



Advanced Computer Networks
Dr Sameer Kulkarni
Department of Computer Science Engineering
Indian Institute of Technology, Gandhinagar
Lecture 50
Data Center Networking - Characteristics and Challenges

(Refer Slide Time: 0:18)


DATA CENTER NETWORKS

- 10's to 100's of thousands of hosts, often closely coupled, in close proximity:
 - e-business (e.g. Amazon)
 - content-servers (e.g., YouTube, Akamai, Apple, Microsoft)
 - search engines, data mining (e.g., Google)
- **Challenges:**
 - **Scale:** multiple applications, each serving massive numbers of clients
 - **Performance:** managing/balancing load, avoiding processing, networking, & data bottlenecks
 - **Power:** minimize carbon footprint; effective cooling and heat dissipation.
 - **Fault-isolation:** localization and containment
 - **Failure resiliency:** operate even when the failures occur.



Inside a 40-ft Microsoft container, Chicago data center

Data Center Networks

Advanced Computer Networks



So, we looked at the data centers in terms of how they are built and what are the core components. Now, when we consider the networking part of the data centers, we have to consider many other aspects in terms of how the networking is laid out, and what are the challenges that we should be addressing when we are trying to deploy the networks for data centers.

So, primarily, we spoke about different kinds of applications that these data centers support in terms of e-business applications like for example, consider Amazon as a website, and you see a host of applications that they provide. And likewise, the Content Service Servers or Content Service Providers that Akamai etc facilitate for shipping multiple of the web servers onto their domain and facilitate connectivity from the nearest possible location. Likewise, for the content you see as YouTube, Microsoft services, and even search engines, etc. So myriads of applications, not just the web servers, but also the content some are say latency sensitive, some are throughput sensitive, all kinds of applications exist on the network that are being facilitated through these data centers.

Hence, the foremost of the challenge that we would also need to think in terms of first is how to enable the scale for multiple applications, where each is serving a massive number of clients. That means the network should not be a bottleneck in facilitating the requirements of connecting multiple users, or in terms of if it translates to network, it is equivalent to saying a large number of flows. And when we say multiple of data, maybe database servers, data like streaming servers that are going to be accessing the data, the scale also means in terms of the bandwidth that we would want to support.

Again, the second aspect is the performance, where if there are multiple of the servers that are serving similar kinds of content, how the network would be able to help facilitate or manage to balance the load across different servers? And if in any way, a network component is a bottleneck, what are the means to mitigate and overcome such bottlenecks and evade or avoid such data bottlenecks?

And third, is the power I mean, when we have lots of networking devices, also, we spoke earlier about the middleboxes, which also are often present in large numbers in such data centers. So, when they are powered up, our goal also is to minimize the carbon footprint or be powered in a more efficient manner. And also whatever the heat that they are going to generate and the cooling mechanisms that need to be addressed for these specific devices be done appropriately.

And often, if there is no network load or no traffic that is coming, you would want to shut off such a device in terms of the network and bring it up only when there is a need for having access to such a device. So, power management at the whole data center level as well as at the networking level becomes crucial for better data center management.

And third and most important is typically whenever there are failures, you would want to identify and detect such failures quickly and contain those failures. So, localization meaning trying to identify, detect, and pinpoint where the failures have happened and isolate such failures from propagation, is also a critical aspect. When you look at lots of cyber threats that these data centers or in general the networks are open to. Hence, the containment of such failures also becomes a critical aspect. And especially when it comes to the networks, we are now talking about the network links in the orders of several thousand to hundreds of thousands.

And whenever one of the links fails, rerouting ensuring that the updates are being done in a timely fashion and you avoid the root loops and you avoid black holes becomes a critical concern when it comes to networking devices. Also, when such failures are seen, especially lots of links flap, and whenever there is such a link flapping or temporary intermittent failures you would want to operate and ensure that the services still function without having to see any downside of having one or several of such links to be down that means we need to build the resiliency in the system, which can enable us to overcome any of the intermittent failures and yet facilitate the service to run as if no failures have happened.

