An Evaluation of Load Balancing between Server Platforms inside a Data Center

Amer Tahseen Abu Jassar
Faculty of sciences and information technologies, Irbid National University
Jordan/Irbid

Abstract

The overview of principles of network load balancing between server platforms in a data center network is given in the work. The data center model in term of Infrastructure-as-a-Service is proposed in the work. The CPU utilization of server platforms and throughput capability which depended from the amount of servers is suggested to use as a main performance characteristics. The evaluation model of data center network and the formulas for CPU and throughput evaluation are also suggested in the work. The analysis of load balancing between server platforms was performed according to the proposed formulas in case of use a clustering.

Key words: *load balancing, throughput, CPU utilization, data center, server platform.*

Introduction

The rapid growth of online services and applications requires the improvement of multi-service networks functionality. The permanent increasing of network load without changes in management and transport network technologies is leading to the number of adverse effects: lower productivity, an increasing delays and response time and, as a result, partial degradation of the provided service. More and more organizations are being use data centers concept for supporting a large number of distinct services. Data centers provide a high-level quality of service and computation resource flexibility for enterprises and providers of services. The Data Centers based on Infrastructure-as-a-Service cloud service model obtain a wide popularity [1]. This approach allows integrating the existing network technology and responding more quickly to business demands, increased effectively and networking scalability. Multiple cloud technologies of data storage and computing, network virtualization and data flows aggregation was provided to the IT industry with aim to increase the scalability and flexibility of network solutions. However, each new approach makes the infrastructure of data center more complex [2].

The different load balancing technologies are using to achieve a high level of processed load between components of data centre. It is help to increase a scalability and fault tolerance. But, load balancing between server platforms is a difficult task, which today has a number of solutions. The effectiveness of the solution depends on the technical characteristics of network elements. Two independent tasks should be solved in order to achieve maximum efficiency in the allocation of network resources. The first task is the load distribution within the network (between server platform and another computing component).

The second task is the management and distribution of incoming traffic.

The uniform principles of network load balancing between servers platforms is considered in the work.

Such parameters as CPU overhead and throughput are suggested to measuring for evaluating efficiency of load balancing of server platform.

The obtained calculation results will give ability to increase effectively of future data center network solution in the future.

The model of data centre network (in term of infrastructure-as-a-service) and basic components of network, requirements for effectively services providing are given in Section I.

The basic tasks of load balancing are also considered. The logical structure of the load balancer and its location in the network architecture is given in Section II. Model of network load distribution for main network indicators such as the number of server platforms and its technology characteristics and comparison experimental and calculation result are given in Section III.

1. A model of data center network

The modern data centers consist from a set of management component (task scheduler, load balancing components), storages and computing platforms (servers, computation nodes and VMs), which simultaneously load a fragment or full copy of the application [3]. Depending on the business demand, services may be replicated using server platforms with different functions. This concept is more popular today [4]. It allows realizing next benefits: increasing scalability flexibility of network, quick adaptation for new services needs, decreasing amount of executing limitation. The data centers network in terms of Infrastructure-as-a-Service (IaaS) includes the following set of components: different types of servers data storages and network components, applications are running above of them and produce a lot of concurrent processes that can be provided to the customer with dynamic intensity. The structure and main component of data centre architecture is depicted on Figure 1. The major components of architecture are [2]:

Client or and users that interact with the clouds to manage information related to the cloud.

Servers that hosting different applications: a server, which actively checks and executes the services of their hosts. The main compute resources is belong to the servers and VMs [5]. All online services and applications are executed across them. The cloud infrastructure configuration and future maintains in the most cases is determine by the provided applications and the processes.

In general, two types of servers are existed in data centre [1, 2]. Front End server receives and handles incoming traffic and multiple application Middle End servers that running applications. The Front End servers [1] provides reception requests from users, translation requests to Middle End Servers and delivery of application or service to the user. In this way Middle End servers are the computing components.

Management and forwarding components: a hardware and software which manages, distributes traffic and delivers services. These components interact with all cloud executed component and obviously include persistence and session management and task schedulers. The load balancing has the most significant weigh in the services provision.

