International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Comprehensive review of datacenter architecture evolution

Abhishek Jayakar Shetty[1], Ganashree K. C.[2]

[1] B.E. Department of Computer Science and Engineering, R.V. College of Engineering
[2] Assistant Professor, Department of Computer Science and Engineering, R.V. College of Engineering

---

Abstract - *Today, data centers have become an important part of the information-centric world. They have evolved from large and power-hungry infrastructure to far more power-efficient entities. They essential for business continuity and with an increasing need for processing and storage due to the exponential increase in the data produced and information created. They house redundant computing, storage and network components to keep the infrastructure up and running. The emphasis is currently on making these data centers more power efficient and allowing better resource usage to improve the overall system throughput. Datacenters have evolved from bulky mainframes to servers with centralized storage to converged and hyper converged infrastructure. Each stage in this evolution makes up for the shortcomings of its previous generation but comes with its own sets of advantages and disadvantages. Datacenters have evolved from massive mainframes to centralized storage servers to converged and hyper-converged infrastructures. Each stage in this evolution addresses its preceding generation's shortcomings but comes with its own set of advantages and disadvantages.*

Key Words: **3 tier architecture, virtualization, rack, hypervisor, converged infrastructure, hyperconverged infrastructure, Storage Area Networks, Direct Attached Storage**

1. INTRODUCTION

With the everchanging landscape of data consumption and to better optimize and organizations resources, every datacenter architecture has been improving upon its preceding generation. In the initial days systems were optimized to ensure high availability and throughput leading to creation of mainframes. Later when focus shifted to optimizing capital and organizational expenses standalone servers with less expensive components came into being. When organizations needed to improve the utilization of a datacenter resource a shared pool of storage was introduced leading to centralized storage architecture.

Server virtualization completely changed how datacenter operates and led to the most efficient resource utilization and increased its cost effectiveness. Further improvements such as converged and hyper converged infrastructures have led organizations to decrease their deployment time and ease the maintenance of datacenter. In this paper all these architectures have been discussed along with their respective pros and cons and improvement it brings over the preceding architectures.

2. MAINFRAMES

Mainframes were one of the earliest forms of bulk computing. They have been around since the 1950s. A mainframe is a centralized computing system in an organizations data processing center and much less powerful devices such as workstations are linked to them. Earlier mainframe systems were housed in large metal boxes taking up a bulk of room space. Over the years they have been reduced to size of a refrigerator. The focus of a mainframe is throughput, reliability and scalability. Thus, a mainframe has relatively larger number of I/O (Input/Output) units than compute units. This is because mainframes were designed to handle large number of transactions in bulk rather than fewer computationally intensive tasks [1]. In addition to the general-purpose CPUs (Central Processing Unit) mainframes house at least one SAP (System Assistance Processor. The SAPs are invisible to the Operating Systems and applications and assist in bulk data computation by leveraging the multiple channels available for the data transfer. They contain memory capable of in memory error correction improving the yield and reliability of high-density memory design [2]. Storage is built into the mainframe itself. Mainframes are considered to be highly reliable systems because of high degree of internal redundancy built into them. These redundancies are achieved by having multiple units of critical components such as CPU and storage to nullify the effect of any hardware failures. Due to this reliability and ability to execute large number of transactions they are still used to run mission critical applications in areas like order processing, handling financial transactions, airline ticket reservation. Mainframes are also easy to vertically scale up by adding more compute resources if the to meet the organization needs. Major disadvantages of mainframe systems are the high cost due to high quality components and high number of these components used to ensure high availability. They also need special operating system to function and they are complex systems requiring highly trained staff to operate thus increasing

organization operational expense. They also create a highly siloed environment due to their inability to be accessed over a network.

3. STANDALONE SERVERS

To overcome some of the drawbacks of using a mainframe, standalone servers came into existence. These servers consisted of CPUs for compute, ECC (Error-correcting code) memory and an internal DAS (Direct Attached Storage). This type of storage is directly attached to the server internally avoiding the need for accessing storage over a network. Few protocols used by these DAS devices are ATA (Advanced Technology Attachment), SATA (Serial Advanced Technology Attachment), SCSI (Small Computer System Interface). These servers are much less expensive than the mainframes because they are not designed to be highly available, thereby do not need redundant high-quality components. But some components such as dual PSU (Power Supply Unit), RAID (Redundant Array of Independent Disks) Disks and ECC memory increase the fault tolerance of the system. These servers are more flexible than mainframes because they can be accessed over a network thus are not as siloed as mainframes. Standalone servers serve as a relatively low-cost deployment solution for their mission critical and business critical applications while still ensuring minimum downtime. Major disadvantage of the standalone servers is low / unequal resource utilization since and need for multiple servers for multiple applications, say a separate server for a database and to serve web requests. This eventually increases capital and operational expenses for the organization. Since the storage is directly attached to the server itself, this is inaccessible by other servers creating a siloed environment and inefficient usage of storage resources. Standalone servers also have a single point of failure for both compute and storage. Figure 1 shows a typical datacenter with standalone servers.