And hence, building such a failure resiliency in terms of the data center networks becomes another crucial aspect. So all these, in effect, are also the requirements. And at the same time, doing those for thousands of devices is also a major challenge.


(Refer Slide Time: 5:45)


DATA CENTER COSTS

Amortized Cost*	Component	Sub-Components
~45%	Servers	CPU, memory, disk
~25%	Power infrastructure	UPS, cooling, power distribution
~15%	Power draw	Electrical utility costs
~15%	Network	Switches, links, transit

**3 yr amortization for servers, 15 yr for infrastructure*

- Total cost varies
 - upwards of \$1/4 B for mega data center
 - server costs dominate
 - **network costs are significant (both capEx and opEx)**
- Long provisioning timescales:
 - new servers purchased quarterly at best





Data Center Networks
Advanced Computer Networks

So, in this aspect, let us try to understand a bit about the data center costs in terms of where the network point of view stands and what is the overall cost factors for which data centers account. So, this was a study done in early 2010. And what it showed was that the servers are the CPU, memory, and disk components of a server machine in a data center, that is hosted and costs to roughly around 40 to 50 percent of the data center cost.

And the power being the another important component that added to the cost roughly around 20 to 25 percent while the network added to around 15 percent, roughly, for the overall data center cost. And when we say network, that is basically the switches, the links, the transit links, the management of all of them. So, we can see 15 percent is not a small component when you are talking about billions of dollars that you had to invest for bringing up the macro data centers.

So, it becomes important to manage these devices properly, so that we are able to both reduce the capital expenses, as well as operational expenses. And in this regard, we saw how NFV helps lower capital expenses and SDN with centralized control helps lower operational expenses, ensuring that you have centralized control over all of these devices. Think of millions of such routers, millions of switches, and routers that are being configured and controlled, you would not have enough manpower to handle or manage these unless you centralize and handle them more efficiently.

Thus, SDN and NFV, we can see that they have really played a great fit towards the data centers and also enabled us to minimize these costs. And another important aspect is that when we think of servers or networking components, every year, there is an improvement, and the servers typically get better and better as we progress. And typically within 3 to 4 years, you would have to replace it with a new server.




So, we would want to ensure that the monetization cost of the servers, that is, to reclaim the amount of cost that we would have put or whatever the cost that we put, how quickly does it zero out in terms of providing the value worth for the bulk is in terms of typically around 3 years. But when we look at the infrastructure as a whole, it takes a lot more time because you invest a lot more of the investments towards the infrastructure, and it takes around roughly 15-20 years to amortize the cost on the infrastructure.

So, what this tells us on the other side is that we are going to be upgrading our servers, upgrading our networking devices more often in a given data center. So, when we are going to upgrade, it should not take a long time to provision and patch the updates or replace the hardware. Hence, whenever the new service are purchased or new networking devices are upgraded, we need to ensure that we are able to patch them up and update with as soon as quickly as possible without having to have longer provisioning time scales. This is also a crucial aspect, which would add

otherwise to the cost and any downtime incurs a negative penalty for any service. So, hence, we want to also ensure that we do not pay cost in these aspects.



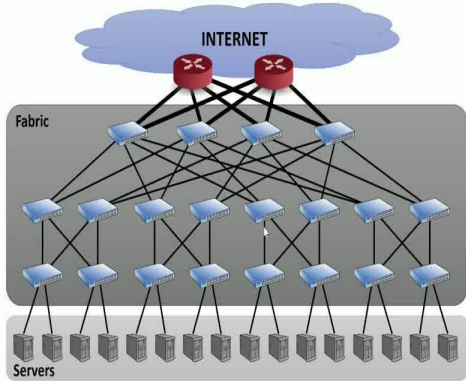
(Refer Slide Time: 9:31)

WHAT'S DIFFERENT ABOUT DCNs?



