SCALABILITY AND FAULT-TOLERANCE IN DATA CENTER NETWORKS: ARCHITECTURES WHICH ARE PROMISING TO ACHIEVE THESE NETWORK CHARACTERISTICS IN DATA CENTER

Fawad Ahmed, Muhammad Alam, Nazia Shahzad

Abstract — Data centers are considered central hub of information and communication. In last recent years,as application size increased and application became business-critical, services down time became unacceptable and increasing fame of cloud- computing need, forces scientist and developers to explore more efficient data center network architectures to achieve fault-tolerance and scalability. This paper discusses and analysis some data center architectures such as: Fat-tree, Aspen-tree (Modified Fat-tree), VL2,Portland and Dcellat structure level with their characteristics and techniques related load balancing, redundancy and network virtualization for fault-detection and recovery in much lesser time to gain high availability, fault-tolerance and scalability.

Keywords- Fault-tolerance; Scalability; Routing Algorithm; Load balancing; Agility

I. INTRODUCTION

Growing needs of data center services and demand of highly availability appeals for exploring new architectures for data center networks, current trend in design and architecture prevents agility [6], efficient fault-tolerant and scalable mechanisms. Different architectures have different interconnection mechanism techniques. One of the proposedarchitecture: VL2 overcomes these limitationsand achieve agility, uniform bandwidth and perform isolation. Modified multi-rooted fat-tree named as Aspen-tree which have minimum convergence time in fault-tolerance and fault-recovery and removes overheads [7].

anuscript Received:12-24-2016; accepted: 2017; date of current version JUNE 2017

Fawad Ahmed is with IOBM, Korangi Creek, and Karachi, Pakistan
(email:std_18916@iobm.edu.pk)
Muhammad Alam is with IOBM, Korangi Creek, and Karachi, Pakistan
(email: malam@iobm.edu.pk)
Nazia Noor is with Indus University, Pakistan (email: nazia.noor@gmail.com)

Dcell is architecture from Novel with good interconnection technique efficient and promising fault-tolerance mechanism, D cell contains different levels of connection: each server is connected to each level but in equal manner [9].

After describing the comparative analysis of different architectures available, this paper also describes the emerging DCNA like: Portland, which design and implementation overcomes the limitations of layer2 and layer3 existing network protocols in terms of VM migration, inflexible communication, fault-tolerance and scalability [12]. This paper also tabulated data center architectures in comparison matrix: which helps in better understanding regarding fault-recovery and also in scalability, which are the most challenging concerns in future data centers.

II. DATA CENTER NETWORK ARCHITECTURE AT GLANCE

The most hierarchical data center architectures prevent the technique of services agility which describes any server for any services, which increases cost and risk in several ways [6]. Like existing architectures, rely on tree-based structure built from expensive hardware impact the cost and scalability. To address these limitations in conventional design for large scaleVL2 architecture proposed. This architecture achieves the objectives like: Uniform high capacity among network levels, Performance isolation between servers and virtual Layer-2 meaning [6]. VL2 also provide advantage to data center operators in terms of server pool fragmentation over bandwidth constraints. Finally, evaluation of the performance of the VL2 architecture through macro and micro level experiments, which produces remarkable results in terms of throughput, Scalability, and robustness and of course agility management.

As data centers are scaling because of cloud services, which help large organizations, groups or enterprise to highly available its businesses and make it disaster recoverable. Networking requirements, related to size of data centers, methods for network virtualization and methods for building physical networks [5]. Mainly the approaches of building physical networks such as : MPLS, L3 Switching and Fabric Switching,these

 ISSN 1816-613X

methods supports for virtual networks for large-scale DCNA, as density of server machines and storage devices increases so as the ultra-high speed interfaces will as be the requirement. Traffic mechanism is also main concern between VMs and switches. Lastly to implement large-scale data center research is still continue to make its operation stable and smooth.

