduled on that resource. The resource ready time is updated and except the entire unexecuted tasks are scheduled the procedure is repeated.

Main problem of this algorithm it's chooses small tasks to be finished firstly, which in turn long task delays for very long time. Min-Min algorithm is did not utilize resources competently which lead to a load imbalance.

E. Max-Min Algorithm:--Max-Min algorithm is close to equal as the min-min algorithm. The core difference among Min-Min and Max-Min algorithm is following: in this algorithm first finding out minimum execution times, then first the most extreme value is choose Then the performance time for all tasks is updated on that machine, this is done by adding the performance time of the assigned task to the performance times of other tasks on that machine. Then all assigned task is erased from the list that executed the system.

F. *Randomized:*--This is a static algorithm in nature. In this randomized algorithm a procedure may also be care of by a specific node n with a likelihood p. This algorithm functions admirably when every single procedure are of equivalent loaded. Issue emerges at the point when burdens are of various computational complexities. This randomized algorithm is not keeping up deterministic methodology .

G. First Come First Serve Algorithm:--In this algorithm jobs are served in the direction wherein they arrive i.e. Jobs are queued and served in structure of FIFO. This algorithm is discreet and really quick but doesn't provide so much effectively to job and resource optimization .

H. Shortest Response Time First:--In this algorithm every procedure is assigned a need which is permitted to run and equivalent need procedures are scheduled in FCFS demand. The SJF algorithm chooses the occupation with the most limited preparing time first. In this SJF algorithm shorter jobs are executed before long jobs. In this algorithm, it is significant to know or evaluation processing time of every job which is large SJF problem .

I. Equally Spread Current Execution:--This is an algorithm of dynamic load balancing, which handles the method with priority. It chooses the need by checking the scale of the system. This similarly spread present execution algorithm disseminates the load randomly by checking the extent of the system after then transferring load to a VM (Virtual Machine). On this algorithm load balancer spreads the load on to various nodes, so it's knows spread spectrum methodology.

J. Resource Awareness Scheduling Algorithm:--Resource Awareness Algorithm it's a blend of Min-Min and Max-Min algorithm and has no time consuming instruction. The time complexity of this algorithm is O (mn2) where n is number of jobs and m is number of resources.

K. Major goals of load balancing algorithms:

a. *Cost effectiveness:* Load balancing help in provide better system performance at lower cost.

b. *Scalability and flexibility:* The system for which load balancing algorithms are implemented may be change in size after some time. So the algorithm must handle these types' situations. So algorithm must be flexible and scalable.

c. *Priority:* Prioritization of the resources or jobs needs to be done. So higher priority jobs get better chance to execute first.

V. SCHEDULING AND LOAD BALANCING

Figure 6. Shows that the scheduling and load balancing design, in which the scheduler has the logic to find the most suitable VM and assign the tasks to VMs based on the proposed algorithm. The scheduler places the run time arrival jobs in the most suitable VMs based on the least utilized VM at that particular job arrival time. Load Balancer decides the migration of task from a heavily loaded VM to an idle VM or least loaded VM at run time, whenever it finds an idle VM or least loaded VM by utilizing the resources current status information. Resource monitor communicates with all the VMs resource prober and collects the VM capabilities, current load on each VM, and number of jobs in execution/waiting queue in each VM. The task requirement is provided by the user which includes the length of the tasks to be executed and transfers the requirements to the scheduler for its operative decisions.

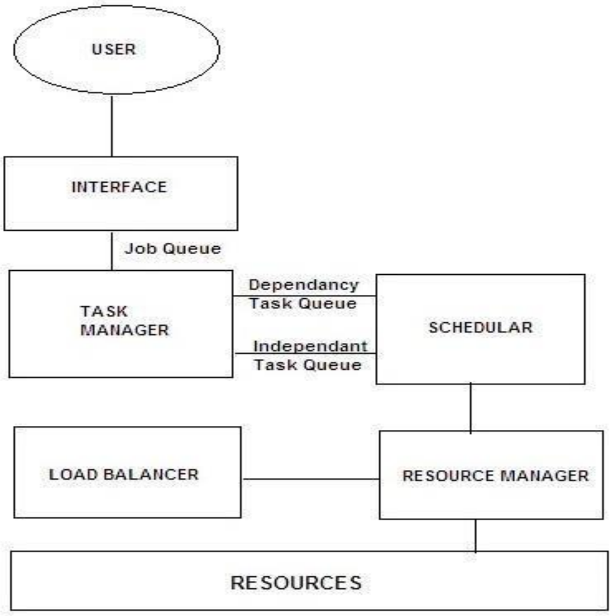


Figure 6: Scheduling and load balancing design.