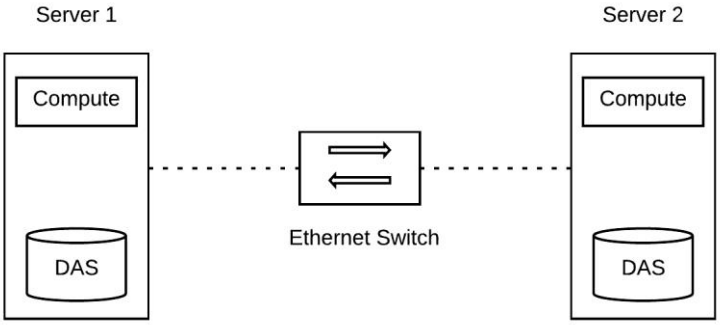

Fig –1: A datacenter with standalone servers

4. SERVERS WITH CENTRALIZED STORAGE

A centralized storage paved way for information centric storage approach for data center, this overcame the limitation of a standalone server where data sharing was not possible due to the local nature of the storage. A pool of centralized storage with multiple RAID Disks is served to the server by means of a storage controller or an array of storage controllers to ensure redundancy also called as external DAS (Direct Attached Storage). The servers may also be attached to a NAS (Network Attached Storage) appliance via an ethernet NIC (Network Interface card) using protocols such as NFS (Network File System) and CIFS (Common Internet File System A NAS system is usually a special-purposed device designed to provide clients with files on a LAN [3]. Technologies such as SAN (Storage Area Networks) also provide a centralized storage system with a need for special storage switches and a Fiber Channel (FC) connectivity [4]. This implementation of datacenter considerably increases storage utilization and increases data sharing among the compute servers. This arrangement also has considerable disadvantages. A centralized storage adds more complexity to the architecture when compared to a DAS. This means a separate highly trained staff is needed to maintain the storage. Also, this is relatively expensive than standalone servers because of high cost of specialized FC switches and fiber cables. This central system also serves as a single point of failure thus cutting access to all the compute server in case of a catastrophe. Having multiple storage controllers and switches to increase the availability also adds to the capital expenses of organization. Most significant drawback would be the latency of I/O (Input / Output) operations since they are now happening over the network rather than through local buses in case NAS systems and other IP based protocols such as iSCSI (SCSI over internet). The servers also need a specialized HBA (Host Bus Adapter) to connect to the FC switches. Even though storage utilization was improved, improving compute resource utilization still remained a challenge in this datacenter architecture. Figure 2 shows a datacenter architecture of servers with centralized storage.