Data Center Networks Advanced Computer Networks

WHAT'S DIFFERENT ABOUT DCNs?



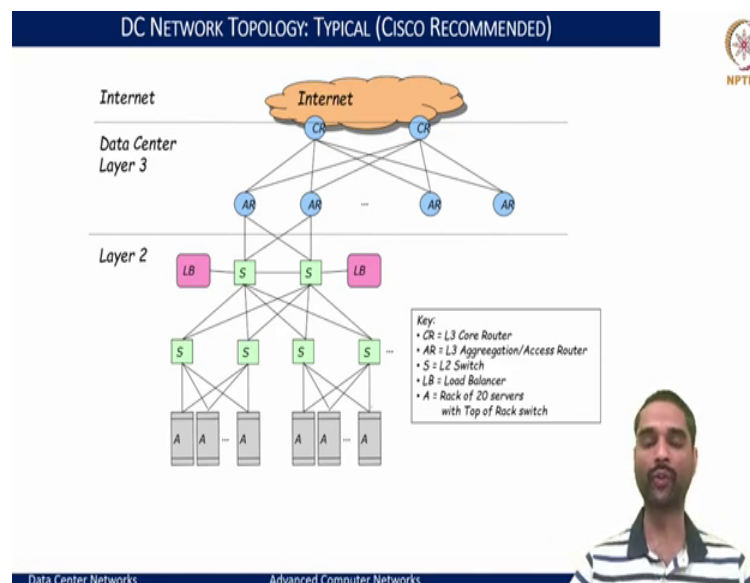
Data Center Networks Advanced Computer Networks

So, the question now comes like, what is so much different about data center networks? if we are already managing so many of the web servers, the file servers, and data center networks, which way do they differ? And if we look back in this, we are talking about a huge networking infrastructure, which is enabling access to the servers from across the internet. And if we think of just the inside of a building where the data centers are hosted, we would see that the links

themselves are in the span of several hundred thousand of networking links and networking devices that we are operating with.

And this is basically a mini internet kind of framework that we have to deal with. But all under one control, meaning we have the flexibility to weave how the network components should be connected as data center owners have the flexibility to decide how the servers would be connected to such networking infrastructure, what is the networking infrastructure, we would want to build inside of the data center, and how we would want to enable connection to the outside world through the use of the routers and connect them with the internet.

(Refer Slide Time: 10:56)



So, these aspects typically were studied in the early 90s. And by the early 2000s, late 90s, in 2000s. Typical network topology in terms of what should be the way the network should be built, if the data centers have to be scaled, and if the data centers have to have the means to upgrade and add and scale the components within this data center. And Cisco recommended this three-tier or multi-tier architecture. And this multi-tier model has been the most common model that is been used in enterprises today.

And this design can cater to any of the web applications, database servers, and various of the platforms, in terms of what kind of servers we would want to deploy heterogeneous mainframes multiple of the rack servers, all of these could be constituted in this architecture or a topology as

shown here. The data centers, aspects if we have to consider what it shows at the top most is the internet layer through which it facilitates connectivity for a data center to the external world. And to do that, at the boundary, we see what we call as these core routers.

And these core routers are basically the bridge that facilitate us to connect to the external world and connect the underlying network infrastructure that we will want to weave for our data center. So, these are basically the topmost points for starting the data center fabric. By fabric, we mean what kind of network topology, what is the link bandwidths, what kind of switches, and what are the capabilities of each which is all that the data center owner decides.

And this is where the data center's layer three networking starts. And these core routers, which connect to the internet inside the data center, would connect to what we call as the aggregation routers, where you can see that now, we are spreading this network to multiple of the aggregation routers so that we are able to disaggregate the incoming traffic and channelize the traffic for a particular operation and also scale these aggregation routers as and how we want to scale the data centers.

