POLITECNICO DI TORINO

Master's Degree in Communications and Computer Networks Engineering

Master's Degree Thesis

Hybrid Optical and Electrical Data Center

Supervisors

Prof. Paolo GIACCONE

Candidate

Razieh HEIDARIAN

09 2021

Summary

Nowadays, big companies like Google, Yahoo, and Amazon are constructed by mega-data, which contains hundreds and thousands of servers. Therefore, the standard data center with leaf and spine topology within electrical switches is limited to support this massive amount of data within a complex architecture, limited bandwidth, complicated cable organization, energy, and high power consumption.

Introducing optical switches provide this opportunity to resolve these issues by flat architecture and less power consumption to leverage these switches in the core of the network of data centers.
However, current data center architecture is created by electrical and optical switches, which are called "Hybrid optical and electrical data center", which divided the bandwidth according to the factor of (α) to improve efficiency. A question regarding this matter may be brought forward to the reader as, why establishing the electrical switches is still commonly used. In contrast, we can use this opportunity to leverage the full optical switches. In the conventional servers, the link bandwidth and traffic are divided between CPU-to-memory and CPU-to-storage to enhance performance. Furthermore, the peak of data rate between CPU-to-memory is much higher than CPU-to-storage, so optical switches are the best choice to have the highest possible speed with unlimited bandwidth to use them as a connection between CPU-to-memory and electrical switches link between CPU-to-storage. Moreover, I would like to clarify and indicate the significant reasons why electrical data centers are still prevalent in this day and time. First, a lack of buffer in optical switches is an essential element to use electrical switches in ToRs and servers to avoid losing the packets, and the second reason is the high cost of optical switches does not allow us to use full optical switches in the current data centers. The main focus of this thesis shall strongly be based on how these two data centers operate; I will describe how these two data centers work.

A brief description of each chapter:

- In the first chapter, I will introduce how electrical and optical switches work in data centers.
- The second chapter shall relate to each device of the data center and its core function.
- The third chapter shall indicate the relative implementations in regards to the vast data center by taking into consideration two phenomena; electrical switches and hybrid switches.
- The fourth chapter shall shed light on a hypothesis of software created by myself and how it stimulates the leaf and spine topology.
- The last chapter shall be dedicated to outlining such software by portraying a graphical view of topology.

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Professor Paolo Giaccone, whose expertise was invaluable in formulating the research and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I want to thank my dear parents, who were always by my side and did not leave me alone to guide me to be a better child. You are always there for me. Finally, I could not have completed this thesis without the support of my husband, Lucio, who provided inspiring discussions and happy distractions to rest my mind outside of my research.

Table of Contents

Chapter 1

Introduction

Data center networks have been rapidly evolving in recent years. The traditional enterprise client-server workloads and also the modern data center's workloads are dominated by server traffic [1]. Servers are the smallest physical unit of the data center that can hold CPU, memory, and storage. The full data center build by interconnecting hundreds of thousands of servers and storages system with complex topology inside the rack, which works with high speed. It means that amount of computing and memory resources depends on the proportionality of these resources in the individual servers [2]. in the data center a massive amount of data travel from east-west is more than north-south. This large communicating data is between servers and storages in the data center rather than inbound and outbound traffic. To increase the number of hosts, we need to increase the number of switching stages, as well.

Data center power consumption is a problem of significant importance. It produces more than 100 MW challenges for data center operators. The operators need are responsive to the climate change and environmental problems and the latest environmental footprint of data center activities. With this matter in mind, several big businesses have made great efforts to decrease energy consumption. Expected data center interconnects (DCIs) will therefore be expected to provide more data while utilizing less the energy consumed while carrying a unique bit over a link will have to be decreased to 1 PJ from many tens of PJ today. These demands provide reliability and availability of communication bandwidth inside of the data center network [3].

The development of this extensive range of data centers has increased essential engineering requirements such as the requirements for holding the servers up and working with minimum human interference, checking the data losses and sufficient consuming the heat produced by hundreds of thousands of servers. These

repository of data centers need simple high-bandwidth DCIs that can guarantee the servers' connectivity and allow enhanced source utilization. At these ranges, even small improvements in performance or utilize can significantly impact the overall network [3].

1.1 Standard data center

From where the spine and leaf design comes? It has based on the origins of a Clos network. Clos systems are called later Bell Labs researcher Charles Clos introduced the design in 1952 to succeed in the review and cost-related requests of Electro Mechanical switches used in telephone networks. Clos applied a mathematical method to demonstrate that obtaining non-blocking performance in a switching design (now known as a fabric) was possible if the switches were designed in a hierarchy [4].

Clos network is a multistage circuit switching network which illustrates a technical idealization of practical multistage switching systems. Clos topology is applied to build a leaf and spine architecture of interconnecting leaf switches (data center access switches or ToR switches) together in spine switches.

Leaf and spine topology, as presented in figure 1.1, is including a folded clos full mesh topology in which each leaf switch is joined to each spine switch. The leaf switches are the switches that are immediately attached to servers. In this position, the request ought to pass through only one switch if the source and destination servers are joined to the same leaf switch. Thus, this topology is additionally a specific case of clos topology in which both exit and entry are the leaf switches, and spine switches act as the central stage.

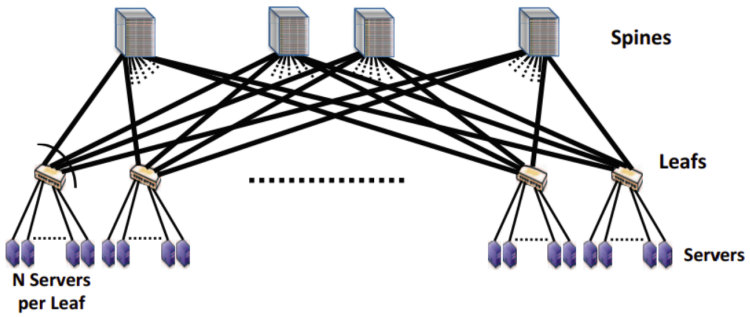


Figure 1.1: The leaf and spine topology reproduced by [1]

1.2 Modern data center

Given that by increasing the traffic, optical technology maintains large bandwidth that is much larger than global IP traffic. Optical switches play a key crucial role in the data centers, meaning that they are transparent to the bit rate of optical switches, which is one big difference between the hybrid data center and the electrical data center. Besides, optical switches for the hyper-data centers are expected to break the power and bandwidth barriers raised by electrical technologies in the future [5]. Another significant point of optical networks is their ability to dynamically reconfigure optical routes between electrical switches joined using an optical switch. This ability can be used to determine one major challenge in data centers VM (virtual machine) employment problem. As optical tracks between edge-switches (top-of-rack switches, to which server machines are connected) can be generated on-demand, there is more adaptability in putting VMs of a request than in an electrical data center network.

The key attributes of such optical circuit switches are the available port count; if we can utilize large-port count optical switches, flat and single-stage optical switching will become possible, eliminating complicated traffic congestion control needed in the multi-stage networks. Indeed, present mega data centers often have latency levels of dozens of microseconds [6].

On the other hand, creating an all optical data center that produces synchronous connectivity between every two edge-switches is costly and impossible for large data centers hosting tens of thousands of servers; therefore, electrical switches are better adapted in short and bursty traffic. Moreover, a hybrid optical and electrical network architecture the best option for prospective data centers. A hybrid data center provides adaptability in joining edge-switches with great connection demands dynamically utilizing the optical network while managing contacts between edge-switches with burst traffic using the electrical network.

1.3 Issues and challenges

Identifying the best architecture seems to be too early as it depends on scale, switching rate and desired bandwidth which will be influenced by the applications. Some general estimates of the introduced designs are given here.

Scalable control plane

One of the leading design variations among electrical packet switches and optical switches is that they are typically bufferless. While the absence of buffers avoids

each queuing delay also decreases peer-to-peer latency, it remarkably intricates the design as flows need to be precisely arranged to check any collision, which would occur in signal degradation and packet loss. Regularly, this is obtained employing a centralized scheduler that, given a request traffic matrix, estimates the flow responsibility and the optical switch configuration [7].

Cost

One of the significant advantages of electrical switches, which we can benefit from economies of scale. Optical switches are expensive because the manufacturing process is less mature and requires expensive packaging and testing procedures. Further, the single switch device's cost and power must consider all additional costs derived from implementing an operational optical network. Finally, depending on the technology used, network architects may need to over-provision the number of optical switches or transceivers to compensate for the loss of throughput due to switching time or the inter-packet gap. To produce an optical data center with cost-effectiveness against its electrical equivalent, all these contributing factors should be appropriately accounted for during the design and minimized cost [7].

Reliability

Although price is a crucial metric, cloud providers' top priority is to secure high availability and continuous service. Despite minor outages negatively impact first and third-party companies and eventually result in sharp money loss and decreased market share. According to the tighter coupling exhibited by optical-switched networks, they are intrinsically more lying to failures, and special attention obligation is used to protect upon such situations. For instance, centralized schedulers or a control-plane system with no (or limited) repetition represent individual failure points and should be evaded. Furthermore, the time synchronization protocol needs are robust against single node crashes or network distributions [7].

Despite all the challenges we discussed earlier, we conclude that optical switching is growing developed. It can transform the cloud foundation by producing anticipated and uniform high achievement (bandwidth and latency) beyond the whole data center. Therefore developing today's silos (e.g., a single server or a rack) with essential advantages in fault tolerance, source administration, and application performance. Moreover, any of the technology evolved to maintain optical switching could further build new possibilities to rethink other parts of the stack. For example, a tightly programmed network would decrease the reliance on distributed congestion control protocols like TCP, making it more comfortable

to perform complicated QoS policies and implement supported performance to applications and services operating in the cloud [7].

Chapter 2

Data Center Networks

2.1 Electrical switches

In this section, we want to talk about the rule of each device in the data centers. Figure 2.1 represents multi-stage electrical data center topology. As you see in the figure, the boxes at the bottom of the picture are racks whose number depends on the requirement bisection bandwidth, and they are connected to the one Ethernet switch called ToR or EoR.

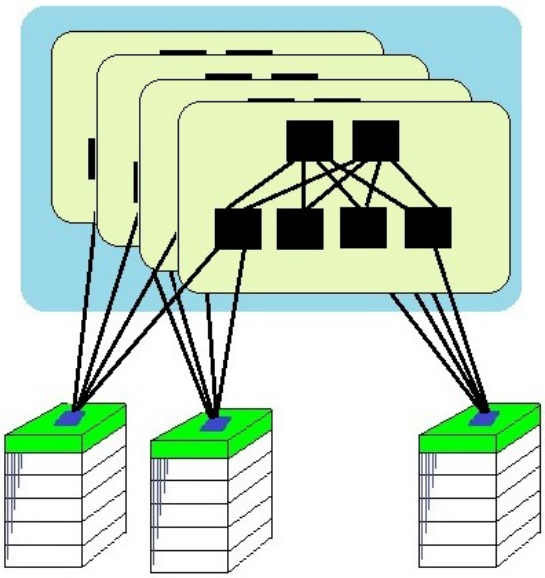


Figure 2.1: Multistage electrical data center network

2.1.1 ToR or EoR

There are two main deployment plans, either from the top level or end level in rack designs, one or two Ethernet switches are installed inside each rack to provide a local server connection. While the name top of rack means to put the switch on top of the physical rack. Also it can be at the bottom or middle (usually above the rack easier access point and cable management for substitute). The most important advantage of positioning your switches inside the racks is connecting the servers to the switch. This implementation style eliminates the requirement for cabling panels, which need additional shelves and a large amount of cabling between the shelves [4].