VI. SCHEDULING AND LOAD BALANCING DESIGN

Job request is given by the user through the interface and passed to task manager for dependency and independent task analysis. This module receives the job and verifies whether the job is a complete independent task or contains multiple tasks. If it contains multiple tasks, then it verifies the interdependency between the multiple tasks. The dependency task queue and independent task queue are found [5]. The dependent tasks will be notified to the scheduler so that parent tasks are scheduled after child tasks are executed. Dependency task queue will contain the tasks, which depends on the other tasks present in the VMs. Once all the child tasks present in this queue completed its execution the parent task will be taken for the execution by assigning it to the VM, whereas independent task queue contains independent tasks. Independent task queue and dependency task are input to the scheduler. The scheduler selects the appropriate VM based on IWRR algorithm. This scheduler collects the resources information from the resource manager. It calculates the processing capacity of each of the VMs and then it applies the proposed algorithm to find the appropriate VM for the given job. Additionally, every VM is maintaining the JobExecutionList, JobPauseList, and JobWaitingList information specific to it. The JobExecutionList contains the current executing job list and the JobPausedList contains the temporarily paused jobs in the VM. Similarly the JobWaitingList Queue contains the waiting jobs on the specific VM, but this will be executed upon receiving the JobExecutionList, JobPauseList, and JobWaitingList from each of the VMs; calculation of the least utilized VM is carried out for every request arrival. Then, this least utilized VM information will be returned to the scheduler. Resource manager communicates with all the VMs to collect each of its capabilities by getting its number of the processing elements and its processing capacity to each of its elements. This resource manager additionally calculates the weightage to each of the VMs based on the processing capacity allotted to it. This also identifies the memory configured available in each of the VMs. Load balancer calculates the ratio between the number of jobs running and the number of VMs. If the ratio is less than 1, then it communicates the scheduler to identify a VM for the job; else it will calculate the load on each of the VMs using the job execution list of the VMs. If the utilization is less than the 20%, then the least utilized VM will be allotted; else the scheduler will be communicated to identify the most suitable VM for the job. Once the appropriate VM has been identified, the Job will be assigned to that VM. The configured data centers include hosts and their VM with corresponding processing elements form the set of resources available for computing. The resources are probed for idleness and for heavy load so that the job requests are effectively allocated to an appropriate resource. The round robin algorithm allocates task to the next VM in the queue irrespective of the load on that VM. The Round Robin policy does not consider the resource capabilities, priority, and the length of the tasks. So, the higher priority and the lengthy tasks end up with the higher response time The weighted round robin considers the resource capabilities of the VMs and assigns higher number of tasks to the higher capacity VMs based on the weight age given to each of the VMs. Improved Weighted Round Robin Algorithm. The proposed improved weighted round robin algorithm is the most optimal algorithm and it allocates the jobs to the most suitable VMs based on the

VM’s information like its processing capacity, load on the

VMs, and length of the arrived tasks with its priority. The static scheduling of this algorithm uses the processing capacity of the VMs, the number of incoming tasks, and the length of each task to decide the allocation on the appropriate VM.

VII. CONCLUSIONS

Cloud computing mainly deals with software, data access and storage services that may not require end-user knowledge of the physical location and configuration of the system that is delivering the services. In the cloud storage, load balancing is a key issue [3]. It helps in proper utilization of resources and hence in enhancing the performance of the system. This paper presents a concept of Cloud Computing along with load balancing. Main thing is considered in this is load balancing algorithm. There are many above mentioned algorithms in cloud computing which consist many factors like scalability, better resource utilization, high performance, better response time.

ACKNOWLEDGMENT

Authors are thankful to department of computer engineering from Atharva College of Engineering, Malad Mumbai, for giving highly motivational supports. The work could not be possible without the help of the people of the department. The work is supported by such people and authors are giving thanks to all.

REFERENCES

[1] Hung, Che-Lun, Hsiao-hsi Wang, and Yu-Chen Hu. "Efficient Load Balancing Algorithm for Cloud Computing Network." InInternational Conference on Information Science and Technology(IST 2012), April, 2012.

[2] Dobale R.G.,SonarR.P. (2015 February) .*Review of Load Balancing for Distributed Systems in Cloud.* International Journal of Advanced Research in Computer Science and Software Engineering: IJARCSSE, 2015 pp.393-403, ISSN: 2277 128X

[3] Singh A, Juneja D. & Malhotra M. (2015). *Autonomous Agent Based Load Balancing Algorithm in Cloud Computing.* International Conference on Advanced Computing Technologies and Applications: ICACTA 2015,

[4] KalaiSelvi B. Mary L. (2014, August). *A Survey of Load Balancing Algorithms using VM.* , International Journal of Advancements in Research & Technology: IJOART 2014

[5] Kaur R. And Luthra P. (2012) .*Load Balancing in Cloud Computing.* Association of Computer Electronics and Electrical Engineers: ACEE 2014.

[6] Panwar R., Mallick B. (2015, May) .*A Comparative Study of Load Balancing Algorithms in Cloud Computing.* , International Journal of Computer Applications: IJCA, May 2015

[7] Pasha.N, Agrawal A. & Rastogi R. (2014, May).*Round Robin Approach for VM Load Balancing Algorithm in Cloud Computing Environment.* , International Journal of Advanced Research in Computer Science and Software Engineering: IJARCSSE 2014

[8] Shah D.M., Kariyani A.A & Agarwal D.L .(2013, February).*Allocation of Virtual Machines in Cloud Computing Using Load Balancing Algorithm*, International Journal of Computer Science and Information Technology & Security: IJCSITS 2013

[9] Domanal S. G., Reddy G. R. M. (2013, October).*Load balancing in Cloud Computing using Modified Throttled Algorithm. In Cloud Computing in Emerging Markets.*Cloud Computing in Emerging Markets: CCEM 2013, pp.1-5, Bangalore, India

[10] Sharma T., Banga V.K. (2013, March**).***Efficient and Enhanced Algorithm in Cloud Computing.* International Journal of Soft Computing and Engineering :IJSCE 2013

[11] Ajit M., Vidya G. (2013, July).*VM Level Load Balancing in Cloud Environment.* Computing, Communications and Networking Technologies: ICCCNT 2013

[12] Kulkarni A.K., Annappa B. (2015). *Load Balancing Strategy for Optimal Peak Hour Performance in Cloud Data centres.*, Signal Processing, Informatics, Communication and Energy Systems: SPICES 2015

[13] Lua Y. , Xiea Q., Kliotb G, Gellerb A. , Larusb J.R. & Greenber A. (2011, August). *Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services.* , An international Journal on Performance evaluation: IJPE 2011