And both the L3 core routers and L3 aggregation routers facilitate what we call as the data centers layer three mechanism. And this layer essentially is a gateway for connecting the inside and outside worlds. So, for all the devices that are within the data center, the aggregation routers act as the gateway to connect to the outside world, while the core routers act as the means to connect the external entity outside of the internet to the entity that is inside of the network. And this aggregation layer with many access link layers that can be linked further down to provide what we call as the layer 2 network within the data center.

And this layer 2 network is much more critical in the sense of scaling up of the resources that we would want to do be it the deployment of the middleboxes with the deployment of the switches or the server racks that we spoke about earlier in terms of how we want to scale. And at layer 2, all these marked as S are the switches, would enable us to operate in the layer 2 domain, meaning we can operate and ensure the connectivity amongst all of these interval devices. Just at layer 2 as if it is a single LAN, and we could even isolate them by the use of VLANs and ensure connectivity accordingly.

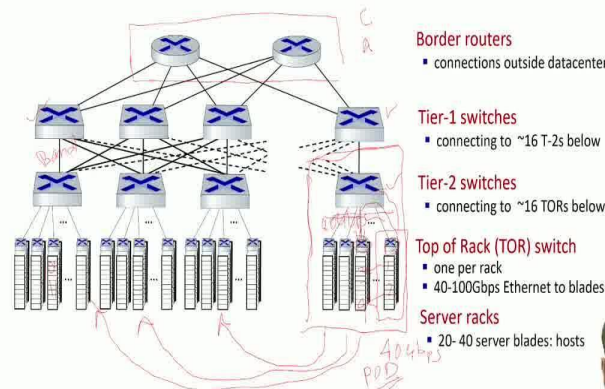
But in terms of connectivity, we could have a simple layer 2 connectivity rather than trying to do layer 3 connectivity through the aggregation routers. What this also means is if I have any of these server racks underneath, we should be able to facilitate communication just at layer 2 without the need to go to the aggregation routers. We would need to go to the aggregation routers only if we have to connect outside of our domain and use the aggregation routers as gateways to facilitate connection to the outside world.

And this mode plays a critical role in meeting several of the requirements such as like if we want to upgrade the hardware switches, we can do them much more easily and provide better bandwidth, upgrade the infrastructure resources within the layer 2 space. And typically this access layer is the first oversubscription point in terms of, how much should be the link bandwidths that we would want to support at each of the layers can be defined with respect to each layer in terms of what is the traffic that we expect to move when it is between the rack when it is feeding the rack and across the rack or in between the racks.

And think of two racks trying to communicate with each other, then the connection has to go through these switches like what we are showing at least at the bottom half the access switches to connect to two nodes that are on different racks. And likewise, we may have to go to the other tier, the higher tier of these switches to connect to the racks in a different region. So, we can scale these resources and we can also set up what should be each of the link bandwidths that need to be set up.

And this is the most typical network infrastructure that was being used in the early 2000s for enterprise data centers. And then most of the links used to be like 10 Gigabit-Ethernet channels that used to enable traffic communications across these devices.

(Refer Slide Time: 17:40)



So, let us now look at the components that we just discussed in such a network topological architecture in more detail. So, at the bottom, what we typically see is a rack, and each rack hosts a bunch of servers. Like we said earlier 42U rack could host around 40 of the 42 single U (1U) racks. But if they are 2 U or 4 U devices, the number would go down accordingly. And typically, we expect around 20 to 40 servers within each rack.

And at each rack, or what we call as a server rack, there will be one switch that is being set, which is enabling connection amongst all of the servers that are present within each rack so that if there are two servers, they would be able to communicate within themselves using this switch that is sitting on top which we call as the top of the rack switch. And this would be one switch per rack, and a typical link bandwidth facilitated through this switch is around 4200 gigabits ethernet.