In the top of rack architecture, servers attach to one or two Ethernet switches installed inside the rack. The phrase "top of rack" has been invented for this design. Nevertheless, the switch location does not certainly require to be at the top of the rack. Other switch positions could be the bottom of the rack or middle of the rack; though, the top of the rack is most popular according to more comfortable and cleaner cable administration. This scheme may also sometimes be introduced as "In-Rack". The Ethernet top of the rack switch is typically determined configuration. The top of rack design's essential characteristic and request are that all copper cabling for servers stays within the rack as relatively short patch cables from the server to the rack switch. The Ethernet switch connects the rack to the data center network, directly connects to the higher layer in the data center [8].

Each rack can be used and manipulated like an individual and modular unit within the data center. It is simple to change out or updates the server access technology rack-by-rack. Any network updates or problems with the rack switches will only affect the servers within that rack, not an entire server row. Given that the server joins with short copper cables inside the rack, there is more elasticity and choices concerning what that cable is and how fast of a connection it can hold. Fiber to every rack produces greater flexibility and investment security than copper because of fiber's unique capability to carry higher bandwidth signals at longer distances [8].

The term end of rack was invented in which an Ethernet switch located at the end of the rack, and it defined a rack or cabinet situated at either end of the server row to produce network connectivity to the servers inside that row. There may be a few network racks located in a small row of their own, collectively giving end of rack copper connection more than one row of servers [8].

The end of rack switch gives a connection to the hundreds of servers within that row. Therefore, unlike top of rack, where any rack is its controlled system, with end of rack, the whole row of servers is arranged like one holistic piece or "Pod"

within the data center. Network updates or issues at the end of rack switch can be service-impacting to the entire row of servers. This scheme's data center network is managed per row rather than per rack [8].

2.1.2 POD structure

Point of Delivery(POD) a module or group of the network, compute, storage, and application components that work together to deliver a network service. The pod is a repeatable pattern. Its components increase the modularity, scalability, and manageability of data centers. By pod layers, we can increase the number of servers that each data center required. We can increase the data center's layer and the pod layers according to the number of servers [9].

2.2 Optical switches

Modern data center networks rely on fiber-optic connections to satisfy bandwidth requirements also to neglect the expensive optical-electrical-optical (OEO) transformation required to join such connections with power-hungry electrical packet switches, researchers have suggested network designs that transfer enough of a data center's traffic quietly using optical circuit switches(OCSes). OCS can hold too high link bandwidths at low per-bit cost and low power consumption because they redirect light from one port to another port, autonomous of data rate. Optical circuit switching confronts two significant obstacles to wide-scale selection in the data center situation. The first obstacle to deployment is the dependent control plane. Current designs to use OCSes in the data center reconfigure optical circuits in reply to traffic demands. Reconfiguration requires managing a network-wide demand data center to calculate a plan of switch configurations, rate-limiting packet communications, and synchronizing the OCSes with each other, the plan, and the end points. This strong coupling among the different network elements offers a significant challenge at scale-up. The second issue with employing commercial OCS devices in data centers is their inadequate scalability, especially their level port number and slow configuration rate [10].

By introducing optical switches, difficulties presented by the typical architecture must be defeated. The lack of a commercially viable clarification for optical random access memory and optical buffering performs it unlikely that true optical packet switching (that is, optical switches which can produce on a packet-by-packet base) will develop in the near term. Optical switches thus cannot be counted as a one-to-one replacement for electronic packet switches. The network design will likely apply optical switches in some organizations with standard electronic switches for enhanced execution. In all these instances, optical switches are utilized to adjust

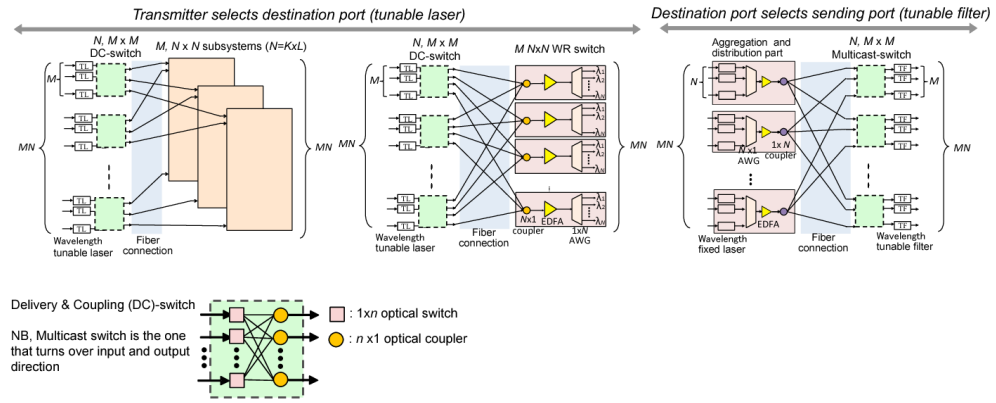
the network to precise traffic patterns. In other words, pairs of racks transferring more important traffic levels can be granted higher bandwidth using the optical network. Reconfiguration of an optical switch breaks previous connections and needs phase locking, handshaking, and modifying the new links' routing tables [3].

2.2.1 Proposed large port count optical switches

Picture 2.2 represents the architectures of our lately introduced optical switches. Switching with wavelength routing can be performed by tuning the optical source wavelengths of the tunable filters at optical receivers. Figure 2.2 (a) and (b) determine the method of tunable lasers that are performed in ToR transceivers, while (c) displays working tunable filters at the receiver side of the function is executed with coherent detection. The architectures in Figure 2.2(a) and (b) differ in wavelength routing switch organization. (a) uses wavelength routing subsystems that consist of cyclic AWGs (Arrayed Waveguide Gratings), while (b) uses aggregates of an optical coupler and a conventional non-cyclic one x N AWG. Architectures (b) and (c) is almost proportional; the input and output directions are reversed, and tunability at the input side is performed in (b) by using wavelength-tunable lasers, while that at the output side is done by using tunable filters in (c) [5].

2.2.2 Combination of Tunable Lasers and Cyclic AWGs

The architecture in Figure 2.2 (b) applies optical couplers and AWGs in the wavelength routing element rather than cyclic AWGs as in Figure 2.2 (a), where EDFAs can compensate the coupler loss. The mixture of an interleaver and AWGs offers the best method to use their advantage of characteristics: the interleaver has some ports, however, a steep filter shape. In contrast, the AWG has a regular filter shape, though its port number can be huge. Thanks to the critical collection of wavelength signals and fine granular wavelength routing, we can reach a massive scale optical switch cost-effectively. The almost expensive EDFA is given by many wavelengths, which efficiency small per-port cost [5].



(a) AWG based wavelength routing switch [33] (b) Tunable laser based wavelength routing switch [35] (c) Tunable filter based wavelength routing switch [36

Figure 2.2: Scheduled MNxMN optical switch designs (a) AWG based wavelength routing switch(b) Tunable laser based wavelength routing switch(c) Tunable filter based wavelength routing switch reproduces by [5]

2.3 Hybrid optical and electrical data center

Optical technologies appear promising for switching large bandwidth traffic. Intra data center traffic is much larger than global IP traffic. Optical switches will play a critical role in intra data center networks in the future. Note that most optical switching schemes are transparent to the bit rates of the optical signals, completely different from electrical switching systems [5]. Besides, the power consumption of optical switching is much smaller than electrical systems. Hence, large bandwidth and low power consumption switch systems are possible, which is in contrast to the Silicon switch chip. The optical switch offers large bandwidth switching capability and eliminates multi-stage switch network architecture needed with electrical switching. Furthermore, the optical switch's single-stage architecture greatly simplifies operating costs, including cabling complexity. It substantially reduces the number of transponders needed. One of the most powerful data center requirements is cost-effectiveness and scalability [6].

In data centers, traffic can be categorized according to flow size: mice flow and elephant flows the difference in size and latency requirements. Elephant flows are associated with virtual machine migration, data backup, large file transfer, high-quality videos, etc. Most are not latency sensitive, while mice flows are very sensitive. Mice flows are dominant in numbers, but elephant flows determine the total bandwidth. Offloading elephant flows from electrical switches to optical switches (electrical and hybrid switch approach) can dramatically reduce the electrical switching bandwidth needed [11]. In optical switch networks, large port

count optical switches are useful since employing small switches demands multi-stage configurations or the traversal of multiple switches between ToR switches.

Reducing switching stages by using large port count optical switches can reduce the number of transceivers and interconnection fibers. Electrical ToR switches are connected through single-stage optical switches or multi-stage electrical switches. For large data centers, the number of switching stages is more than 2 (3 stages, including ToR). Hence, optical switches reduce the number of transceivers, which substantially simplifies network configuration. shows the necessary number of optical switches and bisection bandwidth to interconnect ToR switches.

Please note that this parallel use of optical switches significantly reduces traffic collision at the destinations and sources or the resultant sending delay (ToR switch is electrical and employs buffers). The performance (delay and possible data loss) can be analyzed using the multiple server model in queuing theory, where such parallelism significantly reduces information sending delay. The link speed and the necessary parallelism can be determined by considering the data center's control policy accommodating different applications and link costs. As discussed above, the introduction of optical switches substantially reduces the number of power-hungry electrical switches. Hence, the electricity cost of intra data center networking can be significantly reduced.

Optical interface costs (including transceivers) can be a substantial part of network costs for present data center networks. This indicates the potential network cost reduction possible with the introduction of optical switches. Optical switches substantially simplify the overall switching system. Please note that, with optical circuit switching, only a limited number of electrical switches need to be used to create a hybrid switching system [5].

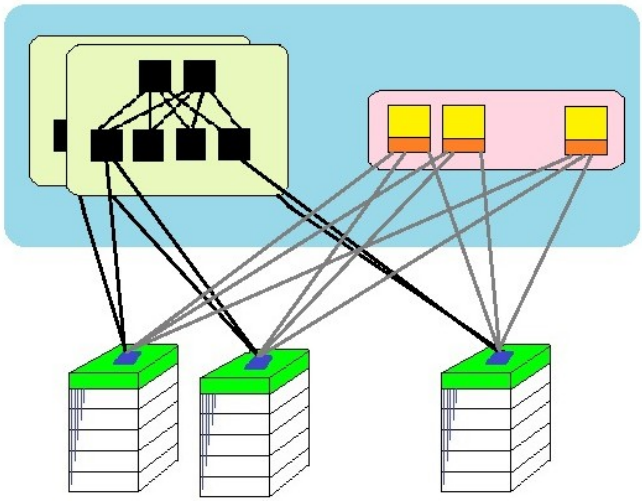


Figure 2.3: Hybrid optical and electrical data center

2.3.1 Disaggregated rack

Disaggregation is a concept in which related sources are merged, and the different sources acting separately updated and the network adaptively configured for optimized execution. The system can be disaggregated at various levels, for instance, at the rack or increasing the servers. The disaggregated data center needs an interconnection fabric that necessity supports the additional traffic produced by the disaggregation and be tremendous bandwidth and low latency to support and enhance execution. The network needs a switching fabric to provide the computing devices. Although packet-switched channels maintain electrical and optical circuit switches are the best applicants for re-configuring sources in the disaggregated system. Notice to any added latency in the interconnect that force lead to performance degradation. Average latency to memory, in standard servers in which the memory is near the CPU, takes tens of nanoseconds [3].

Disaggregated data center explains three principal kinds of blocks, i.e., compute, memory, and accelerator chunks. Figure 2.4 presents a schematic representation of the method design. Bricks are plugged in standard rack-mountable trays; a tray container hosts random mixes of each sign's blocks. Recursively, racks are made out of trays and so the entire data center. Blocks in the equivalent tray are interconnected by a crossbar electrical switch, while a circuit-switched optical network attaches bricks crossed several trays and racks. A software control plane manages the system administration; if a new workload report, the control plane configures the system software also hardware on the bricks and the network switches to build circuits attaching one compute brick through one or more memory bricks [2].

