International Journal of Enterprise Computing and Business Systems

ISSN (Online) : 2230-8849

http://www.ijecbs.com

Vol. 1 Issue 2 July 2011

BALANCING LOAD FOR OPTIMAL PERFORMANCE OF DATA CENTER

Karthik Narayan M[1], Shantharam Nayak[2], Annamma Abraham[3]
[1]4th Sem M.Tech (SE), RVCE, Bangalore – 59.
[2]Associate professor ISE-RVCE and Research scholar Dr. MGR University
[3]Professor & HOD Maths department, BMSIT, Bangalore - 560064

ABSTRACT

Data Center in common words is storage room of computers along with the components of it forperforming the essential tasks such as communication, processing and so on. In order, to perform with optimal efficiency data centers should efficiently cater the requested information by the client. The data center is composed of many processing nodes. A “Hotspot” can occur in the datacenter if some of nodes are overloaded and others aren’t. So, to prevent “Hotspot”, techniques of Load balancing along with Caching with RAID concept could be adopted. Load balancing is necessary as the data center should work in networking environment. Caching is necessary for the speedy information supply. RAID concept is proven as one of the best method in achieving the redundancy in a cost effective manner. Load balancing algorithms not only identify the non overloaded node, but also identify the correct path for the smooth transfer. We propose a novel approach of path finder algorithm which works in the combination of choke packet algorithm and path planning algorithm concept. A cache table can be maintained for the purpose of caching with RAID-2 or RAID-3. By adopting the proposed techniques we can make the datacenters to perform its activities quickly and accurately.

International Journal of Enterprise Computing and Business Systems

ISSN (Online) : 2230-8849

http://www.ijecbs.com

Vol. 1 Issue 2 July 2011

Keywords

Caching ,Choke Packet Algorithm, Datacenter, Hotspots, Path Planning Algorithm, Performance Optimization

I Introduction

Data center in simple terms can be described as a repository which is used store data which has importance to us. A data center also called as server farm is a facility used to house computer systems and associated systems such as telecommunication and storage systems. The main purpose of data center is to provide a reliable and secure infrastructure for information systems. They are the crucial aspect of the IT operations in an organization. They require large storage capacity and larger bandwidth for internet connectivity. Hence performance of the data center is a critical aspect in providing services to the external world. Data center today is being transformed into cloud computing which provides various kinds of services. These are the platform as a service (PaaS), software as a service (SaaS), and infrastructure as a service(IaaS). As many companies are moving to cloud, this puts a demand for the efficient management of the resources at the data center. If the efficient management of the resources are done it may lead to load unbalance at the data center. This may lead to a HOTSPOTS which may lead to the poor performance of the applications and the infrastructure at the data center. In this paper we propose a method by which we can manage the load at the data center and thereby increasing the performance.

II Data center Performance metrics to measure the service Delivery

a. Application Workload **-**The load or "demand" made on the system by users. In a stable system this is also equal to the throughput, and is usually described by the business in terms of business transactions. IT will need to put some effort into translating a business transaction into units of work that are executed within the IT infrastructure, but this is often addressed through established capacity planning and chargeback methodologies.

b. Response Time - The main measurement of performance. This is the time each transaction takes to complete. End-user transactions can be externally "clocked" in many ways, and this is often accomplished through implementation of service level management solutions.

c. Utilization - The effective "busy-ness" of the IT system that services the workload. Once utilization reaches 100 percent, no more work can be done.