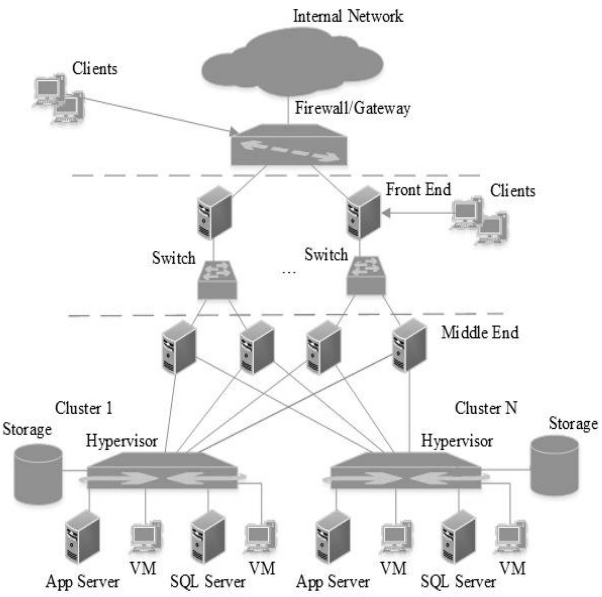

Figure 1 – The main component of a data center network.

The achievement of effective workload distribution across multiple infrastructure components through appropriate paths and optimal utilization of infrastructure recourses are be able due to use a load balancing algorithms. In this way load balancer determines the effective balancing and the optimal distribution proportions of incoming requests and their subsequent dynamic adjustments when changing the system operating conditions (change in the number of nodes, and the composition of the set of running applications, upgrading the hardware platform of separate nodes etc). Load balancing can be implemented by hardware, software, or a combination of both [5].

The major goals and requirements of load balancing algorithms in network infrastructure are [3, 4]:

- Scalability. For the concurrent system, particular in Data Centers, the high scalability is one of the key driver – the total network performance and reliability directly depends on scalability. The load balancing mechanisms allow to maintain the required level of network scalability and flexibility.
- Prioritization: the resources prioritization belongs to the earlier stages of service life cycle. The resources prioritization mechanisms allows to optimize service provision procedures.

Load balancing on server is designed for critical applications. It improves application performance by sharing the load between several different platforms [4]. For example, the strong rule of load balancing requires for the Front End servers (see Figure 1). It is provides performance and reliability for clients. For the Middle End servers Load balancing is performed inside the network. Middle End servers include the same set of load balancing mechanisms as the Front End servers. Balancing on the network layer involves the tasks, which allows to one particular server IP-address corresponding to different physical or virtual machines. Such balancing is highly effective and can be implemented using a variety of different ways.

The main benefit of load balancing mechanisms used on Middle End servers [4] has two possible realization – can be realize both as software of hardware where one of solution can be reserved. This give ability to prevent the failure of services or eliminate the bottlenecks in compute performance. Many Middle End servers use the open source architecture. The significant cost savings is provided due to these solutions.

However, the lake of effective distribution appears in case of data distribution between Middle End Servers. The execution processes are distributed between servers. CPU characteristics and scheduler policies of their servers are different. This is impact to CPU utility, response time and throughput greatly.

2. The main principles of distributed network load balancing

More over network load balancing algorithm uses fully distributed software architecture to maximize throughput and high availability. The identical copies of the network load balancing driver run in parallel on each Middle End servers or load balancing hosts.

Load balancing mechanisms mastermind all hosts on a solitary subnet to simultaneously recognize approaching system activity for the bunch's essential IP address. On each host, the heap balancer goes about as a channel between the system connector's driver and the TCP/IP stack, permitting a part of the approaching system activity to be gotten by the customers. A heap balancer likewise can be utilized to give a virtual IP arrange association with conveying organize asks for over various application occasions [7]. Load balancing process in terms of the stack of layers is represented in Figure 2. This permit to administration organizes activity from customer.

If there should be an occurrence of layer 3 loads, adjusting every server runs a duplicate of an IP-based administration and system stack adjusting circulates the systems administration workload among them. These velocities up typical handling with the g customers see speedier turnaround on their solicitations [7].