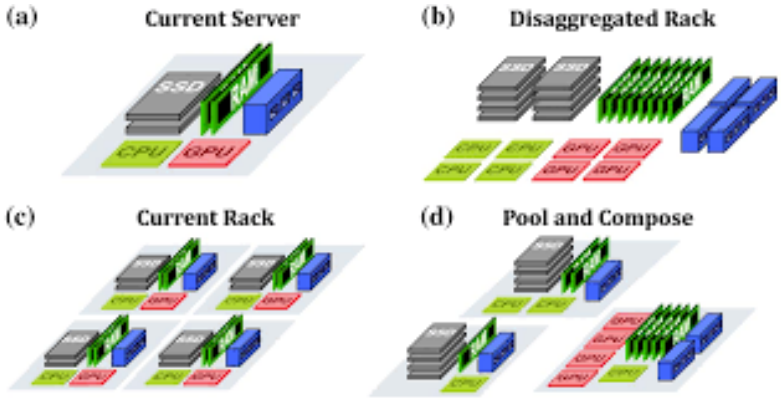


Figure 2.4: Disaggregated rack reproduced by [3]

2.3.2 Bisection bandwidth

In computer networking, the bisection bandwidth of a network topology is the bandwidth available between the two partitions. The bisection bandwidth is the bandwidth that goes from servers to the higher layer in the data center. Equation (2.1) is to compute the bisection bandwidth without over-subscription, in which (N_s) is the total number of servers which the data center higher layer can supports and (S_R) is the server link rate.

$$B_\mathrm{B} = N_\mathrm{s} \times S_\mathrm{R} \tag{2.1}$$

While when we have a value of over-subscription, we have to divide the whole bisection bandwidth to the value ($alpha$), which is:

$$B_\mathrm{B} = \frac{(N_\mathrm{s} \times S_\mathrm{R})}{\alpha} \tag{2.2}$$

where:
B_B = total amount of bisection bandwidth
N_s = total number of servers
S_R = server link rate
α = the over-subscription ratio

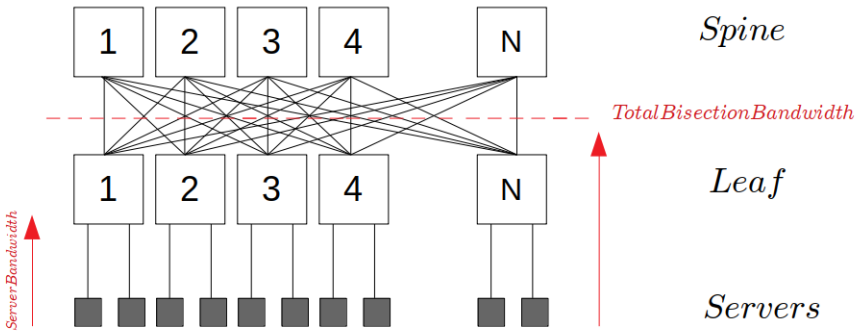


Figure 2.5: Total bisection bandwidth

Chapter 3

Design Methodology

3.1 Electrical data center without over-subscription

Recent data center designs regularly utilize a two-tier LAN topology design, where the lower layer is called a leaf, and the upper layer is called the spine. Moreover, leaf switches which are the lower layer switches attach to the racks. Each rack contains servers interconnected to the electrical ToR switches. The traffic from leaf switches goes through spine switches in the upper layer. Besides, in standard designs, most of the traffics transferred between North to South, from servers to the Northbound, where the growing requirements of applications to scaling up networks needed considerable resources to increase network performance. Nevertheless, current designs are more focused on East to West traffic flows between VM in the racks and exchanging the data to increase the network performance [12].

In the figure, 3.1 shows the architecture leaf and spine topology constructed by the principal characteristic of which leaf switches attach to the spine switches. L_{s} is the number of switches in the leaf layer which amount of these switches depends on the number of servers to support. Moreover, it is essential to have the same quantity of spine switches as the number of up-links ports of each leaf switch, where the relevant number will be named S_{s}. The deeper layer that hosts contained is called P_{s}, its number is not fixed, and it would vary according to the quantity of bandwidth and other over-subscription rates considered. Moreover, the whole number of hosts available in the topology will be achieved by multiplying L_{s} by P_{s} in data center layers.

The ideal situation is an over-subscription ratio of 1:1, signifying that the total bandwidth available in the down-link ports is the same as in the up-link ports [12]. However, this condition means having greater power consumption and more expensive costs and a loss of bandwidth if not every host is working at the

full rate. Therefore, the advantage of a leaf and spine structure is that according to massive transmitting data is between east to west, hosts can be communicating, and there are two hubs in between, which are two ToR inside the racks. This is a great advantage compared to similar topologies such as Fat Tree. Unfortunately, that very same benefit carries an inherent drawback, as there may appear scalability issues when the number of hosts is high; it needs a considerable number of cables and, as a consequence, high power consumption due to the high quantity of switches.

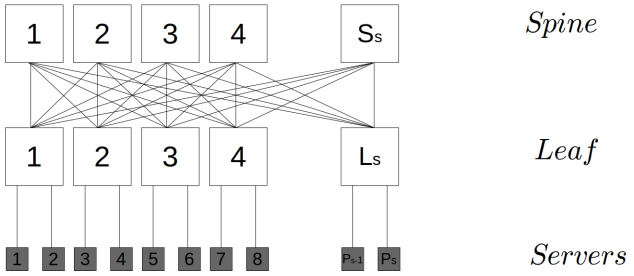


Figure 3.1: Block diagram of the leaf and spine data center

3.1.1 Two layer data center

The two main features of a data center are switching time and performance. This section describes a new method to satisfy this two keys and the exact required bandwidth. In figure 3.2, you can see the leaf-spine architecture entire structure with all the notations that you can easily remember.

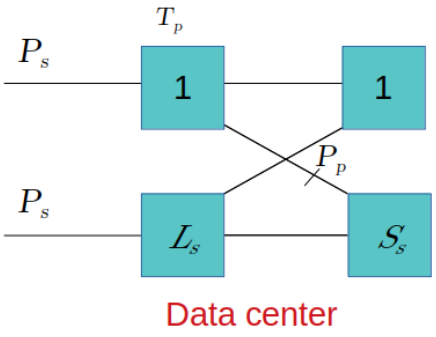


Figure 3.2: Block diagram of two layer of data center

I have estimated all the following equations to have a controllable data center and avoid wasting switches and cables. As you can see in the formulas, we need to estimate the number of leaf and spine switches, total server ports, and parallel cables. For all equations, I applied ceiling functions to make sure all servers can be supported.

All these equations work for small and massive data centers. Equation (3.1) estimates the number of switches in the leaf layer. We need to divide the total number of servers by the half of the total ports per each switch, and we know how many leaf switches are needed to connect to the servers. Equation (3.2), as before, by dividing the number of servers by the total number of ports for each switch determine the spine switches. Equation (3.3) measures ports for servers; we need to divide the number of servers by the number of switches in the leaf layer to avoid using the switches with a large number of ports. Moreover, equation (3.4) is the number of parallelism levels; to avoid wasting switches and cables, we have to calculate parallelism level. All the equations as follows:

$$L_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{H_{\mathrm{p}}} \right\rceil \tag{3.1}$$

$$S_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{T_{\mathrm{p}}} \right\rceil \tag{3.2}$$

$$P_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{L_{\mathrm{s}}} \right\rceil \tag{3.3}$$

$$P_{\mathrm{p}} = \left\lceil \frac{P_{\mathrm{s}}}{S_{\mathrm{s}}} \right\rceil \tag{3.4}$$

Where:
L_{s} = Total number of leaf switches
S_{s} = Total number of spine switches
N_{s} = Total number of servers
P_{s} = Total number of ports for servers
P_{p} = Total number of parallelism livel
T_{p} = Total number of ports per each switch
H_{p} = Half of number of ports per each switch

3.1.2 Two layer POD

As I mentioned above, when our data center is enormous, we have to design a two-layer pod that is the point of delivery, which is a module or group of the network, compute, storage, and application components together to deliver a network service. The pod is a repeatable pattern its components increase the modularity, scalability, and manageability of data centers. By pod layers, we can increase the number of servers that each data center required. We can increase the data center's layer and the pod layers according to the number of servers. Figure 3.3 could see all the notations with the block diagram of the pod's two-tier.

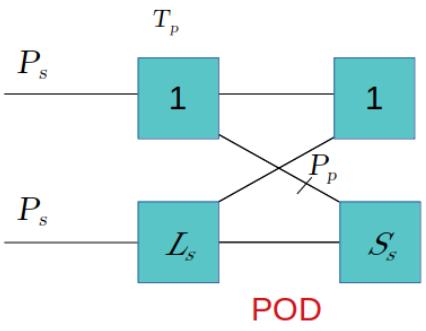


Figure 3.3: Block diagram of two layer of pod

The design of this layer gives us the ability to support many servers in higher layers. In this part, we have to calculate the number of servers and the number of total spine switches for the next layer of the data center. Therefore, I had to use the last part's notations in this section, but with a little modification in the total number of leaf switches, equation (3.5) shows the modified formula. We keep all formulas as the same as before with little modification in the number of leaf switches. In this layer, we can not use the same number of leaf switches for two-layer of the pod as before; thus, we use half of them to construct the pod's two-layer.

To design a three-layer data center, we had to utilize the calculations of this section. Therefore, to know how many servers can be supported in the next layer, we have to multiply the P_s by the number of L_s switches, and also to compute the number of spine switches for the next layer, we have to sum the number of L_s and S_s.

$$L_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{T_{\mathrm{p}}} \right\rceil \tag{3.5}$$

Where:
L_{s} = Total number of leaf switches

3.1.3 Three layer data center

To design a large-scale data center, we oblige to create the third layer. Thus, to produce the third layer for the electrical data center, we need to use all equations to build the layer two data center and the pod to support the required bandwidth to develop an enormous electrical data center.

We keep all the equations as the same as before with a little difference in the number of ports for servers and the number of spines switches which I discussed how we could calculate in the last part. I mention again to calculate the number of ports for servers in layer three of the data center; we multiply the number of leaf switches with the total number of ports for servers in the pod layer. Therefore, equation (3.6) the number of P_{s} is obtained by layer two of the pod in the last section and so to calculate the leaf switches we have to use half of the amount of these ports, and equation (3.7) shows the calculation of spine switches which obtained by dividing the number of servers by total ports for each switch. Equation (3.8) is divided by the total number of servers by leaf switches calculate the number of total ports per each switch in leaf layer. At the end equation, (3.9) is the parallelism level estimated by dividing the total number of ports for servers by the total number of spine switches. All modified formulas are as follows:

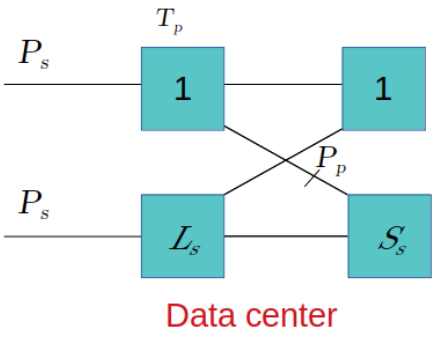


Figure 3.4: Block diagram of three layer data center

$$L_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{H_{\mathrm{p}}} \right\rceil \tag{3.6}$$

$$S_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{T_{\mathrm{p}}} \right\rceil \tag{3.7}$$

$$P_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{L_{\mathrm{s}}} \right\rceil \tag{3.8}$$

$$P_{\mathrm{p}} = \left\lceil \frac{P_{\mathrm{s}}}{S_{\mathrm{s}}} \right\rceil \tag{3.9}$$

Where:
L_{s} = Total number of leaf switches
S_{s} = Total number of spine switches
N_{s} = Total number of servers
P_{s} = Total number of ports for servers
P_{p} = Total number of parallelism level
T_{p} = Total number of ports per each switch
H_{p} = Half of number of ports per each switch

3.2 Electrical data center with over-subscription

A lot of recent researches have concentrated on defining and analyzing new and promising architectures for system level optical interconnects in data centers. Most of the proposed technologies and architectures have been initially developed for the application in access and core networks and slightly adapted to match data centers'

requirements. Hybrid architectures usually rely on a compound of commercial electrical switches for dynamic packet switching and simple yet energy-efficient optical switches providing circuit switching capabilities. While hybrid architecture is somewhat flexible and able to adapt to varying traffic situations, and cost-efficient because they use state-of-the-art commercial technology, the need for electrical commodity switches makes them the less viable long-term solution future data center networks [13].