As advances in applications and data-intensive computing compelling system designers and developers in exploring new ways of interconnection in data center for fault-tolerance and scalability after comparison of different available architectures categories: Switch-only, Server-only and hybrid across the matrices of scalability, fault-tolerance, path diversity, cost and power [8]. Different architectures have different interconnection techniques and mechanism. Concluding, designers and data center operators must optimize the characteristics of the available architecture before deployment and implementation in data center.

A modification in multi-rooted fat-tree based on fat-tree architecture, which is named as Aspen tree. Aspen tree topology addresses the minimum of convergence time at failures, network sizes, fault tolerance at local hops to avoid message overhead of alternative paths [7]. As critically known, for high availability there is 5-nine (99.999) percent allowed downtime in data center, means 5 min per year [7]. Aspen tree added the concept of redundant link, which leverages fault-tolerance at local link. Consequently, as evaluated Aspen tree reduces drastically the convergence time and overheads, improve fault tolerance and scalability.

The goal of data center network is to maintain connectivity between data center servers, and provide efficient and fault-tolerant routing mechanism. The study of saving power in large-scale DCNA with number of network devices in a routing perspective know as energy aware routing [2]. Main idea is to use less network devices to provide the routing service, without sacrificing the network performance, the network devices are not being used can be shut down or put into power saving mode for energy saving. Model of energy aware routing and design heuristic algorithm show data on simulation, which shows that routing can effectively save power consumed by network devices when the network load is low [2].In near future, high density data centers will be able to consume less energy over routing in low network load period.

Data center performance is based upon the load balancing of network traffic in real time. To achieve agility and provide high bandwidth between servers hosted, load balancing is essential in scale-out topology based on data center networks. It has been investigated that two load balancing schemes flow-level VLB: deals with highly volatile data center network traffic or works better traffic flow is uniform, while packet-level VLB: which performs better than flow-level in non-uniform traffic flow or large flow. To overcome the related problems in previous mentioned schemes, there proposed two advanced load balancing schemes named: Queue length directed adaptive routing and Probe- based directed adaptive routing these two schemes performs similar with different traffic pattern in different experimental settings. The result indicates or concludes that proposed schemes are best performer than flow-level and packet-level load balancing [4].

As cloud computing causes the creating of large-scale data centers, which interconnects millions of servers, routers, switches and storage devices that provides highly available distributed services to customers**,** partners or group of enterprises. There must be a need of data center centralized management system which maintains availability of services and helps in network issues and troubleshooting as well as network SLAs. Pingmesh: a centralized system for large scale data center which addresses last concerns. Pingmesh is being used by Microsoft for certain years and also by network software developer and application developers and operators. Latency and packets drop are main issues in maintaining any network quality, especially in large-scale data center where seconds of downtime matters. Addressing these issues there are many data center centralized management system being used: Autopilot, Cosmos, SCOPE etc., while Pingmesh has built over Autopilot framework. Pingmesh measures and calculate network latency distribution between servers, network devices and applications in large-scale data center using visualized patterns. Pingmesh, among its properties and advantages it has some limitations, helps in improving network quality and fixing black holes and packet drops in data centers networks [3].

The increasing need of fault tolerant data centers with scaling properties also increased the research efforts in exploration of new topologies routing protocols and centralize manageable systems to improve the fault tolerance and scalability. Mainly F10, a novel topology, a modified Fat-tree architecture with good fault-recovery properties, load balancing, rerouting techniques [10] has been discussed and evaluated. Results show that F10, contributes 30% of improvement in making DCN a robust, fault-tolerant and scalable [10].

 ISSN 1816-613X

As number of services and servers are increasing in data center so there must be a need of efficient architecture. DCell, a novel network structure with good interconnection technique,efficient and promising fault tolerant and scalable mechanisms, contains different levels of connection and each server is connected to each level with multiple links but act equally. DCell uses fault tolerant routing algorithm DFR, which is without global state & support shortest path first routing [9]. Results show that DCell is suitable interconnection structure for data center networks at scale.