III Assumptions

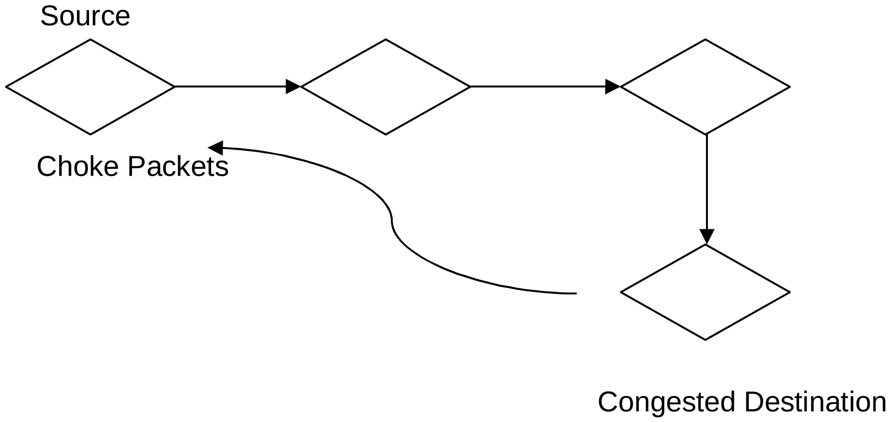
a. It is assumed that all the nodes that provide a similar kind of service are grouped into one cluster
b. It is possible to estimate the load at a node.
c. It is possible to preempt the already assigned job at the node.

IV Problem Statement

Data center provides a customized service for the IT enablement of the organization. Such a data center may have several nodes. These nodes may be serving different parts of the world. They may have different set of users. They may have different capacities. Hence the load across the various nodes may vary from time to time. This will create various kinds of congestions at the node. They can be classified as resource congestion and the network congestion. The resource congestion happens at the node itself, which happens when the number of applications running on the node overshoots. The network congestion happens when network connecting to the node may be overloaded. This kind of overloads will lead to some of the nodes overburdened and others idle. This is in turn will lead to reduce in the response time for some parts of the data center. This is not desirable for mission critical applications using data center. In this paper we try to apply the standard choke packet like algorithm used to resolve the network congestion to reduce the load at the data centers and ensure a uniform load to increase the performance. We also use the principle of geo-aware caching to reduce the load.

V Choke packet algorithm

This section applies how choke packet algorithm can be applied When the number of packets in the part of the subnet too many then performance of that part of the network degrades. This situation is called congestion. There are many approaches to address this problem.



In the choke packet approach the congested router sends a choke packet back to the source host, giving it the destination found in the packet. The original packet is tagged (a header bit is turned on) so that it will not generate any more choke packets farther along the path and is then forwarded in the usual way. When the source host gets the choke packet it is required to reduce the traffic sent to the specified destination by certain percent. In our algorithm we use the choke packet algorithm to detect an overloaded node and distributed the load across an idle node.

VI Path Planning Algorithm

This algorithm is used to detect a shortest path between a given source and destination. In our algorithm we apply this algorithm in order to find the best node to which the job can be transferred. Here the shortest means the most cost effective node to which the load can transferred