So, if two devices that are on this rack want to communicate with each other, they will get around 100 gigabits per link bandwidth to ensure that they can communicate with each other. And we may have like 40 of them each connected through this switch, each getting around 4200 gigabit links connecting them. And these become the units of scale at the lowermost end, where if we are speaking of hosting around hundreds or thousands of servers. So, if we speak around thousand servers, then we would need at least 50 of such racks so that you are able to host 100 servers, considering approximately 20 servers per rack.

And if you are speaking of orders more like 10,000 then accordingly the number of server racks that you would want to deploy would increase. And as these server racks increase, now if we want to communicate amongst these racks, then this is facilitated through what we call as a tier 2 switch that sits and connects with multiple of these racks, so we can see that this tier 2 switch now is able to connect to around multiple of these approximately consider 16 top of rack switches can become connected through these tier 2 switches.

And now, we can enable connections to ensure that we can have communication between 1 and 3 or 2 and 4, to go through these tier 2 switches, and to facilitate this connection. Now, the question comes, What should be the bandwidth for the links of this tier 2 that are having the down links that are connecting with each of these top of the rack? Should they be again 4200 gigabits, or should they be more? So, this is where the principle of over-subscription comes. And if we say that each of the links that are down here are 40 gigabits, and each of the links that are connecting to this tier 2 is also 40 gigabits.

Then, we basically have the same bandwidth, but in terms of the communication now, we can have if we had just two of these servers, each between 1 and 2, it would get 40 gigabits per second. And between 1 and 3, it will also get 40 gigabits per second that it will communicate. But now, as you scale the racks, this communication is going to be accordingly shrinking in terms of what bandwidth they would get when all of them would go. And this results in what we say: the more you add, the subscription becomes much more on those links, which is much more than the thing that you could basically support.

So, if all of these 1 and 3 also want to communicate and 2 and 4 also want to communicate, you have a requirement of 80 gigabits per second bandwidth.

But since these links are only 40 gigabits, now, we have all subscribed it by 2, so 1 is to 2 and likewise, if you increase the number, you will see that it will start dropping. And that is how the oversubscription typically happens. And if you want to avoid, then we would want each of these links at this point to be 80 gigabits per second so that they are able to support the traffic. And that is where we will see that there is an interplay of how you want to build the links and what kind of bandwidth they should be supporting for each of the connections.

And like we said, we are not confined to a small set of racks; we are talking about hundreds to thousands of such racks. So, what this would mean is we would be replicating such a unit multiple of the times to ensure that we are able to meet and scale out the required resources.

And when we scale out now, if we want to interconnect and communicate between this, we need another mechanism. And that again, we can think on the same lines to be facilitated through the tier 1 switches wherein you have a tier 2 switch connected to this tier 1 switch. And this tier 1 switch would then connect to every other tier 2 switch in the network. So, all the tier 2 switches are now connected with tier 1 switches. And ensure that I can communicate amongst any of these through the pair of tier 1 and tier 2 switches that I want to go.

So, if I had 12, and I say I have 10 and 11, if 1 has to communicate with a device 10, it would go through the switch, and then it has an option to go either to this switch and then come here or it has an option to go here and then connect to the devices 10 and 11. So, likewise, now we are trying to build lots of possible routes through which we are able to connect the two ends of these and typically what we call this instance of a scale as a POD, which constitutes of multiple of such racks connected with a tier 2 switch as a scaling unit that would then be connected to a tier 1 switch.

And again we can scale these tier 1 switches and if you consider 16 of such things now we have 16 x 16 x 40 of such servers that we can basically connect in this network. And again, as this network has to connect to the outside world, we would need these border routers, and again, these border routers can be layered as we saw as either the core and aggregation routers or simply the core routers that enable us to connect outside.