Load balancing may be performed at different layers of the OSI model. However, the layer 4 load balancing is normally available free of charge and at the same time have a good characteristics and less complicity than another. The layer 3 stack balancers play out the accompanying capacities: stack equalizations at a basic level and coordinates activity in light of server reaction time; every server has a similar IP address; the slightest occupied server reacts to the substance request.

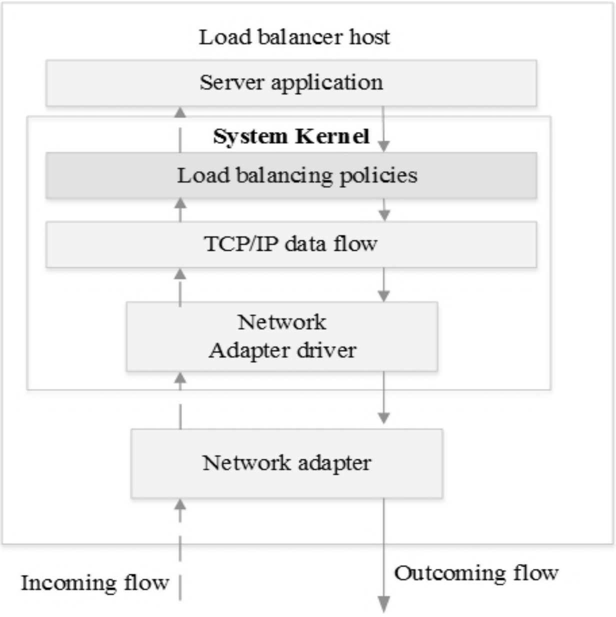

Figure 2 – The place of network load balancer

One of the best-known Layer 3 stack balancers is Microsoft Organize Stack Balancer or NLB [8], this is center of system load balancer programming that is accessible to clients of basic applications. However, the functionality of network load balancing is greatly depending on characteristics of Middle End server platforms.

3. Analysis of network load balancing between server platforms

The network load balancing question should be addressed to the early stages of data center planning. However, this requires an a priori value of server performance. The network load balancing performance can be measured as CPU overhead and throughput [1].

Server CPU utilization is the main criteria that show optimal load distribution within the cloud. The value of CPU utilization in the operating range (70-80%) is directly depends on number of user requests that particular server execute. The CPU utilization impacts to the response time. It increases with the non-overlapped portion.

An analysis of performance of network load balancing based on a network fragment as shown in Figure 3.

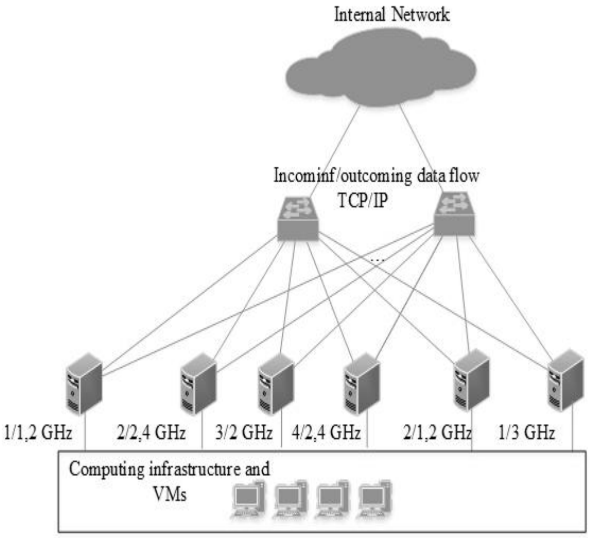

Figure 3 – Fragment of experimental network

Network Load Balancing of Windows Server [8] put as a baseline of experiment. Servers with different technical characteristics take into account. The correct value of the coefficient of proportionality load, which is caused by the difference in the CPU frequency and the number of CPU cores, it is the main task of modeling and analyzing. The assessment of efficiency coefficient is given in a number of technical works [1, 2, 9]. The efficiency coefficient per data center user, greatly depend on the used software, the characteristics of the processor architecture, hardware, etc., but this does not affect the dependence of it change.