Optically switched interconnects can be seen as a promising candidate for future data centers because they offer the highest capacity and bandwidth density and the potential for the lowest latency among all interconnection options. When implemented in a pure circuit switched manner by using large optical switches such as, e.g., optical MEMS switches, the system can be built to provide high scalability, low energy consumption, and a relatively low cost [6].

However, the applications requiring dynamic switching cannot be optimally supported because of the large reconfiguration overhead of circuit switching, which leads to a low transmission efficiency. On the other hand, architectures providing fast all-optical packet switching are usually more complicated and expensive and typically less scalable. Additionally, the lack of practical optical buffering technologies limits the achievable performance of large all-optical packet-switched networks. Thus, the architecture of choice needs to provide excellent scalability as well as high efficiency and reliability. The term efficiency is to be broadly construed and includes transmission, energy, and cost-efficiency [14].

Over-subscription is commonly used to take advantage of network traffic patterns. In this part, by over-subscription ratio, we can divide the bandwidth between electrical and optical switches. We know that optical switches provide a high amount of bandwidth and low power consumption, thus the vast amount of bandwidth dedicated to the optical switches rather than electrical switches.

3.2.1 Two layer data center

In this section we want to design a data center with an over-subscription ratio by an (α) factor. It is an over-subscription ratio to divide the bandwidth between electrical and optical switches. In this part, I will explain why we should use over-subscription to design the data center. As mentioned above, we divided the bandwidth fairly between the input and output ports for the electrical data center. However, in this section, we have to dedicate the bandwidth to the electrical switches and the optical switches, with over-subscription 10:1, which is the amount of bandwidth that goes to the optical data center is ten times bigger than the amount of the bandwidth which goes to the electrical data center. Thus, I have to redesign the full electrical data center and rewrite all the equations with over-subscription ratio. Figure 3.5 shows the topology of the hybrid optical and electrical

data center which the gray one is the optical switches and green is the electrical switches.

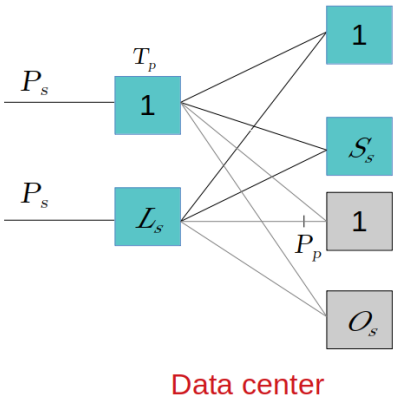


Figure 3.5: Block diagram of two layer of data center

Electrical switches designed with multi-stages leaf and spine topology and optical switches are flat single stage with large port counts. In these formulas, our prototype information is the total number of servers (N_s), the total number of ports for each switch (T_p), the size of the optical switch (O_p), and the over-subscription ratio (α). As you see in equation (3.10), we can calculate the total number of ports for servers by dividing the total number of servers by total ports, equation (3.11) shows the calculation of the number of leaf switches obtained by the total number of servers divided by the total number ports for servers (P_s) that we calculate it in the last equation. The equation (3.12) represents the number of electrical switches in the spine layer, which can be achieved by the total number of ports for servers (P_s) multiply by ($1/\alpha$). The equation (3.13) is the whole number of optical switches connecting to the leaf layer, obtained by multiplying the number of ports for the server by ($1 - 1/\alpha$). The last equation (3.14), which is parallelism level, can be obtained by the size of optical switches divided by the number of leaf switches. All equations are as follows:

$$P_s = \left\lceil \frac{N_s}{T_p} \right\rceil \tag{3.10}$$

$$L_s = \left\lceil \frac{N_s}{P_s} \right\rceil \tag{3.11}$$

$$S_s = \left\lceil P_s \times (\frac{1}{\alpha}) \right\rceil \tag{3.12}$$

$$O_{\mathrm{s}} = \left\lceil P_{\mathrm{s}} \times (1 - \frac{1}{\alpha}) \right\rceil \tag{3.13}$$

$$P_{\mathrm{p}} = \left\lceil \frac{O_{\mathrm{p}}}{L_{\mathrm{s}}} \right\rceil \tag{3.14}$$

Where:
L_{s} = Total number of leaf switches
S_{s} = Total number of spine switches
N_{s} = Total number of servers
P_{s} = Total number of ports for servers
P_{p} = Total number of parallelism level
T_{p} = Total number of ports
O_{s} = Total number of optical switches
O_{p} = Total number of optical ports
α = over-subscription ratio

3.2.2 Two layer POD

In this part, to expand the data center, we have to design the pod layer. Therefore, I had to use the last part's notations in this section, with a little modification in the total number of leaf switches. We can not use the same amount of leaf switches that we used in layer two of the data center for the pod layer; for this reason, we have to use half of the leaf switches used in the second layer of the data center for the pod layer. Equation (3.15) shows the calculation of leaf switches.

$$L_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{T_{\mathrm{s}}} \right\rceil \tag{3.15}$$

Where:
L_{s} = Total number of leaf switches

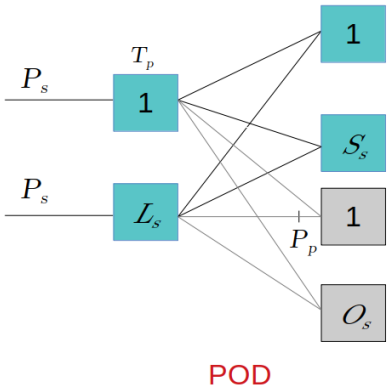


Figure 3.6: Block diagram of two layer of pod

3.2.3 Three layer data center

To design a large data centers, we must to build the third layer data center. To provide the third layer for the electrical data center, we need to utilize all equalization to make the data center layer two, to support the required bandwidth to develop an enormous electrical data center. Thus, we have to use the equations that we used for layer two of the data center with a small variation in the number of ports for servers and the number of spine switches we obtained in layer two of the pod. We apply the exact amount of leaf switches to apply for two layer to design this hyper-scale data center. To determine the number of spine switches, we multiply the leaf switches with spine switches in the previous pod layer. Therefore,

as I explained in the previous section, to know the number of ports for servers in the third layer of the data center, we need to identify the number of servers that we got in layer two of the pod. Calculating the number of full switches requires multiplying the number of leaf switches with the entire amount of switches that we got in the pod layer and add the number of spine switches in the third layer. Figure 3.7 is the three layer data center topology.

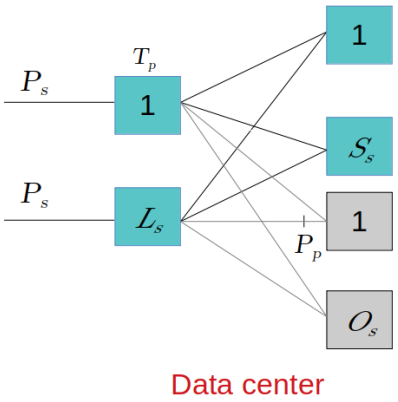


Figure 3.7: Block diagram of three layer of data center

$$P_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{T_{\mathrm{p}}} \right\rceil \tag{3.16}$$

$$L_{\mathrm{s}} = \left\lceil \frac{N_{\mathrm{s}}}{P_{\mathrm{s}}} \right\rceil \tag{3.17}$$

$$P_{\mathrm{p}} = \left\lceil \frac{O_{\mathrm{p}}}{L_{\mathrm{s}}} \right\rceil \tag{3.18}$$

$$O_{\mathrm{s}} = \left\lceil P_{\mathrm{s}} \times (1 - \frac{1}{\alpha}) \right\rceil \tag{3.19}$$

$$S_{\mathrm{s}} = \left\lceil P_{\mathrm{s}} \times (\frac{1}{\alpha}) \right\rceil \tag{3.20}$$

where :
L_{s} = total number of leaf switches
S_{s} = total number of spine switches
N_{s} = total number of servers

P_{s} = total number of ports for servers
P_{p} = total number of parallelism level
T_{p} = total number of ports
O_{s} = total number of optical switches
O_{p} = total number of optical ports
α = over-subscription ratio

Chapter 4

Numerical Results

4.1 Implementation design

4.1.1 Implementing electrical data center

In this section, you will see how we designed a large data center. As you see in figure 4.1 shows the whole electrical data center. It shows the leaf and spine architecture for the electrical data center. The purpose of this section is to achieve the 13.11Pb/s bisection bandwidth, which is almost massive bandwidth, the first step is to know how many servers we need to calculate by equation (2.1), so the whole number of servers is 131072. The second step is to choose the electrical switches with a large port count. The electrical switches with 256@100Gb/s ports are the best option to use to build a large data center and also can support a large number of servers. At this point, we can design our block diagram; as I discussed in the last section, the (L_s) is the number of switches in the leaf layer, which is 256 switches. To compute the spine switches (S_s), we have to divide total servers 32768 by the total number of ports I obtained 128 spine switches. Also, we have to calculate the number of ports for servers, which can calculate it by dividing the total number of servers by the total number of leaf switches which is 128@100Gb/s ports for servers. To compute the entire cables in this layer of the data center, we have to count all the wires, which are 65536 this number obtained by summation of total cables from leaf switches to ToRs and total cables connected from leaf switches with spine switches.

To build the pod two-layer, I used the same notations and calculations as the last part, but with one difference in leaf switches, we got 128 switches with 256@100Gb/s ports. In the two-layer of pod, the total amount of servers are 16384, which we can use for a total number of ports for servers in the next layer of the data center; as I mentioned in the last chapter, the total switches in this layer are

obtained by summation of leaf switches and spine switches which is 256 switches for this layer of the pod. Furthermore, we have to compute the total number of spine switches in this part which we got by multiplying the number of leaf switches by spine switches which are 16384 switches as well. The last computation is the total number of cables which get like as last part, which is the summation the cables go from leaf switches to ToRs and go to the spine switches which is 32768 cables.

To build the three-layer data center, we used the last computations from the previous layers. In this layer, the total servers are 131072, which can support 13.11Pb/s bisection bandwidth and whole switches compute by multiplying total leaf switches with the all switches which we obtained in the two-layer of the pod and sum it with the spine switches in the three-layer data center that is 2560 switches with 256@100Gb/s ports. The final computation is total cables which are 12582912 cables in the three-layer data center.