Using of commodity switches (inexpensive switches) in place of high-end switches to maintain bandwidth and control bottleneck. Mainly, specifies tree-base topology (Fat-tree), its architectural design, routing algorithms, oversubscription ratio, two-level routing, fault tolerance mechanism and scalability. As cluster size increase which limits port density in high-end switches incurring high cost so commodity switches can provide scalable bandwidth to large scale cluster at lower cost [11]. Consequently we can say that commodity switches may also perform better in data center for communication. Using commodity switch also effects the scalability (becomes fixed 24, 48, and 63 as per switch ports) and also fat-tree faces single node failure [13].

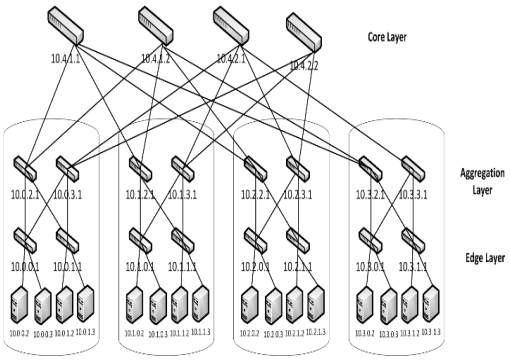

Figure 1: Traditional Data Center Architecture

III. PROBLEM STATEMENT

In data Center networks, to achieve high performance, availability and efficiency, fault-tolerance and scalability are significant concerns. But currently, the data center network architectures more or less help in scaling and failure recovery to make services available, while as cloud-computing is becoming famous, services and application requirements are growing now it is a challenge to find more advance and robust architecture schemes those promises for fault-tolerance and scalability.

IV. SCOPE OF CURRENT RESEARCH

There are several concerns in data center networks architecture, while fault-tolerance and scalability considers the most important. How DCNA effect these characteristics while achieving high performance and network operations continuity? As emerging cloud computing causing large-scale DCN and high uptime of critical applications and services compels the designers and developers to search for more advance schemes and techniques to achieve fault-tolerance and scalability in data centers.

V. TRADITIONAL DCNA

The traditional architectures are no more the requirements as cloud computing services are growing. As Figure 1 shows, there are several short comings in traditional architecture in terms of fault-tolerance, single node failure, scalability, bandwidth bottleneck from lower to upper layer, mixed hardware in environment and expensive switches installation in network and to increase the processing capacity of devices requires power consumption, and also causes oversubscription, which opens the way for alternative architectures.

VI. MODERN DCNAs WITH COMPARATIVE MATRIX

In this section we will define and analyze the comparison of different architectures of data center networks.

a. Fat Tree Architectures

The architecture has been proposed by Al-Fares [11].Fat tree is promoted as an effective DCN architecture and it usestructured commodity switches to provide more uniform bandwidth across each network level and also control oversubscription. Fat-treestructure has(n) pods and each pod contains (n) servers and (n) switches which are organized in two layers of (n/2) switches. Every lower layer switch is connected to (n/2) hosts in each pod and (n/2) upper layer switches (making aggregation layer) of pod. There are (n/2)2 core switches which in turn connect to one aggregation layer switch in each of pods. Below Figure 2 explains:

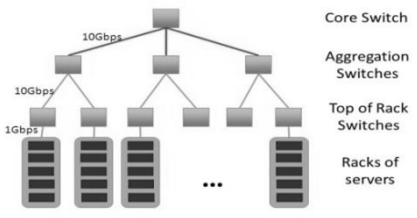

Figure 2: Fat-tree Data Center Architecture [13]