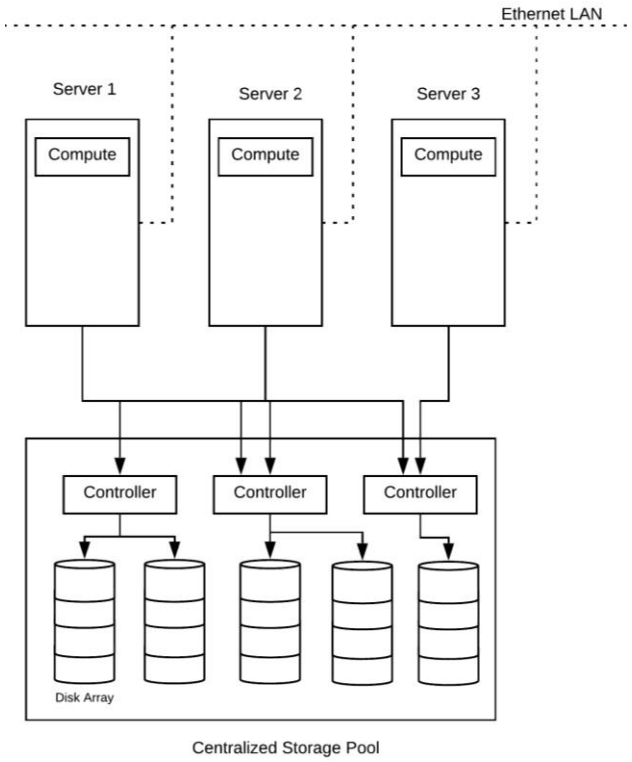

Fig –2: Servers with centralized storage pool

5. SERVER VIRTUALIZATION

After solving the problem of inefficient storage utilization, the focus shifted to improving consumption of compute resources such as CPU and memory. Also, multiple physical servers consume a significant amount of rack and datacenter space also subsequently more power. This adds to the capital expenses and a larger team to manage the servers adds to the operational expenses. Server virtualization addresses the above issue by abstracting the physical hardware and creating virtual instances of the server called virtual machines using a special piece of software called hypervisor running directly on the hardware (Type 1 Hypervisor) or running on a Host Operating system running on the hardware (Type 2 hypervisor) [5]. Hypervisors, also known as VMM (Virtual Machine Monitor) abstracts the underlying hardware and provisions it to the Virtual Machines. The hypervisor is transparent to these machines and acts as a resource allocation manager. This isolation creates a layer of security between the virtual machines and the host and between the virtual machine itself. The hypervisor has full control over the compute and storage resources and distributes it among the virtual machines based on the demand. They also claim back the allocated resource once the virtual machines cease to exist. Virtualization not only creates an isolated environment for the virtual machines but also enables aggregation of multiple hardware instances into one and sharing a single hardware instance among multiple virtual machines. Server virtualization in the end results in better utilization of compute and storage resources and also provides significant energy savings by eliminating the need for larger number of physical servers [6]. The drawbacks of server virtualization include higher cost of implementation and operational expense due to requirement of skilled personnel to handle the virtualization. It also creates a scalability issue since the resources are shared among the virtual instances any spike in virtual machine resource utilization may lead to failure if enough resources are not reserved. Still the traditional 3 tier architecture remains complex due to presence of multiple links, centralized storage forming a single point of failure, separate maintenance teams for compute, storage and network, scaling expense when space in a constraint and lower return on Investment. Major drawback of this 3-tier architecture being long deployment time which may go up to weeks or months Figure shows a traditional 3 tier server virtualized datacenter environment where the compute, storage and the network components can be physically distinct (vendor agnostic) and managed separately from each other. Figure 3 shows a typical 3 tier datacenter architecture.

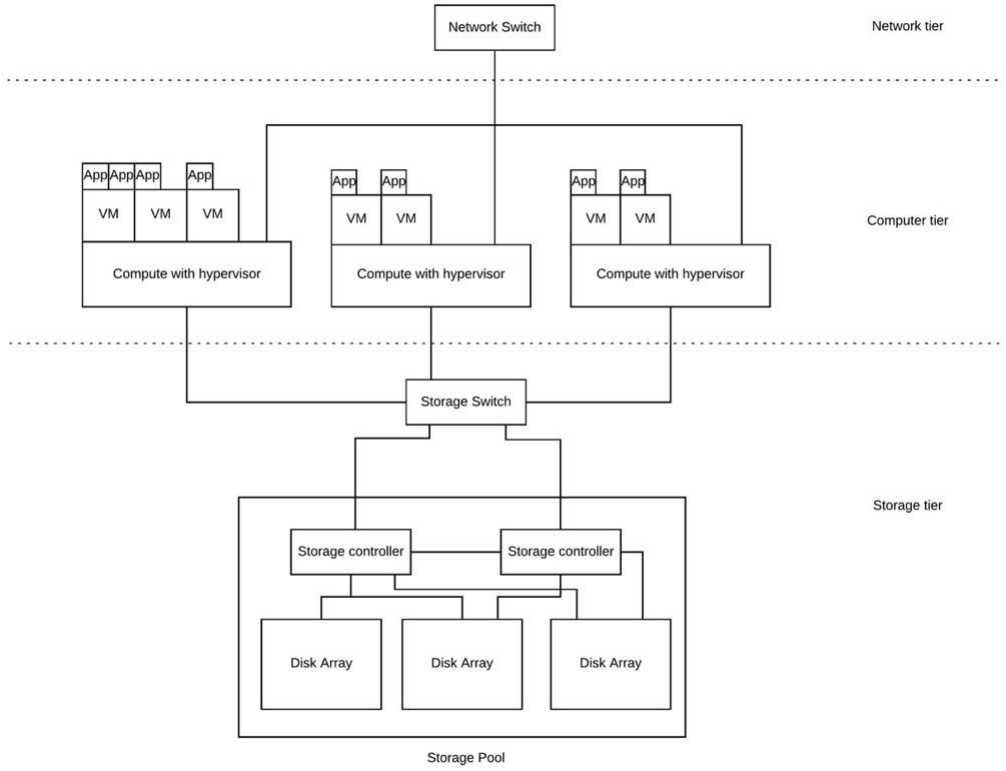

Fig –3: Traditional 3 tier server virtualized datacenter architecture.