So, this is a common means of setting up the data center network elements. And we spoke about this top-of-the-rack switch, which enables such as this module for each rack to connect to the tier 2 switches, and tier 2 switch as an entity that would allow us to scale at a level of POD. And tier 1 switch allows us to communicate amongst multiple of such PODs, and the border routers facilitate to connect to the outside world. And thus, we can see that we can operate with scales of tier 2, and tier 1 switches to ensure that we can connect each and every entity within the data center and also with the outside world.

And also, when we speak of these core routers again, the bandwidth for these links, should they be, again, 40 gigabits or 100 gigabits? How should we determine? This becomes a concern and that is where many of the different kinds of topologies were brought to see which one would fit better.

And we can also observe another aspect like when we want to communicate within a rack, we have one hop devices going through just the switch. And when we want to communicate within a POD, we have two hops, that is, one top of the rack switch, then going to the tier 2 switch and then coming back to the top of the rack switch for that corresponding destination so that you have at least two hops before you connect to the end host.

And when we have multiple of these PODs that want to connect, we are having one more hop of two hops that is being added by the tier 1 switches to go what is the source tier 1 switch through which you can reach the other end. So, you have one top of the rack going to one of the tier 2, then going to tier 1, then coming back to tier 2, and then to the corresponding top of the rack and then to the device.

So, you can see that in terms of latency when we increase or scale out in this fashion, we would also incur additional excess latency and additional hops through which we have to navigate. And this is pretty common when we consider this kind of model.

(Refer Slide Time: 27:18)

ToR vs EoR/MoR

- Top of rack (ToR): is also known as In-Rack design. The network access switch is placed on the top of the server rack and all servers in a rack directly connect to the network access switch.
- End of Row (EoR): there is a direct connection of each server in the rack with the end of row aggregation switch. This eliminates the need of in-rack switches.
- Middle of Row (MoR): Similar to EoR, instead the Aggregation switches are placed in the middle of the row to avoid long cabling.

Data Center Networks Advanced Computer Networks

And it is not just that we have top of the rack as a means to connect. There were also other models that were identified and see which would suit up, and to put it like what we have seen as a top of the rack, each rack has a switch as a top element. And all of these servers would connect with this switch first. And then this is the one that connects to the aggregation switch or the tier 2 switch that we just saw.

So, this is also called as In-Rack design. And in this, the network access switch is always placed at the top of the server, and all the servers interact directly and connect to this network access switch. Although the top here means in the rack, it is not necessarily that you place the switch just at the top; it becomes simpler as a convention to put it, but you could as well put a switch somewhere in between the rack there is no such binding that it has to be just at the top, but the convention that is being used with each rack hosts a switch, and this switch is then connecting with the aggregation switch.

And what you can see now like as I scale my racks, I need one switch to be purchased and set up. This in terms of cost will be expensive when in terms of management also, we need to handle a lot more of the switches. So, alternatives were thought and two of the models were proposed one is called end of the row wherein when you connect each of these racks, the racks need not have any switch which interconnects them. Rather, I would bypass this top-of-the-rack switch and connect all of these devices directly to the aggregation switch.

So, what that means is you will have lots of wires that will go from each of the racks and connect to the aggregation switch, which might be placed at the end of the row so that you have one place where the aggregation switches are placed and you are cabling all of the racks, server cables to connect to that aggregation switch. So, what this means is now there is a direct connection for each server with respect to the aggregation switch. And now, if any of these devices, whether it is inside of the rack or whether it is across the rack, if they have to connect, they just have one hop that is the aggregation switch.

So, we have in essence eliminated the additional top of the rack switch as a hop to facilitate this communication. And that also eliminated the need for these in-rack switches, saving a lot on the cost. But this has its own consequences, when we think of like now we are cabling, we are sending a lot of these wires, and making these connections, a lot more on the aggregation switch.

So, the number of ports that you would need for the aggregation switch, you can see here we needed just around 3 ports for connecting all the racks now you would need as many as if there are 10 of these, then you would need 30 of these ports on the aggregation switch. That is also an expensive affair in terms of setting up.