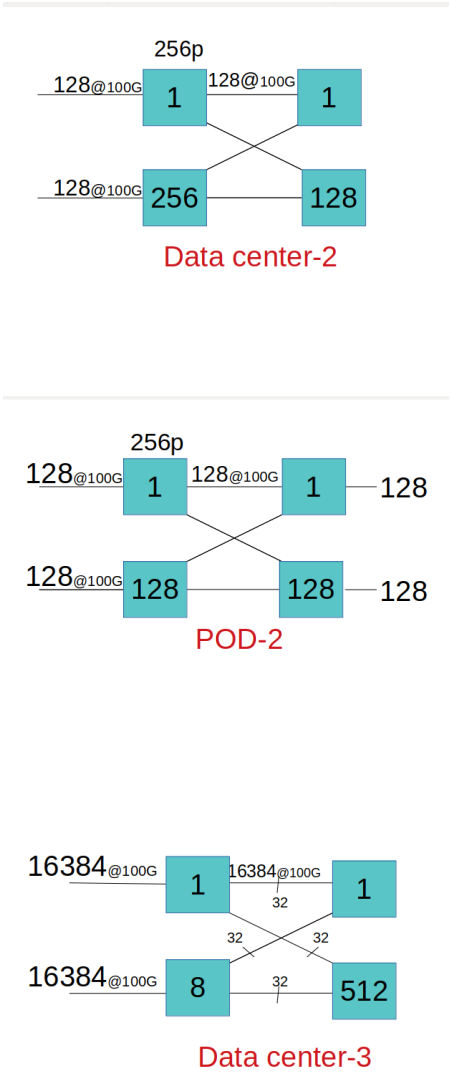


Figure 4.1: Block diagram of electrical switches without over-subscription(total number of ports per each electrical switch is 256@100Gb/s)

Table 4.2 has shown all the calculations of three layers. This table is the estimation of the large electrical data center. As you see in the table, the total switches are 2560, which is quite a significant amount of switches that consume a lot of power consumption. Moreover, the number of cables in this data center is too much according to the multi-stage architecture. This data center supports 13.11Pb/s bisection bandwidth, which supports 100% with electrical switches.

	servers	Spine switch	total switches	total number of cables	Total number of spine switches(without ToR)
Two layer data center	128 x 256 = 32768	-	128 + 256 = 384	(256 x 128)+ 32768 = 65536	-
Two layer POD	128 x 128 = 16384	128 x 128 = 16384	128 + 128 = 256	(128 x 128)+16384= 32768	-
Three layer data center	16384 x 8 = 131072	-	(256 x 8)+ 512 = 2560	(256 x 16384)+(256 x 32768) = 12582912	(8 x 128)+ 512 = 1536

Figure 4.2: Calculation of electrical data center without over-subscription

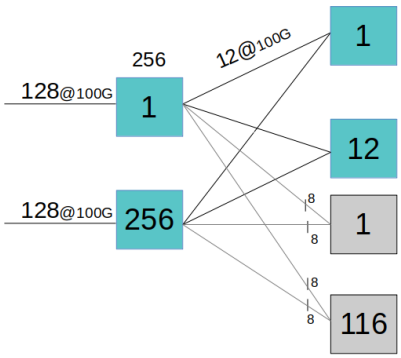
4.1.2 Implementing hybrid optical and electrical data center

Optical fiber can be used in data center networks to interconnect servers and switches to simplify cabling and avoid electromagnetic interference. The network engineers also study higher data rate and switching capacity (e.g., 40G, 100G) in the future data center design. Nevertheless, it is hard to produce a sizeable electric packet switch operating at high data rates due to the bottleneck of input and output bandwidth and the chip's power resources. As a result, many electronic switches need to be expanded to scale out the number of servers in the data center, which causes a severe scalability obstacle to the data center network in terms of cost and power consumption. Optical connections architectures that can produce ultra-high transmission speed and switching capacity in a price and energy-saving way are granted a hopeful solution to discuss the limitations of electronic packet switches (EPSs)data centers. By replacing EPSs with optical switches, the decreased power-demanding electrical to optical and optical to electrical (O/E) conversion are required to reduce data center networks' power consumption dramatically. Different optical interconnect architectures for data centers have been introduced in the research in recent years. However, these architectures operate all-optical switches based on different topologies and technologies at the core layer but rely on conventional. EPSs at ToR connects to the servers in the racks. Despite this, the EPSs at ToR are qualified for many overall data center traffic, and the EPSs at ToR add the most power consumption to the data center network. Therefore, energy-effective optical connection designs are expected for the way tier in the data centers [15].

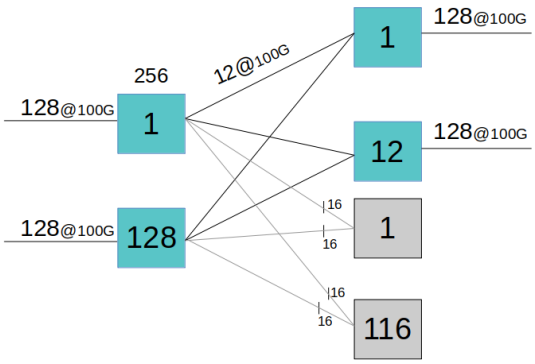
In this part, I will describe the hybrid optical and electrical data center; as I mentioned in the previous chapter, the lack of buffer in optical switches obliges us to use the electrical switches in ToR inside the racks. But we have to use optical switches for less power consumption and high-speed performance for the spine layer. In the two-layer data center, 256@100Gb/s electrical switches for the leaf layer, which 128@100Gb/s ports, go to the servers and the 128@100Gb/s ports dedicate to the electrical and optical switches which the bandwidth divide according to the (α) factor. Therefore, to calculate the spine layer, we follow the equations in the last chapter which the total number of optical switches is 116, and 12 electrical switches.

The two layers of pod follow the two-layer of the data center with one difference in the leaf switches because we want to create a pod layer, we have to use half of the number of switches that we used in the previous layer, which is 128 switches which they are connected to the higher layer of data centers.

To design the third layer of the data center, we have to use the calculation that we did in the last layer of the pod, the number of ports for servers which calculated by multiplying the number of leaf switches by the number of ports for servers in the previous layer, and we use it for layer three of the data center. The spine layer, which highest layer in the leaf and spine topology, is built by electrical switches and optical switches that the number of each data center depends on (α) value. The total number of spine switches is 16384, but due to the (α) value 1536 dedicated to the electrical data center, we used the parallelism level, and you see in the design 48 electrical switches and 116 dedicated to optical switches.



Data center-2



POD-2

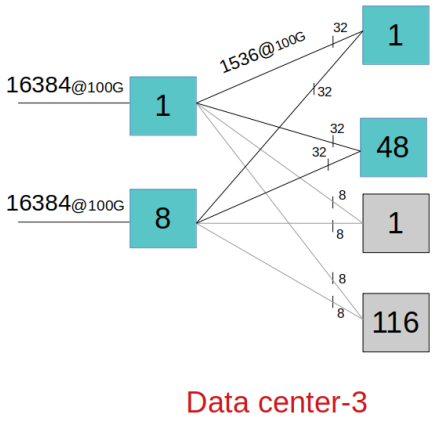


Figure 4.3: Block diagram of electrical switch with over-subscription(total number of ports per each electrical switch is 256@100Gb/s and total number of ports per each optical switch is 2048@100Gb/s)

The table below shows the total calculation of whole three layer hybrid data center. The two-layer data center can support 32768 servers that are not enough servers to support 13.11 Pb/s bisection bandwidth. Therefore, we need to design the pod two-layer; by pod layer, we can extend our data center, and we use it for the next layer data center. In the pod layer, we used half of the leaf switches which we utilized for two-layer data center, and to calculate the number of servers to use for the next layer data center, multiply the P_s by L_s which is 16384, the full servers that we obtained in the pod layer can use it in the next layer of the data center. In a three-layer data center, to support a certain amount of bisection bandwidth, we use parallelism to avoid wasting the switches and cables. To calculate the total switches in three-layer, we have to multiply the total switches in the leaf layer with total switches in the pod layer and summation with switches in the spine layer; in total, 1168 electrical switches need to build this massive data center which is much less than switches that we obtained in the electrical data center.

	servers	Spine switch	total switches	total number of cables	Total number of spine switches(without ToR)
Two layer data center	128 x 256 = 32768	-	256 + 12 = 268	(256 x 12)+ 32768 = 35840	-
Two layer POD	128 x 128 = 16384	128 x 12 = 1536	128 + 12 = 140	(128 x 12)+16384= 17920	-
Three layer data center	16384 x 8 = 131072	-	(140 x 8)+ 48 = 1168	(256 x 1536)+(256 x 17920) = 4980736	(8 x 128)+ 48 = 1072

Figure 4.4: Calculation of electrical data center with over-subscription

4.1.3 Comparisons of system configurations

The table below shows the comparison with the electrical data center and hybrid data center. A multi-stage electrical data center can support massive bisection bandwidth as 13.11 Pb/s, but the number of switches in the electrical data center is 1536, except for ToR, which is a large number of switches and, consequently, high power consumption. On the other hand, the hybrid data center exactly supports the same amount of bisection bandwidth with fewer switches 1072 electrical switches except for ToR, and 116 optical switches have 2048@100Gb/s ports that have a massive number of ports.

	Multi layer electrical switch	Electrical and optical hybrid switch	
		Electrical switch	Optical switch
Bisection Bandwidth	13.11 Pb/s	1.23 Pb/s	11.80 Pb/s
# of optical switches	-	-	116
# of optical links	-	-	(2048 x 2048)@100Gb/s
# of electrical switches (except for ToR)	1536@100Gb/s	1072@100Gb/s	-
# of total links	12582912@100Gb/s	4980736@100Gb/s	237568@100Gb/s

Figure 4.5: Comparison of hybrid optical and electrical data center networks

4.2 Functional experiments

To illustrate my point, let us look at some line graphs. According to the number of servers, the graph shows the software's behavior to choose the number of leaf and spine switches, the number of ports for each switch, and the total number of servers. The calculations are given as the exact numbers that we require to support the servers.

As you see in figure 4.6, the horizontal axis shows the total number of servers and the vertical axis is about the total number of leaf and spine switches, the parallelism of ports, and the total number of ports for servers. I figure out the switch with a small number of ports to be clear the charts. These charts show the calculation which we did for the switches with 30 ports.

Overall, it can be seen the number of leaf switches raise slightly from 1 to 30 because the maximum number of switches that we have is 30 switches. Consequently, the total number of the spine switches is raised according to half of the spine switches and the number of parallelism levels is decreasing by increasing the spine switches; they have an indirect relationship. However, while the number of parallelisms decreased, the number of leaf switches increased to maintain the exact number of servers we need to support. If you look at the trends over many servers, we can see the number of ports for servers changed by the required servers. For instance, to support the 30 servers, instead of using switches with thirty ports, we use the switches with ten ports for servers and three leaf switches to support the exact amount of servers. Furthermore, I have to explain that to find the exact amount of ports for servers to avoid wasting cables and switches; we used the parallelism method; we have to divide the number of servers into leaf switches. Besides, figure 4.7 shows how to layer two of the pod which the server ports are increasing and decreasing widely over 225 servers, and the number of ports using supplements remained reasonably static at approximately 15 switches in leaf layer because we can not use the exact amount of switches which used in the two layer of data center. Between 15 to 225 servers, the number of parallelisms fluctuated up and down depending on the switch's number. Following that, it is stated on two parallelisms.