Utilization factor of multi-core systems depends of the possible problem to parallelization. In the work assume that software tools enough optimized for parallelization level core/processes (up to 80% of the requests can be distributed between the core/processors) and by increasing the number core utilization factor of the values is given in Table 1 [1].

Table 1 - The characteristics of server platforms

Number of core/processors	The core utilization	Processor clock rate
1	1	$1{,}2*10^9$
2	1,8	$2{,}4*10^9$
4	3,1	$2*10^9$

The formula for the CPU utilization as a function that depends on the number of user requests (*N*), represented as follows:

$$C(N_u) = K_{os} + K_s + \frac{C_{Ni} \cdot N}{F \cdot M} . \quad (1)$$

Where K_{os} - the part of CPU utilization conditioned constant process in operation system (practically does not depend on the number of processed user requests);

K_s - the part of CPU utilization conditioned additional processes on server platform (not depend on the number of processed user requests, but depends on the extra running. processes);

ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.20943/01201702.6771

C_{Ni} - the number of machine cycle time required for processing one N user session;
F - the processor clock rate;
M - the utilization factor multiple cores in a multi-core / multi-processor systems
N_u - the number of users requests.

The calculation of CPU utilization for a uniform network load balancing between six servers with different hardware resources (Fig. 3) in situations where the servers are running as Middle End. The calculation was performed for different number of clients (from 100 to 4000). The dependence of CPU utilization for analyzed servers platforms as a function of number of clients depicted on the Figure 4.

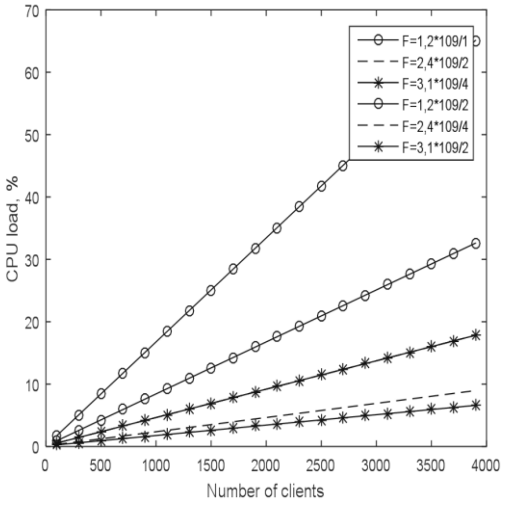

Figure 4 – Server platforms CPU utilization that depended on the number of clients

The obtained results demonstrate that the model is adequate work out the difference in CPU load when the differences in hardware Middle End server platforms (CPU < 80%).

The total throughput of the data center network is dependent on the number of servers that usually represent cluster. Clustering is applying to improve the efficiency of load balancing. Multiple servers combined in a cluster. The load is distributed between them using a set of special techniques called balancing. In addition to addressing the problem of high loads clustering helps provide backup servers at each other.

Efficiency of clustering depends on how distributed (balanced) load between the elements of the cluster. In this way throughput can builds extra customer activity that the bunch can deal with preceding soaking the servers (calculation segments).

For example, the maximum throughput on a single load balancing cluster (*Hi*) depends of amount of server and Virtual Machines (Images) that can process the end user requests. In the event that the end client ask for rate develops with the quantity of hosts, throughput scaling for N maximally stacked group hosts is restricted by development in CPU sifting overhead and other criteria.

The criteria for compliance throughput for a data center network that used distribution load balancing clusters can be the maximum transfer rate (*C*) and probability of service failure (   ), which depends on the CPU overload (*p*) [7]. The maximum throughput per host in an *N*-host cluster can be calculated as:

$$Th(Nh)=\sum C(Ni)\left(1-\left(Nh\left(p\frac{R(N_i)}{Th(Ni)}(1-\rho)\right)\right)\right)$$

Where $C(Ni)$ - maximum transfer rate for
Ni load balancing server platforms;
Nh – the amount of servers or computing nodes;
R(Ni) – the transfer rate;
Th(Ni) – the throughput for Ni load balancing nodes;
$C(N_u)$ – the CPU utilization. The value of server CPU utilization has been calculated in
– the service failure probability. For calculation examples assume that $\rho=0$.