6. CONVERGED INFRASTRUCTURE

The server virtualization solved the problem of inefficient compute and storage resource utilization. But the issue of infeasible upscaling of infrastructure and slower deployments remained. Converged infrastructure addresses these issues but converging compute, storage and network infrastructures in a datacenter into a rack implementation. It must be noted that the storage here is directly attached to the physical server and not shared among the multiple compute nodes. Here the converged components are physically discrete but have a single control plane. The advantage of this infrastructure is faster deployment due to lesser cabling and assembly and easier management. Also, since the constituent components are still physically discrete it is possible to scale them up independently. Horizontal scaling can be done by adding more components to a rack and vertical scaling by adding additional rack units. The drawback of this infrastructure is it leads to vendor lock in, which can result in fewer functionalities and limited options for customization. Also, adding new components to an already existing converged infrastructure is a complex and expensive process. Converged infrastructure solutions are largely adopted by larger enterprises moving from a traditional 3 tier datacenter architecture. Even though it requires specialized staff to manage the infrastructure it scales out horizontally very well. Figure 4 shows constituents of a rack in converged infrastructure.

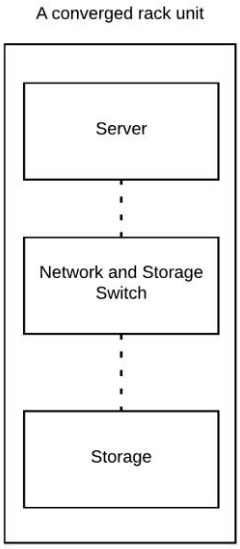

Fig – 4: Rack constituents in converged datacenter infrastructure

7. HYPERCONVERGED INFRASTRUCTURE

Hyperconverged and converged infrastructure share same goals but differ in the implementation. The major distinction between two solutions is hyper converged infrastructure is a software defined solution and are made of software building blocks rather than physical hardware. In hyperconverged solution compute and storage are combined into a single solution and have a single control plane whereas network still remains physically discrete and has a separate management plane. The important feature of an HCI (Hyper Converged Infrastructure) is the shared storage. Even though the disks are directly attached to each compute server it is shared among multiple servers in a distributed fashion [7]. In some solutions the software controllers are connected to a SAN. The raw disk capacity is not always the available disk capacity as the system maintains multiple replicas of data based on a pre-configured replication factor. This is done for disaster recovery and enable high availability in case of any node or disk failures. Software defined techniques such as compression and deduplication can be implemented to increase the overall available storage. Being a software defined storage solution, the storage controller runs as a service in every single node. Hyper Converged Infrastructure is built on commodity hardware which can be replaced easily and being a software defined solution addresses the vendor lock-in problem of converged infrastructure solutions. Keeping storage close to the compute and distributing its access over the group of compute servers not only keeps the storage IO latency in check but also increases storage utilization. The flexibility of adding any vendor hardware and simple management added with much faster deployment (in hours) attracts various small to mid-level enterprises. The infrastructure scales out linearly just by adding a new node (computer and storage) to the existing cluster (group of nodes). But horizontal scaling is expensive since a whole new node has to be added even if you want to scale up just compute or storage capacity, if the Operating System licensing is per CPU core this is a disadvantage to customer who only needs storage. Due to multiple management software the resource consumption overhead exists in Hyper Converged Solutions. Figure 5 shows a sample hyperconverged cluster.