ISSN 1816-613X

i. Scalability

Fat tree scalability is fixed due to commodity switches available ports 24, 36, 48, 64 which limits host size support : 3456, 8192, 27648, 65536 respectively. Cost to scaling and achieving uniform bandwidth among thousands of nodes is significant

ii. Fault tolerance

Fault tolerance in fat-tree is achieved due to redundant links or path among multiple network devices. A failure broadcast update switches to bypass the faulty link. Each switch maintains bidirectional detection session [11] with other switches get failure information. Failure can be in two areas: (a) between lower- and upper-layer switches inside a pod and (b) between core and an upper-level switches. Clearly, the failure at each level of hierarchy continue its operation due to BDS updated and redundant paths.

b. Aspen Tree Architecture

Aspen tree, a modified version of fat-tree illustrated in Figure 3, form a multi-rooted hierarchy containing n-levels, K-port of switches and host connected with leaf switches. Aspen tree has denser network link, it may vary level to level in the hierarchy [7]. Density of links at each level maintains fault tolerance and minimize convergence time, using fault tolerance vector (FTV), but effects scalability.

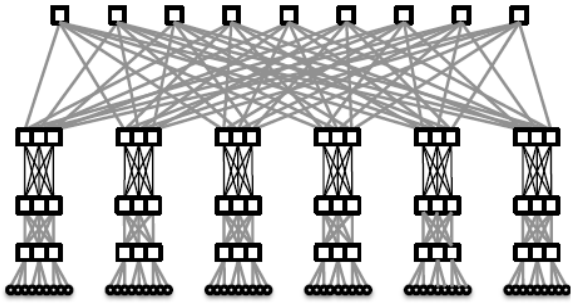

Figure 3: Aspen-tree FTV=<0, 2, 0> [7]

i. Scalability

Scalability is effected due to denser links for achieving fault tolerance. As redundant link will add at each level reduces the number of host supported by the architecture [7]. Below equation describe the number of host supported by aspen architecture, where S is number of core switches at core layer, and Ports of lower level switch and DCC is the number of path between switches That is:

Hosts=k/2xS =$k^n/2^{n-1}$x1/DCC……eq (1)

Fault Tolerance		S	Switches	Hosts	Hierarchical Aggregation			
FTV	DCC				L_4	L_3	L_2	Overall
<0,0,0>	1	54	189	162	3	3	3	27
<0,0,2>	3	18	63	54	3	3	1	9
<0,2,0>	3	18	63	54	3	1	3	9
<0,2,2>	9	6	21	18	3	1	1	3
<2,0,0>	3	18	63	54	1	3	3	9
<2,0,2>	9	6	21	18	1	3	1	3
<2,2,0>	9	6	21	18	1	1	3	3
<2,2,2>	27	2	7	6	1	1	1	1

Figure 4: All possible 4-level 6-port Aspen-tree [7]

ii. Fault tolerance

Aspen tree has efficient fault tolerance using FTV technique, at each level containing denser links. Aspen tree also reduces the convergence time of fault detection and message update among switches.

c. VL2Architecture

This architecture has been proposed by Greenberg [6], it is also fat-tree based with difference of connection of servers in virtual layer 2 located in same LAN [13]. VL2 uses Valiant Load Balancing (VLB) for routing load balancing and also implements ECMP (Equal Cost Multi-Path Routing) to forward data over multiple optimal path to resolve address redistribution in VM migration. The provision of path diversion promises many concurrent problems such as: agility, oversubscription. VL2 requires directory service and server agents for routing mechanism which are location addresses LA and application addresses AA. Where LA separates server name from location for agility and assigned to all switches and interfaces, while AA are only used in applications.