Figure 5 below shows throughput scaling for maximum throughput for the number of load balancing nodes (N=6) that depends of different CPU utilization ($C(N_u)$=0,1 $C(N_u)$=0,3, $C(N_u)$=0,8).

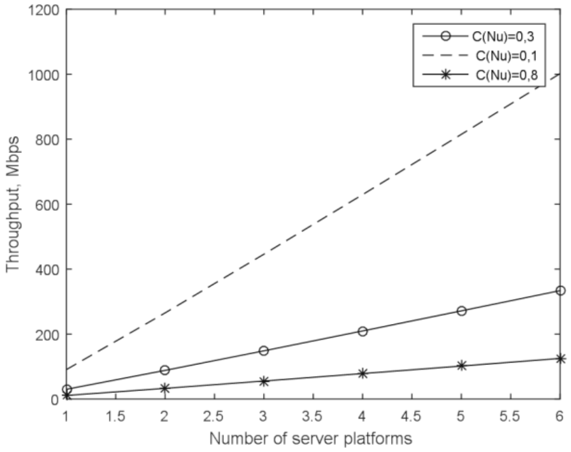

Figure 5 – Maximum throughput for different numbers of servers

As shown on Figure 5, throughput moves off quickly as hosts running at this demand rate are added to the bunch. Nonetheless, if the application devours four times the CPU stack per customer ask for, system stack adjusting separating CPU usage drops to percent, and throughput stays inside 80% of the perfect N-overlay throughput increment for six servers. With a similar server stage and registering hubs, the demand rate would be lower by four-overlay. To keep up the demand rate appeared in the diagram, four-center processor could be utilized.

The analysis shows that such method of network load balancing provides a satisfactory solution for load distribution between server platforms. These principles of network load balancing will be taking into account in future research.

https://doi.org/10.20943/01201702.6771

Conclusion

Load balancing inside data center networks is challenging. The dynamic nature of the request, business demand, differences in the technical characteristics of the computing elements (in particular the server platforms) are the reason of this challenge. The main technical characteristics that affect the network load are the processor CPU and throughput. Formulas for calculating the CPU utilization and network throughput for analysis network load balancing in case of clustering proposed in the work. These formulas allow us to estimate the effectiveness of load balancing between servers. Calculated results showed that the loading of such network structures is within normal limits (CPU < 80%). And more over, the CPU utilization less than 10% in the case when four-core processors with core utilization 3,1 is using.
Satisfactory characteristics were also obtained for evaluating throughput. The analysis of results showed that increasing amount of servers leads to a significant throughput increasing. Thus the proposed structure of the network and formulas can be used in the construction of new data centers.

List of References

[1] W. J. Dally, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann Publisher, 2004, 550 p.
[2] T. Benson, A. Anand, A. Akella, M. Zhang, *Understanding Data Center Traffic Characteristics*. ACM SIGCOMM, 2010, pp. 92-99.
[3]. Cisco. *Evolving to a Unified Data Center Architecture*, July 2007 [Online] Available: at 28.12.12015
[4] Microsoft. *Best Practices for Performance in ISA Server 2004*, USA, 2012 [Online]Available: at 26.12.2015.
[5] An Oracle White Paper. Making Infrastructure-as-a-Service in the Enterprise a Reality, 2012, 14 p.
[6] *Load Balancing Microsoft IIS Deployment Guide*. rev. Microsoft, 2015 [Online] –Available: Microsoft_IIS_Deployment_Guide.pdf at 28.12.2015.
[7]Microsoft. *Overview of Network Load Balancing*, 2015 [Online] – Available: at 28.12.2015.
[8] Microsoft. *Windows Server. Create a network load balancing cluster*, 2015 [Online] – Available: at 28.12.2015.
[9] O. Tkachova, Abdulghafoor Raed Yahya *An Analysis of Load Balancing Algorithms in Cloud Environment* // Proceedings of XIVth International IEEE conference TCSET'2015, 23-26 February 2015, Lviv-Slavske, Ukraine.

Dr-Amer Tahseen Abu Jassar
Doctor of Philosophy in Engineering Sciences / Computer Science, working an assistant professor in the sciences and information technologies "Computer science department" in Irbid National University/ Jordan.