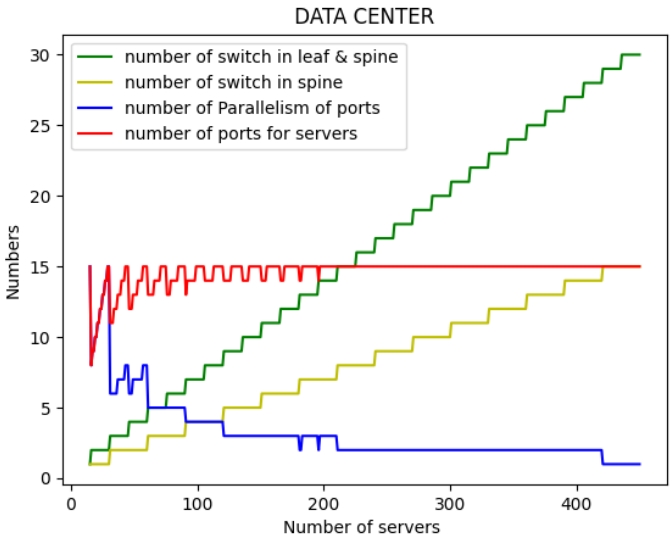


Figure 4.6: Two layer of data center in electrical data center (in this example the total number of ports per each switch is 30@100Gb/s)

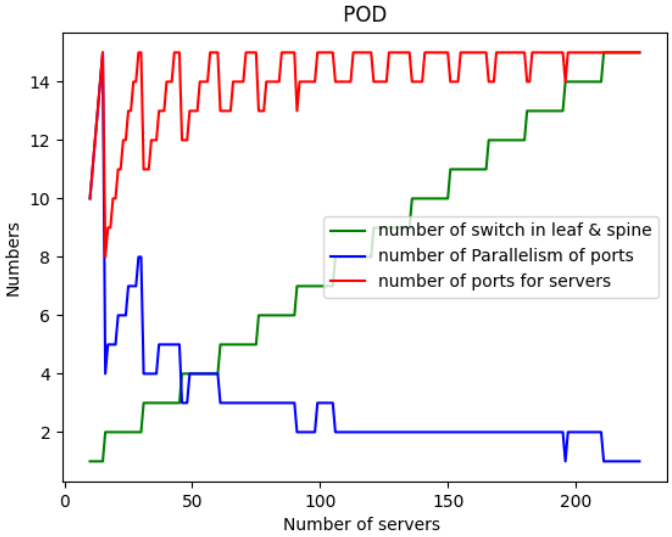


Figure 4.7: Two layer of pod in electrical data canter (in this example the total number of ports per each switch is 30@100Gb/s)

The line graphs in figures 4.8 and 4.9 give you the total amount of switches required to build a hybrid optical and electrical data center, to support between 15 to 450 servers in layer two data center and from 15 to 225 for two layer of the pod, which each switch is with 30@100Gb/s ports.

Looking at the graphs, the total number of switches in the leaf layer increasing slightly by increasing the number of servers in two layer data center increases until 30 switches while in two layer pod increases until 15 switches. In contrast, spine switches in both graphs almost remain constant over the whole number of servers because of value (α) and small port number of switches, which determines large amount of bisection bandwidth is dedicated to the optical rather than electrical switches.

Moving to the number of ports for servers has fluctuated line in both charts which depends on the total number of servers needed, as you see it remains fixed after 200 servers because the number of servers is relatively high and the data center has to support requirement servers. The optical switch has fluctuated line as the same as ports for servers because they are dependent on each other according to the formulas in chapter two of this paper.

In conclusion, the leaf layer's total switches are the same as the electrical data center because we use the same leaf and spine topology; however, spine switches are much less than switches in the electrical data center because of value (α) most of the bisection bandwidth is dedicated to the optical switches. Also, ports for servers and optical switches depend on each other because of the (α) factor, and then after 200 servers, they remain constant according to the server requirement.

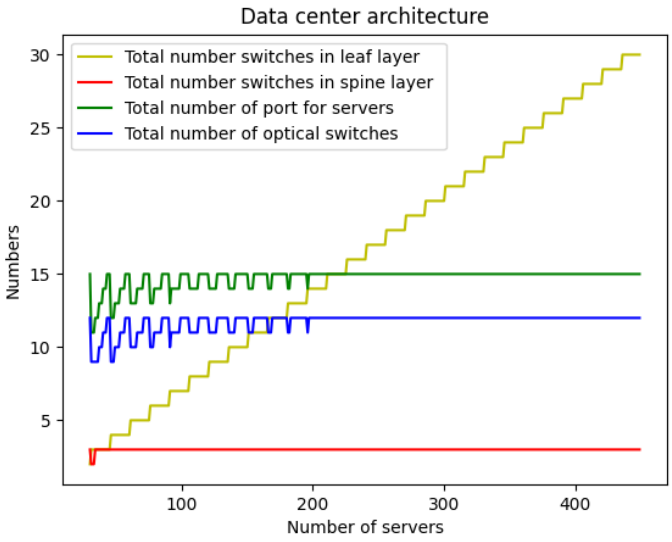


Figure 4.8: Two layer of data center in hybrid data center (in this example the total number of ports per each switch is 30@100Gb/s)

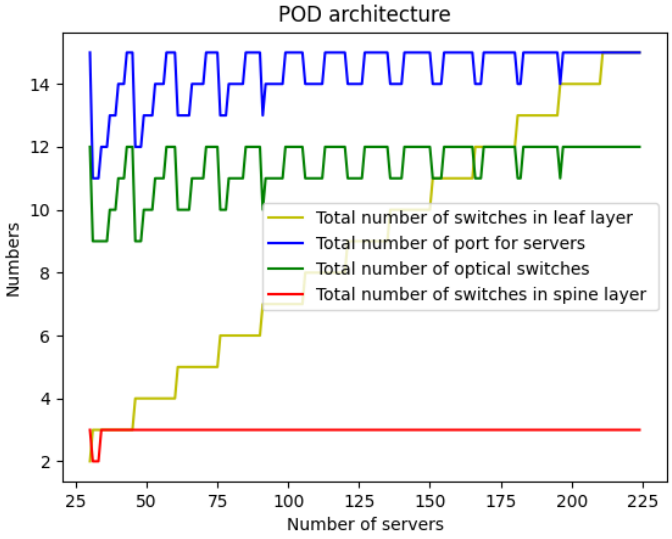


Figure 4.9: Two layer of pod in hybrid data center(in this example the total number of ports per each switch is 30@100Gb/s)

Chapter 5

Design Software

5.1 Software experiment

5.1.1 Graphical view of electrical topology

This chapter shows how the software work. To design the graphs, I used the Python programming language with Networkx library which the nodes are the switches, and the edges are representative of connection cables. As discussed before in chapter two, the graphs are designed by leaf and spine topology. For example, in the figure, 5.2 shows the two-layer of the data center, which supports 50 servers where the nodes in the left part are the total number of servers and the nodes in the middle are the leaf switches that are connected to the higher layer and servers. Spine switches are on the right side of the figure, and the figure 5.2 shows the three-layer of data centers and their connection. The number on edges between leaf and spine switches is the number of parallelisms, which is one because the switches have ten ports that five ports go to the servers and five ports go to the spine switches. Table 5.1 shows all the calculation:

	Number of servers	Spine switches	Total switches	Number of cables between electrical switches
Two layer of data center	10 x 5 = 50		10 + 5 = 15	(10 x 5) + 50 = 100
Two layer of POD	5 x 5 = 25	5 x 5 = 25	5 + 5 = 10	(5 x 5) + 5 = 30
three layer of data center	10 x 25 = 250		10 + 5* = 15	(10 x 25) + (10 x 30) = 550

*: Here, instead of 25 spine switches, we can use five switches by parallelism method, which is shown in figure 5-4.

Figure 5.1: Calculation of three layers of electrical data center

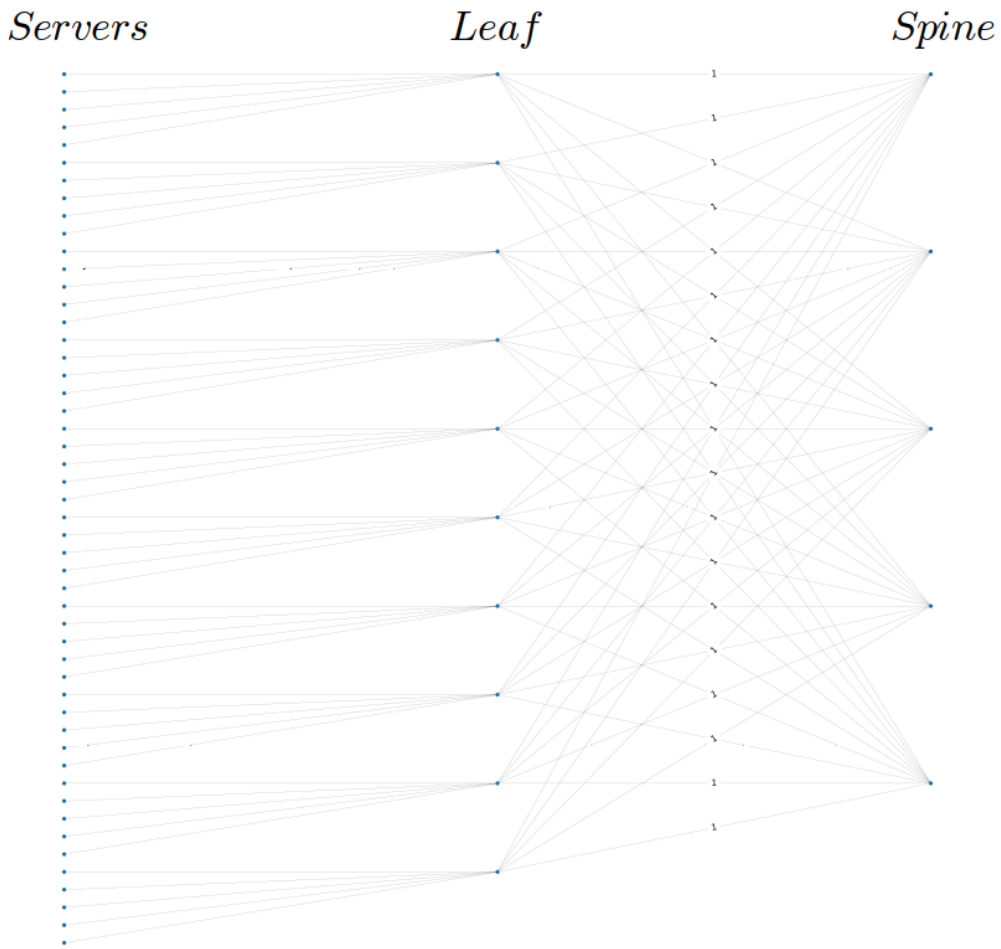


Figure 5.2: Leaf and spine topology two layer data center (total number of ports per each electrical switch is 10@100Gb/s)

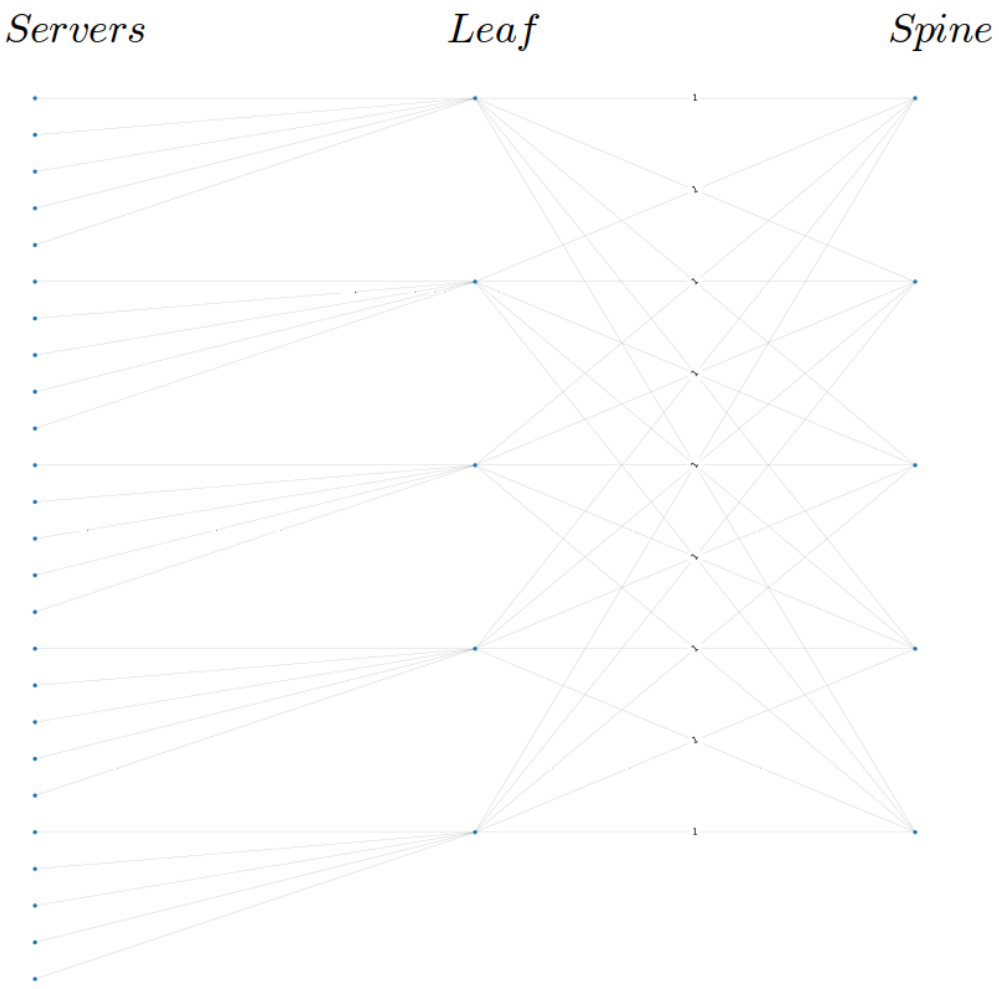


Figure 5.3: Leaf and spine topology two layer pod (total number of ports per each electrical switch is 10@100Gb/s)