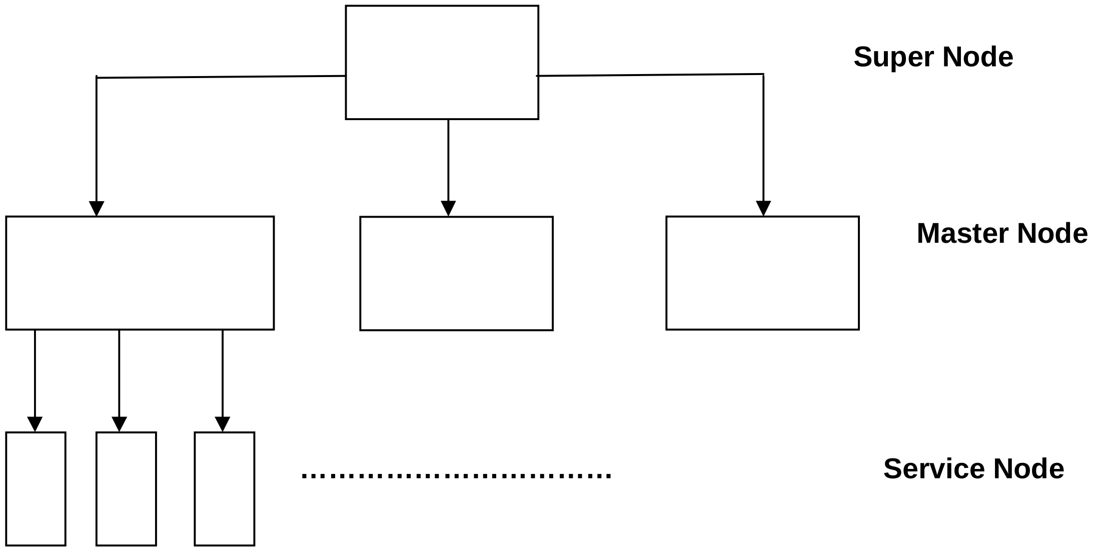
VII Caching

A cache is a component that transparently stores data so that future requests for that data can be served faster. The data that is stored within a cache might be values that have been computed earlier or duplicates of original values that are stored elsewhere. If requested data is contained in the cache (cache hit), this request can be served by simply reading the cache, which is comparatively faster. Otherwise (cache misses), the data has to be recomputed or fetched from its original storage location, which is comparatively slower. Hence, the more requests can be served from the cache the faster the overall system performance can be improved.

The client who is requesting data from a system is not aware that the cache exists, To be cost efficient and to enable an efficient use of data, caches are relatively small. Nevertheless, caches have proven themselves in many areas of computing because access patterns in typical computer applications have locality of reference. References exhibit temporal locality if data is requested again that has been recently requested already. References exhibit spatial locality if data is requested that is physically stored close to data that has been requested already.

VIII Proposed Solution

Basically a data center can be viewed as a server farms. Such a farm can have a Master node whose main task is to balance the load across the nodes. Here it is assumed that the Master node manages all the nodes which provide the same type of service. All the Master nodes are in turn managed by another super node. Hence the nodes are arranged in the hieratical manner.



The services provided by the data center are categorized into different classes. The super node is responsible for registering the various kinds of service. It keeps track of master nodes responsible for that particular service. The master node runs a load balancer algorithm to balance the load among nodes that provide a service. The integral part of the load balancing algorithm is customized choke packet algorithm and the path planning algorithm. Each service node also runs a daemon which will estimate the load at that particular node. This load is calculated using various parameters such as processor utilization, number of applications running on that node, and various types of resource utilization.

The load balancing algorithm will involve these steps

a. The Master will receive a request for a service. The master will allocate this service request to one of the nodes in a round robin fashion.
b. When a node is allocated a job, a check is made whether the node can take this amount of work .
c. If the node can accept the work then it will take up the work and finish it off
d. If the node cannot accept the work then it will send a the choke packet back to master node
e. Once the master node will receive the choke packet then it will not send any work to it, until it receives a assign packet from the node.

f. If the choke packet is received for a work assigned, then it has to be reassigned to some other node.
g. This reassignment is based on a path planning algorithm.
h. The master node will ask all nodes to send an estimation of the work load at their nodes.
i. When the master node will receive the estimation, it will find out the node with the least amount of workload.
j. Then the work will assigned to that node.
k. The master node will cache all the service request assignment to the nodes.

l. The nodes will cache all the service request results for last unit time duration.
m. When the service request is received then it is checked if it was served earlier, then it can be reassigned to same node that served. This node will send back the cached result.

IX Conclusion

Data centers are being used in more and more applications such as cloud computing, enterprise resource management across various domains. This increases the load on the data center. The increase in the load will lead to decrease in the service delivery parameters considered above. Thus this will instantiate a need for a approach where the load across the data center is uniformly balanced. This paper discussed an approach the load across the data center is uniformly balanced. This paper made use of the choke packet algorithm which is used in computer networks to indicate congestion and the path planning algorithm which is used to plan a path among the group of nodes to which the work can be delegated.

References

[1] Dynamic Load Balancing Multipathing in Data Center Ethernet, Yang Yu; Khin Mi Mi Aung; Tong, E.K.K.; Chuan Heng Foh, IEEE, August 2010

[2] Server-storage virtualization: Integration and load balancing in data centers, Singh, A.; Korupolu, M.; Mohapatra, D, IEEE, November 2008

[3] A Novel Multipath Load Balancing Algorithm in Fat-Tree Data Center, Cuiron Wang, IEEE, 2009