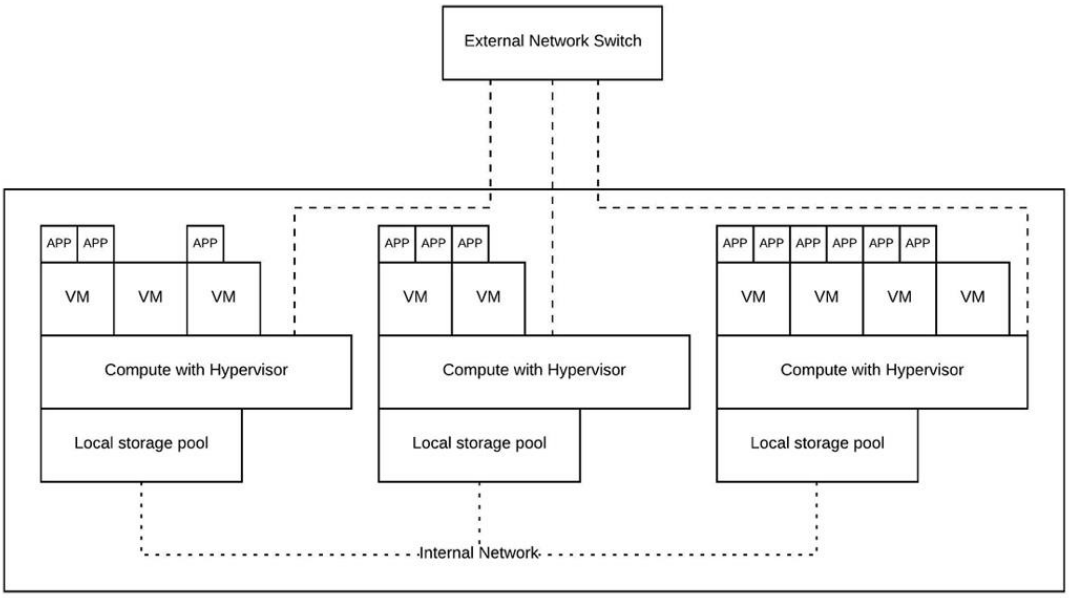

Fig –5: A three node hyperconverged cluster

8. CONCLUSIONS

This paper presents the important distinctions between popular datacenter architectures. Even though datacenter architecture has evolved over the period of time the older architectures are not outdated. Choice of an architecture depends on the organizations need. Even though mainframes were the earliest form of system they are not obsolete. They are used extensively in processing banking transactions and in airline reservation systems. Smaller organizations which want to quickly deploy their solutions at low cost and who value the privacy of their data thus instead choose a converged or a hyperconverged infrastructure solution over a traditional 3 tier architecture. When data is the focus of an organization servers with centralized storage is chosen in order to provide high availability and security to the data with a central management.

ACKNOWLEDGEMENT

I would like to extend my gratitude towards my guide Prof. Ganashree K. C. for her continued support and guidance on my work in this paper. I would also like to show my gratitude to Department of Computer Science and Engineering, RV College of Engineering for giving me the opportunity to write this paper with complete support.

REFERENCES

[1] D. Huang and Z. Gao, "Design of an education cloud based on mainframe," ICSSSM12, Shanghai, 2012, pp. 167-170, doi: 10.1109/ICSSSM.2012.6252214.

[2] M. Lv, H. Sun, J. Xin and N. Zheng, "Efficient Repair Analysis Algorithm Exploration for Memory with Redundancy and In-Memory ECC," in IEEE Transactions on Computers, doi: 10.1109/TC.2020.2996747.

[3] De-Zhi Han, "Snins: A Storage Network Integrating NAS and SAN," 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 2005, pp. 488-493, doi: 10.1109/ICMLC.2005.1526995.

[4] Chao-Yang Wang, Feng Zhou, Yao-Long Zhu, Chong Tow Chong, Bo Hou and Wei-Ya Xi, "Simulation and analysis of FC network," 28th Annual IEEE International Conference on Local Computer Networks, 2003. LCN '03. Proceedings., Bonn/Konigswinter, Germany, 2003, pp. 285-288, doi: 10.1109/LCN.2003.1243142.

[5] D. T. Vojnak, B. S. Đorđević, V. V. Timčenko and S. M. Štrbac, "Performance Comparison of the type-2 hypervisor VirtualBox and VMWare Workstation," 2019 27th Telecommunications Forum (TELFOR), Belgrade, Serbia, 2019, pp. 1-4, doi: 10.1109/TELFOR48224.2019.8971213.

[6] Liang Chen, Longchuan Yan and Yifan Mao, "Research and application of virtualization technology power in data center," 2014 China International Conference on Electricity Distribution (CICED), Shenzhen, 2014, pp. 613-617, doi: 10.1109/CICED.2014.6991785.

[7] S. Bhaumik et al., "NetStor: Network and Storage Traffic Management for Ensuring Application QoS in a Hyperconverged Data-center," in IEEE Transactions on Cloud Computing, doi: 10.1109/TCC.2020.2969154.