And also wiring now, like all of these devices were wired just inside, now you are moving the wiring all the way up to the other end. So, this also adds to the cost on how you would wire and what is the length you would use to connect these wires. And again, like this is again a convention when we say the end of the row being placed at the very last element in the row, but not necessarily that we have to follow this end of the row convention; this could also be the case where we would see this somewhere in between and switch the rack out.

So, that would become essentially what we call as the middle of the row. And this would help avoid the cable length so that either side of the rows can be connected to the middle of the row. And this is exactly the same as what we spoke of this kind of row model. And this is another variant that is also followed, but in very less of the cases, while most of the cases typically use the top of the racks.

(Refer Slide Time: 31:50)

	ToR Design	EoR Design
Network Deployment	Minimum 1 switch per rack	Switches centrally residing in 1-2 racks of the same row
	Each rack is a separate	Module racks work as a group
Required Devices	Switch count is higher	Less number of switches
	Less number of Cables	Higher number of Cables
Power & Cooling	Underutilization of switch	Effective utilization of switches
	High power consumption	Less power consumption
	Greater need for cooling	Lesser need of cooling
Network Expansion	Greater layer 2 data traffic	Lesser layer 2 data traffic
	Network expansion is easy	Network expansion is difficult

Data Center Networks
Advanced Computer Networks

So, if we consider in terms of how these would compare, if we think of network deployment, in the ToR design, you need one switch per rack as an additional entity while in the end of the rack

design, all the switches centrally reside in 1 to 2 racks of the same row that means you only have the aggregation switches that are being placed in one row. And you have eliminated the need for one switch per rack.

But if you look at the modularity that you get with top-of-rack design, each rack is a separate entity that can communicate amongst themselves. And if I want to isolate them from connection with any other entities, I can just take out the connections that connect the top of the rack switch to the aggregation switch, and the entire entity becomes contained in its own, and I can even move this rack around. So, it becomes very modular in terms of treating each rack as a separate entity that can be connected, updated, and done, but this end of the rack design, the module rack works as a group because now you are connecting all of these devices to the aggregation switch and if I have to take out one of the devices, I have or n devices within a rack, I have to basically cut off the links with respect to the aggregation switch. And if we consider in terms of the required devices, yes, ToR has a higher count because you need a switch per rack while EoR has a lower count. That also means that because now we have lowered the devices, the utilization of each of the links would be maximized because the connections would be direct through the server to the end of the rack.

So, effective utilization of the switches is more in the end of the rack design as opposed to the top of the rack design. But top of the rack, although it provides under-utilization it is in a way also giving ample bandwidth to change within a rack in terms of how you want to scale the resources. But for this scenario, until aggregation devices are upgraded, there is no change in the bandwidth for a given end of the row model to communicate.

And in terms of power and cooling, we would also see that there is a lot more higher power consumption because of additional devices that are being put in each of the racks in ToR design as opposed to end of the rack. And this power also means you also need greater amount of energy for cooling. But when we look at just the layer 2 communication, the best part about this top of the rack is because we are contained within a rack and communications can happen much more easily, so, the layer 2 communications that happened within a rack is always more efficient in top of the rack as opposed to end of the rack devices, although the number of hops are lower the means for layer 2 data traffic to move is much more simpler in the top of the rack design as opposed to the end of the rack.

And also the expansion, like we said, the rack has an independent entity that can be plugged in plugged out. And the only way to scale is to adding one more link on the aggregation. But now on the end of the rack design, the expansion is a lot more difficult because if you have to add one new rack in the ToR design, it is equivalent to saying adding one switch and adding one port extra on the aggregation switch. But in the end of the rack, you need to provision again for multiple of the ports on the aggregation switch. And this is where we see that end of the rack has some benefits as well as some drawbacks when it comes to the utility and setup. And most preferred, like I said, is typically this top of the rack design.