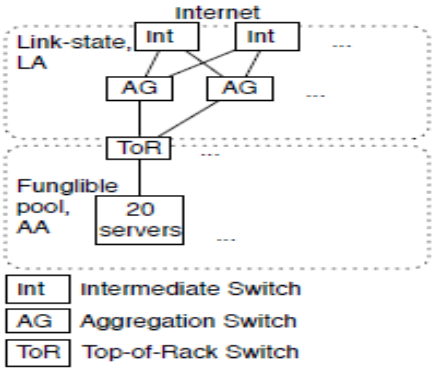

Figure 5: VL2 Data Center Architecture [1]

ISSN 1816-613X

i. Scalability

VL2 is widely used among existing DCNA for large scale however some research and development work is under process to improve its network reliability but still it hasscalability and single node failure issue.

ii. Fault-tolerance

To analyze the fault tolerance VL2 design, failure logs collected for over a year from eight production data centers that comprise hundreds of thousands of servers, host over a hundred cloud services and serve millions of users [6]. Analysis based on hardware and software failures of switches, routers, load balancers, firewalls, links and servers using SNMP polling/traps, syslog, server alarms, and transaction monitoring frameworks. In all, 36M error events from over 300K alarm tickets collected [6].

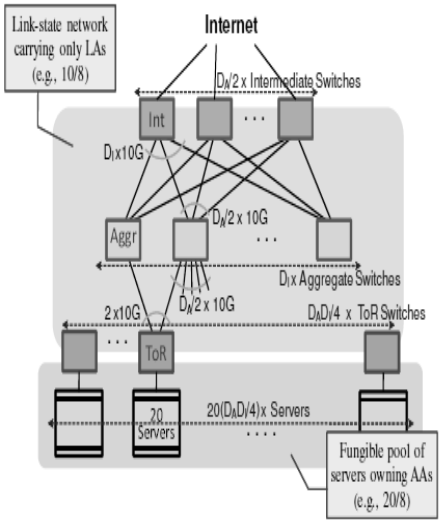

Figure 6: VL2 AAS and LAs illustration [6]

What is the pattern of networking equipment failures?

If communication is unable to response till 30 seconds considers a failure in network. As described in [6],pattern of failures : 50% are small size failure in which less than 4 devices fails, while in 95% failures less than 20 devices fails. Whereas large size failures occur rarely which involve failure of 217 devices.

What is the impact of networking equipment failure?

Traditional architecture apply 1:1 redundancy at higher level to improve reliability, as analysis carried out in [6] shows 0.3% failures are related to redundant network devices or links, which may be because of misconfiguration, bugs faulty components. Still there is no way to eliminate all failure from topmost layer, but VL2 approach somehow muted failures at highest level.

d.Portland

Portland architecture, illustrated in Figure 7, has been proposed by Niranjana Mysore [12], it is also fat tree based, but uses a centralized system, named as fabric manager. Which helps in fault detection on large scale, recover, update the related switches regarding fault. It uses Pseudo-MAC (PMAC), a 48-bit address, to detect a location of device in mapping of certain device MAC address. It deploys new routing mechanism for data forwarding, and supports better fault-tolerance with VM migration and network scalability but risk of single node failure still exist [13].

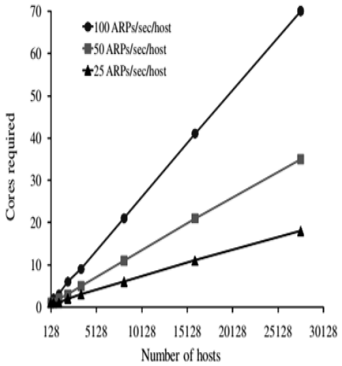

Figure7: Portland Data Center Architecture [12]

i. Scalability

Portland designs have scalability issue for fabric manager in large scale DCN. Figure8 shows the amount of ARP traffic handle by fabric manager and as number of hosts in architecture increase with the cores required to handle the request by fabric manager [12].