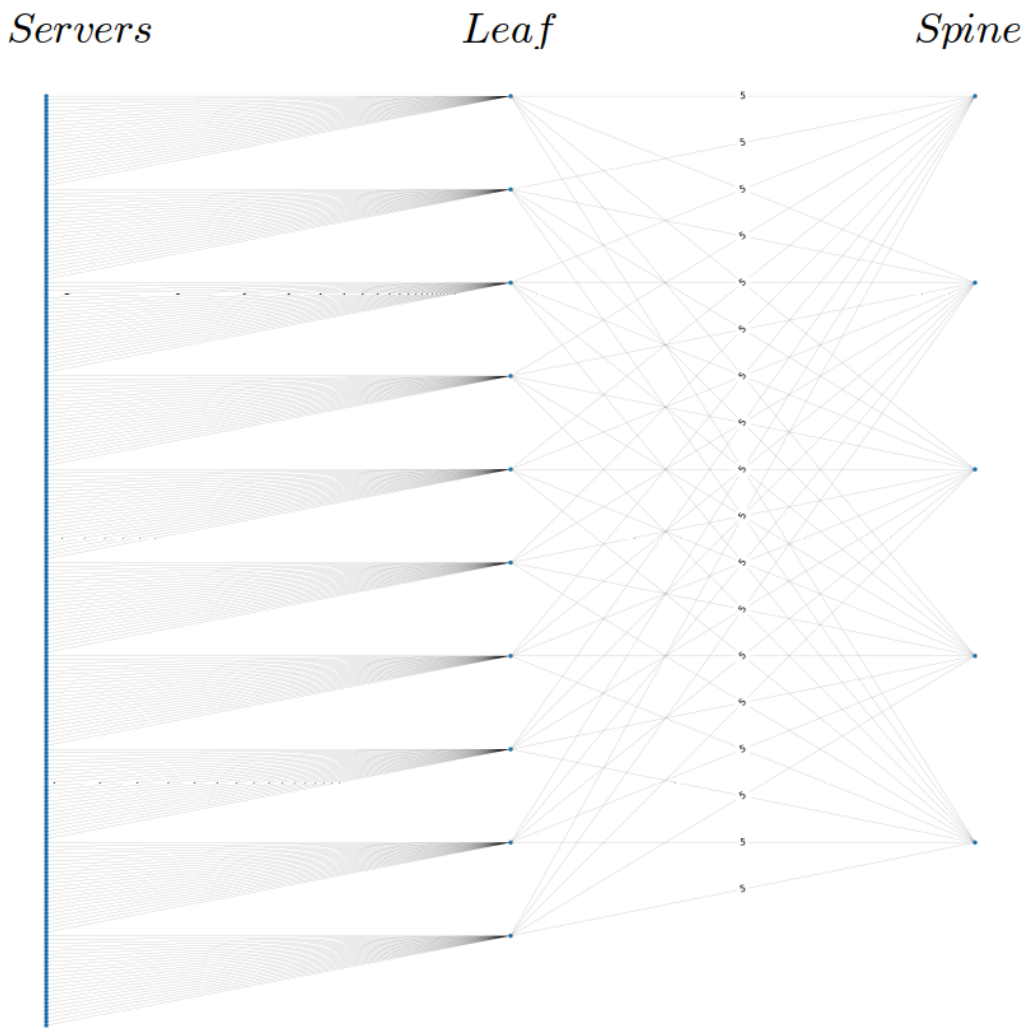


Figure 5.4: Leaf and spine topology three layer data center (total number of ports per each electrical switch is 10@100Gb/s)

5.1.2 Graphical view of hybrid topology

In this part, you will see the hybrid optical and electrical data center graphs, which I show in two colors to simplify the topology and clarify the connection. The nodes are the number of switches that we used to build these data centers, and for the connection between them, I used the edges. The parallelism level is the number on the connection between leaf and spine switches. The black edges connect the leaf to spine switches which construct the electrical data center, and the gray edges are the connection between electrical switches to optical switches, which create the hybrid data center.

According to the design of a practical large data center like Google, let us consider, for instance, an (α) factor equal to four; I justify the choice of this using a real data center. Notice that I apply the (α) factor from server to ToRs. Mention that I avoid designing the server nodes in the hybrid data center because of the large number of servers, and these graphs show the leaf and spine architecture without servers. This data center is designed as the same as the previous electrical data center with one difference which is applying (α) value and splitting the bisection bandwidth between two types of switches. In table 5.5 there are all the calculations to create a hybrid data center. I have to mention that to build this data center, the total number of the port for each electrical switch is 256@100Gb/s, and the total ports per each optical switch are 2048@100Gb/s.

	Number of servers	Spine switches	Optical switches	Number of cables between electrical switches
Two layer of data center	10 x 5 = 50	5 x (1/α*) = 1	5 x (1 - 1/α) = 4	(10 x 5) + 10 = 60
Two layer of POD	5 x 5 = 25	5 x (1/α*) = 1	5 x (1 - 1/α) = 4	(5 x 5) + 5 = 30
three layer of data center	10 x 25 = 250	25 x (1/α*) = 6	25 x (1 - 1/α) = 19	(10 x 25) + (10 x 6) = 310

*: (α) is the over-subscription value which according to the design of a practical large data center like google, let us consider,for instance, an alpha value equal to four.

Figure 5.5: Calculation of three layers of hybrid data center

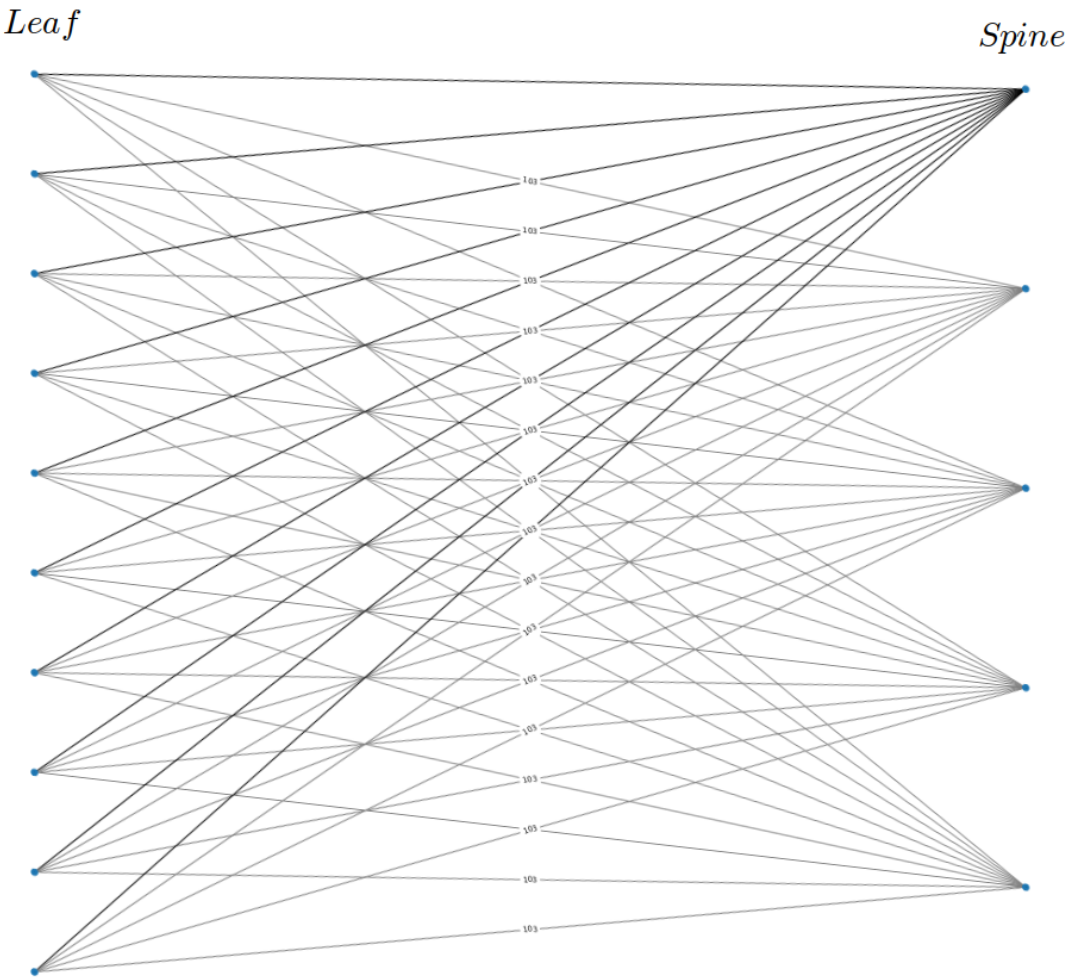


Figure 5.6: Hybrid optical and electrical two layer data center (total number of ports per each electrical switch is 10@100Gb/s)