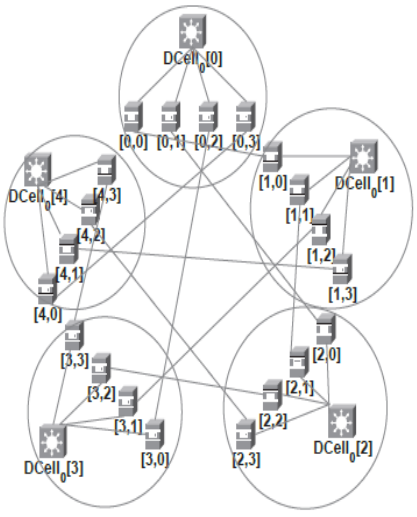

Figure 8: CPU requirements for ARP Request [12]

ISSN 1816-613X

The number of ARPs transmitted per host increases the traffic which also requires CPU time to handle request. As 25 ARP/sec of single host considers extreme in today's DC environment [12] so, Portland's scalability is a challenge in terms of address resolution, routing and forwarding.

ii. Fault tolerance

Primary goal of Portland's fabric manager is to detect and recover failures, for this it deploys LDP (location discovery protocol) to detect and monitor communication session of link or device for a certain period of time. If it sees any failure then update fabric manager, which further update related switches about the failure and then local switches update their routing table to bypass link

e. D cell Architecture

D Cell is lie under server-based architectures which is recursively defined structure introduced by Guo [9], in which many low level D cell units combine and form high level D cell units and at the same time connected with each other. For data forwarding it uses distributed routing algorithm, and as compared to tree based architecture it contains mass redundancy of links for higher band width. It has better capacity of services than traditional tree based architecture. Furthermore, D Cell can be incrementally expanded and a partial D Cell provides the same appealing features.

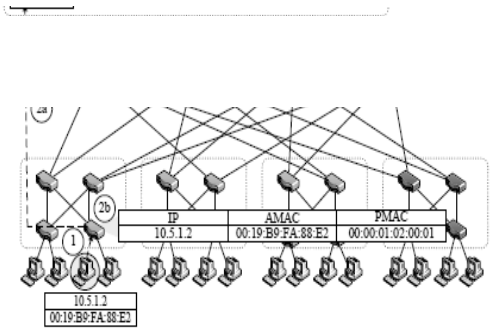

Figure 9: DCell Data Center Architecture [13]

i. Scalability

DCell scales as the node degree increases. It must physically interconnect hundreds of thousands or even millions of servers at small cost it has to enable incremental expansion by adding more servers into the already operational structure. The number of servers scales exponentially, Where number of servers in a D Cell 0 is 8 (n=8) and the number of server ports is 4 (i.e., k=3) -> N=27,630,792 [9]

$(n+1/2)^{2k} - ½ < N < (n+1)^{2k+1} -1$……eq (2)

ii. Fault tolerance

Due to multiple redundant links and efficient recursive structure Dcell has good fault tolerance. Its distributed routing protocol performs near shortest path routing even in presence of severe failures [9]. Both redundancy links and robust routing mechanism makes Dcell ideal for large scale DCN. VL2 partitions the bisection bandwidth.

$N/4\log_n N$…… eq (3)

VII. RESULTS, DISCUSSION AND COMPARISON

The comparison of traditional architecture with other modern DCN architectures in fault tolerance and scalability criteria shows that compared architectures have more or less good or medium level fault tolerance and scalability with some limitations at architecture level.

Fat tree, Aspen-tree, VL2 and Portland are tree based structure. As discussed fat tree overcome the oversubscription, single node failure, uniform bandwidth but with limited scalability. While Aspen-tree architecture addressed, using denser redundant links, to minimize convergence time of failure detection and recovery at each level of structure and used FTV for making efficient fault tolerance but due to complexity of redundant link at each level scalability affected in aspen-tree. At large scale VL2 give idea of service agility and pointed out the bisection bandwidth at topmost level of hierarchy, but still VL2 has single node failure issue. And Portland, being part of tree based structure using a centralized system which automates failure detection at scale.