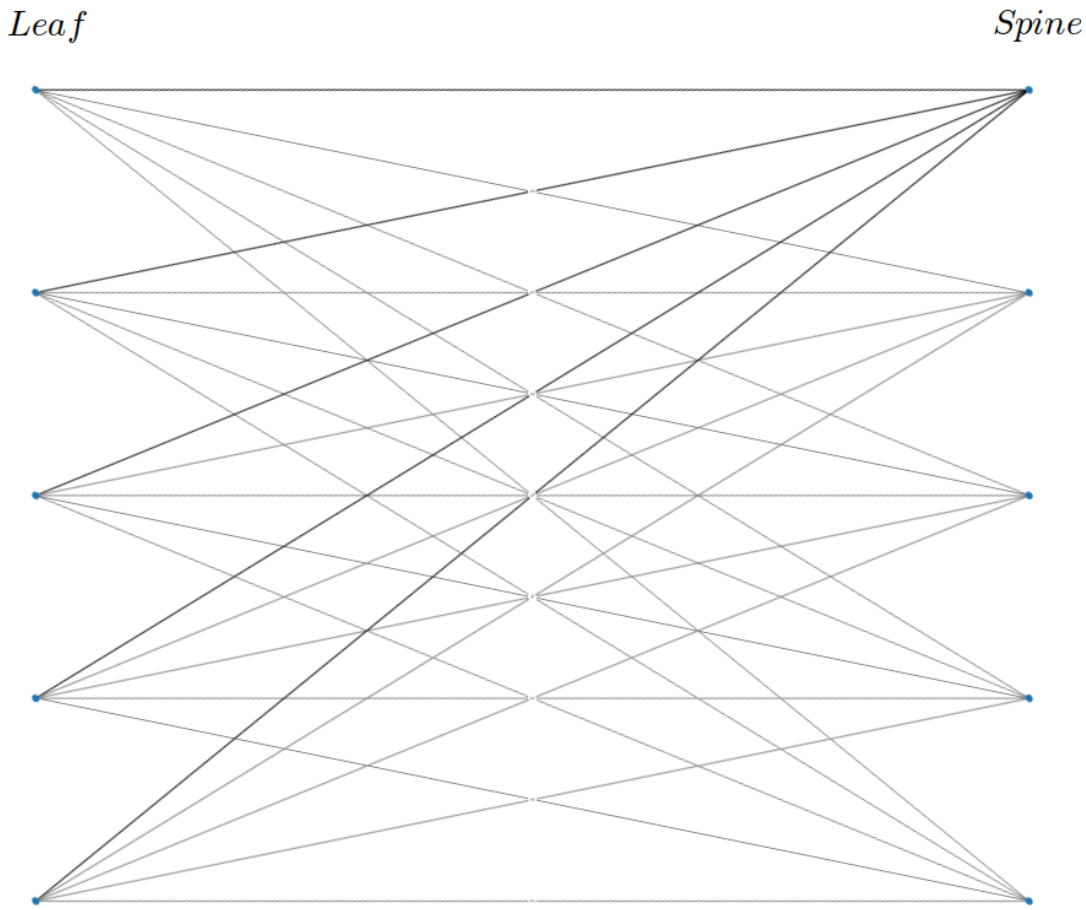


Figure 5.7: Hybrid optical and electrical two layer pod (total number of ports per each electrical switch is 10@100Gb/s)

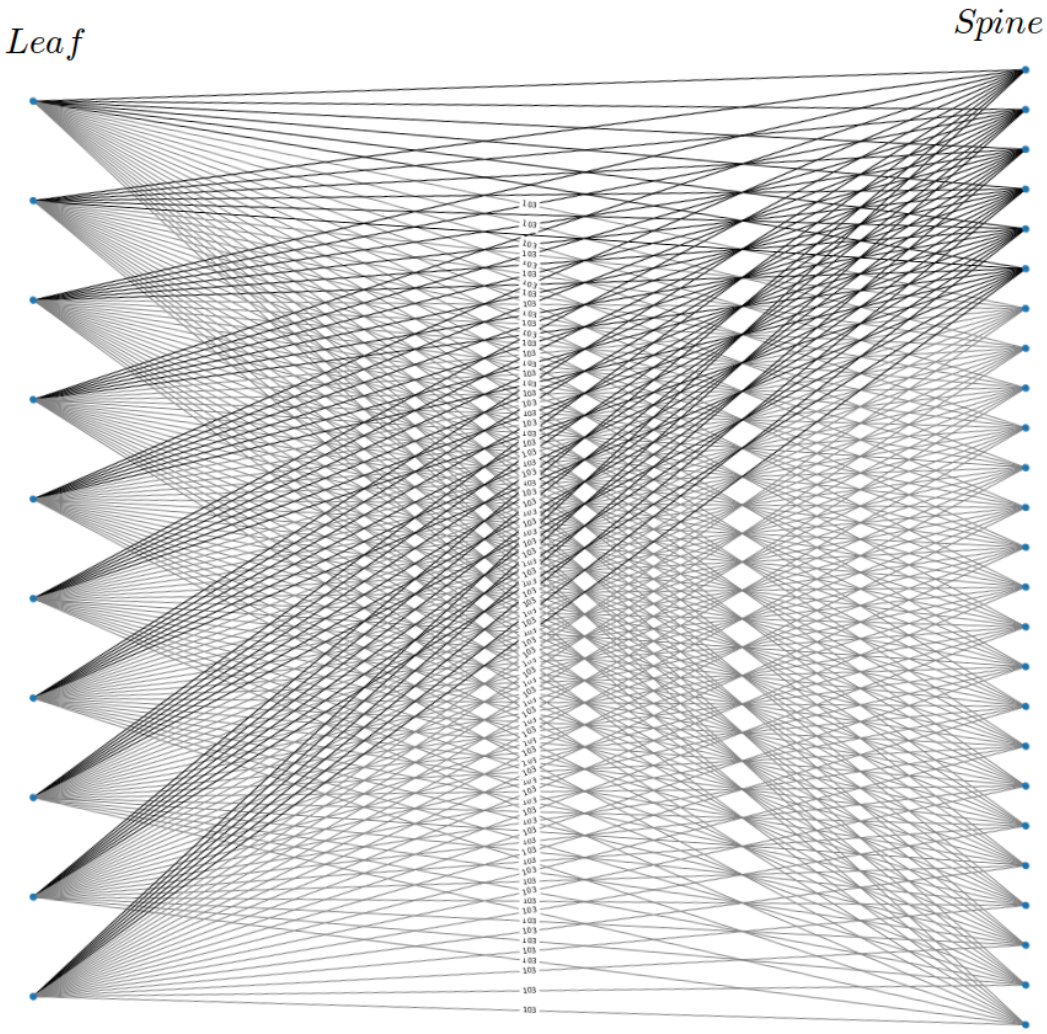


Figure 5.8: Hybrid optical and electrical three layer data center (total number of ports per each electrical switch is 10@100Gb/s)

Chapter 6

Conclusion

In this part, I try to summarize what I learn in this thesis, according to the classical data center, which constructed by full electrical switches has the problems such as the high power consumption and a large number of switches. Optical switches are introduced to solve these problems, with optical interconnects moving towards Tb/s scale throughput to follow up with the requirements of massively growing traffic in the data center. Many optical technologies for current warehouse-scale data centers are evolving, especially technologies that allow optical systems with an extensive integration and technologies that offer large-scale fabrication at low cost [3].

In this paper, the Clos network is a multistage circuit switching network illustrating a technical idealization of practical, multistage switching systems. Clos topology is applied to build a leaf and spine architecture. The leaf layer is the lower layer connected to the ToRs and servers, connecting to the data center's core layer. The leaf layer for both data center networks used the electrical switch due to the lack of buffer in optical switches and high cost-effective. However, instead, we can profit from fact optical technology and leverage them in the spine layer. As you observed in the last chapter, by the factor (α), we can divide the bandwidth into two types of switches to improve efficiency. The CPU traffic to storage is configured with the electrical data centers, and the data from CPU to memory is dedicated to the optical switches.

Disaggregation is a theory in which similar sources are combined, with the various sources being individually updated and the operation configured for optimized execution. The network can be disaggregated at various levels, for instance, at the rack or server scale. The disaggregated data center needs an interconnection fabric that requires additional traffic produced by the disaggregation and high bandwidth and low delay to control and enhance achievement. The system needs a switching fabric to adaptively provision the Disaggregated rack places resources of different types in different parts of the data center than conventional servers

and uses networking to combine and create needed resources unitedly. Although packet-switched networks reside electrical, optical circuit switches are the best candidates for re-configuring sources in the disaggregated network. Data center architectures with optical switch fabrics have been introduced to promote high bandwidth performance and source utilization. Typical latency to memory, which in the standard server where the memory is near the CPU, is on tens of nanoseconds [3].

To built the data center in the second chapter, I followed the paper with the title " How Optical-Circuit/Electrical-Packet Hybrid Switching will Create High Performance and Cost-Effective Data Center Networks " [6], which the required bandwidth to support is 13 Pb/s. To produce this data center, we need to design a three-layer, two layers of the data center, and one layer of the pod. I have to mention that the number of layers depend on the required number of servers the data center has to support. I chose a large number of electrical switches with 256@100Gb/s ports and optical switches with 2048@100Gb/s ports to build the hybrid optical and electrical data center. Moreover, according to the formulas could determine the total number of switches for both electrical and optical also the ports for each switch.

In the last chapter, I simulate the leaf and spine topology (electrical and hybrid data center)by a Python programming language that worked with the Networkx library. In the graphs, you can observe the connection between the switches that entire switches are connected and the data transfer from south to north and east to west.

Bibliography

[1] M. Alizadeh and T. Edsall. «On the Data Path Performance of Leaf-Spine Datacenter Fabrics». In: *2013 IEEE 21st Annual Symposium on High-Performance Interconnects*. 2013 (cit. on pp. 1, 2).

[2] A. Reale and D. Syrivelis. «Experiences and challenges in building next-gen optically disaggregated datacenters : (Invited Paper)». In: *2018 Photonics in Switching and Computing (PSC)*. 2018 (cit. on pp. 1, 12).

[3] Qixiang Cheng, Meisam Bahadori, Madeleine Glick, Sébastien Rumley, and Keren Bergman. «Recent advances in optical technologies for data centers: a review». In: *Optica* 5.11 (Nov. 2018), pp. 1354–1370 (cit. on pp. 1, 2, 9, 12, 13, 49, 50).

[4] *T Day One Data Center Fundamentals*. `https://g.co/kgs/5HBAAK` (cit. on pp. 2, 7).

[5] K. Sato. «Realization and Application of Large-Scale Fast Optical Circuit Switch for Data Center Networking». In: *Journal of Lightwave Technology* (2018) (cit. on pp. 3, 9–11).

[6] K. Sato. «How Optical-Circuit/Electrical-Packet Hybrid Switching will Create High Performance and Cost-Effective Data Center Networks». In: *2019 21st International Conference on Transparent Optical Networks (ICTON)*. 2019 (cit. on pp. 3, 10, 21, 50).

[7] Hitesh Ballani, Paolo Costa, Istvan Haller, Krzysztof Jozwik, Kai Shi, Benn Thomsen, and Hugh Williams. «Bridging the Last Mile for Optical Switching in Data Centers». In: *Optical Fiber Communication Conference*. Optical Society of America, 2018, W1C.3 (cit. on pp. 4, 5).

[8] Brad Hedlund. «Top of Rack vs End of Row Data Center Designs». In: *Apr* 5 (2009) (cit. on pp. 7, 8).

[9] *pod Point of delivery*. `https://www.google.com/search?q=Point+of+delivery+(networking)&oq=Point+of+delivery+(networking)&aqs=chrome..69i57j0i22i30j0i390l2j69i60.528j0j7&sourceid=chrome&ie=UTF-8` (cit. on p. 8).

[10] William M. Mellette, Rob McGuinness, Arjun Roy, Alex Forencich, George Papen, Alex C. Snoeren, and George Porter. «RotorNet: A Scalable, Low-Complexity, Optical Datacenter Network». In: 2017 (cit. on p. 8).

[11] Mihail Balanici and Stephan Pachnicke. «Hybrid Electro-Optical Intra-Data Center Networks Tailored for Different Traffic Classes». In: *J. Opt. Commun. Netw.* 11 (Nov. 2018), pp. 889–901 (cit. on p. 10).

[12] P. J. Roig, S. Alcaraz, K. Gilly, and C. Juiz. «Modelling a Leaf and Spine Topology for VM Migration in Fog Computing». In: *2020 24th International Conference Electronics*. 2020 (cit. on p. 15).

[13] «Dynamic resource allocation in hybrid optical–electrical datacenter networks». In: *Computer Communications* 69 (2015), pp. 40–49 (cit. on p. 21).

[14] M. C. Wu, T. J. Seok, K. Kwon, J. Henriksson, and J. Luo. «Large Scale Silicon Photonics Switches Based on MEMS Technology». In: *2019 Optical Fiber Communications Conference and Exhibition (OFC)*. 2019 (cit. on p. 21).

[15] Yuxin Cheng, Matteo Fiorani, Rui Lin, Lena Wosinska, and Jiajia Chen. «POTORI: A Passive Optical Top-of-Rack Interconnect Architecture for Data Centers». In: *J. Opt. Commun. Netw.* 9.5 (May 2017), pp. 401–411 (cit. on p. 31).