Dcell is recursive based architecture; it used distributed algorithm and redundancy of link for higher bandwidth to maintain scalability and fault tolerance. Below table mention the results in matrix of discussed structures

DCNA	SCALABILITY	FAULT TOLERANCE
FAT TREE	Medium	Medium
ASPEN TREE	Medium	Good
PORTLAND	Good	Medium
VL2	Medium	Medium
DCELL	Good	Good

Table 1: Scalability and Fault Tolerance Matrix [13]

 ISSN 1816-613X

VIII. CONCLUSION

The architecture must be scalable and fault tolerant, using a relatively small interconnects of switches to connect as many end hosts as possible. It should provide as much bisection bandwidth as possible in support of all-to-all communication. As enhancement is in process at architecture level and as well as network hardware level to fulfill future data center requirements. Reactions to architecture changes should happen as quickly as the hardware will allow. In particular, failures information must be updated without broadcast.

IX. OPEN PROBLEMS AND FUTURE WORK

The field of data center networking is still a young area, and new and innovative ideas appear continuously. Here, we discuss the scalability and fault tolerance as well as ideas for ongoing research in the areas of data center enters network topology, communication and efficiency.

ACKNOWLEDGMENT

The authors would like to thank Institute Of Business Management-IOBM,Korangi Creek, and Karachi, Pakistanfor their support in the completion of this research work.

REFERENCES

[1] Juha Salo, "Data Center Network Architectures."
[2] Yunfei Shang, Dan Li, Mingwei Xu, "Energy-aware routing in Data Center Network."
[3] Chuanxiong Guo, Lihua Yuan, Dong Xiang, Yingnong Dang, Ray Huang, Dave Maltz, Zhaoyi Liu, Vin Wang, Bin Pang, Hua Chen, Zhi-Wei Lin, Varugis Kurieny Microsoft, yMidfin Systems, "Pingmesh: A Large-Scale System for Data Center Network Latency Measurement and Analysis."
[4] Santosh Mahapatra Xin Yuan, "Load Balancing Mechanisms in Data Center Networks."
[5] Tatsuhiro Ando, Osamu Shimokoni, Katshuhito Asano,"Network Virtualization for large-scale Data Centers."
[6] Albert Greenberg, James R. Hamilton, Navendu Jain, Srikanth Kandula Changhoon Kim, Parantap Lahiri, David A. Maltz,parveen Patel, Sudipta Sengupta, "VL2: A Scalable and Flexible Data Center Network."
[7] Meg Walraed-Sullivan Redmond, Amin Vahdat, Keith Marzullo," Aspen Trees: Balancing Data Center Fault Tolerance, Scalability and Cost."
[8] Fan Yao, Jingxin Wu, Guru Venkataramani, Suresh Subramaniam,"A Comparative Analysis of Data Center Network Architectures."
[9] Chuanxiong Guo, Haitao Wu, Kun Tan, Lei Shiy, Yongguang Zhang, Songwu Luz, "DCell: A Scalable and Fault-Tolerant Network Structure for Data Centers."
[10] Vincent Liu, Daniel Halperin, Arvind Krishnamurthy, "F10: A Fault-Tolerant Engineered Network."
[11] Mohammad Al-Fares Alexander Loukissas, Amin Vahdat, "A Scalable, Commodity Data Center Network Architecture."
[12] Radhika Niranjan Mysore, Andreas amboris, Nathan Farrington, Nelson Huang, Pardis Miri, Sivasankar Radhakrishnan, Vikram Subramanya, and Amin Vahdat, "Portland: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric."
[13] Han QI, Muhammad Shiraz, jie-Yao LIU, Abdullah Gani, Zaulkanain Abdul Rahman, University of Malay kaula lampur 50603, Malaysia) Email: hanqi@siswa.um.edu.my, abdullah@um.edu.my : Data Center Network Architecture in Cloud computing: Taxonomy and Review.
[14] Costin Raiciu†, Christopher luntke†, ebastien Barre‡, Adam Greenhalgh†, Damon Wischik†, Mark Handley† †University College London, ‡Universite Catholique de Louvain: Data Center Networking with Multipath TCP.

 ISSN 1816-613X