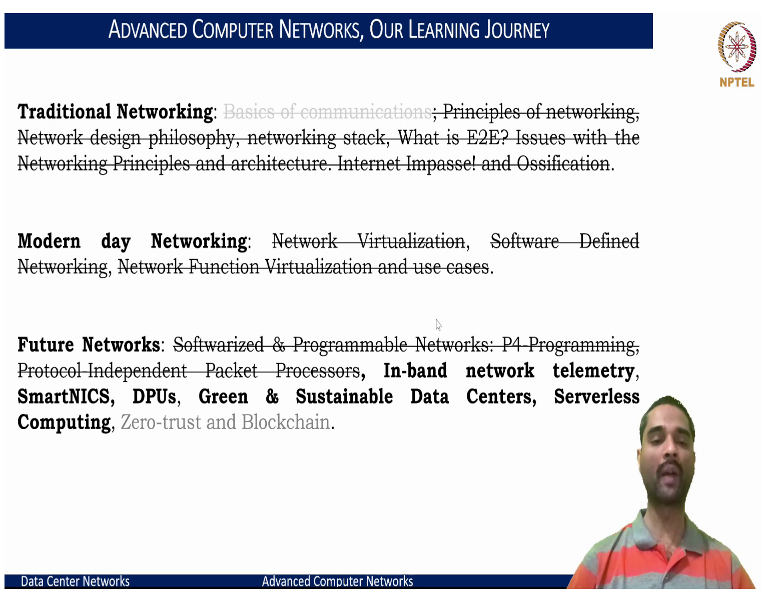
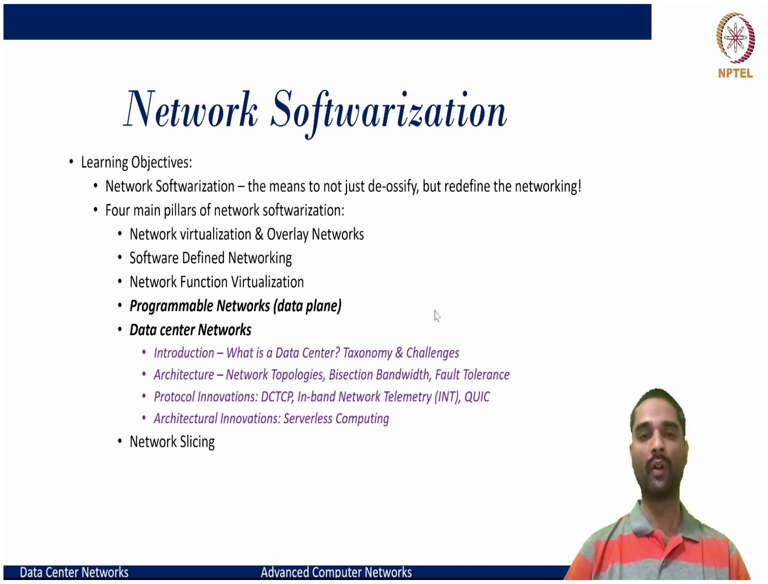
Advanced Computer Networks
Professor Sameer Kulkarni
Department of Computer Science Engineering
Indian Institute of Technology, Gandhinagar
Lecture 48
Data Center Networking - Introduction Part 1

(Refer Slide Time: 00:16)



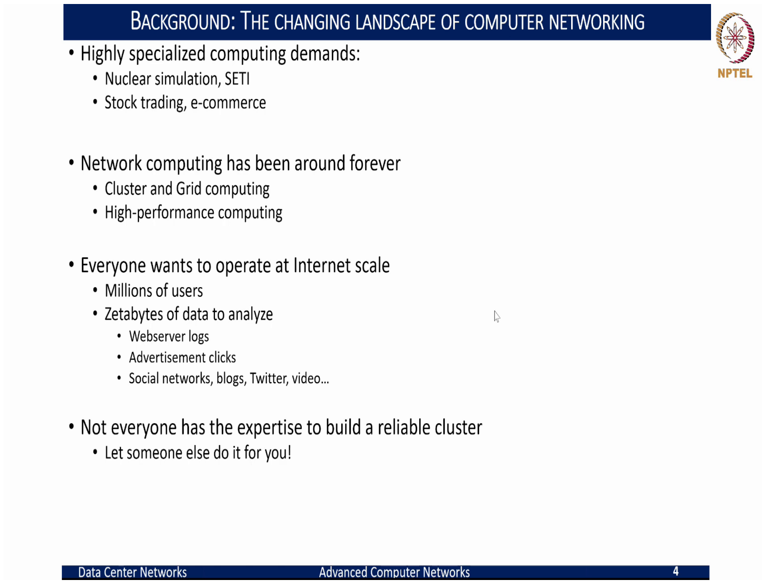
This week we will learn about data center networks. Specifically, what we will try to learn is cover the aspects of what are data centers, how they are built, and then try to revisit what we learned in the earlier weeks of learning about SDN, NFV, and programmable networks and put them in the context of data center networks, and especially, we will learn about the inband network telemetry, and the usage of smart NICS and DPUs towards building green and sustainable data centers. And we will also look into the specific architectural and protocol innovations that happened with data center networks.

(Refer Slide Time: 00:55)



Trying to look into what we call as the DCTCP as a variant of a TCP that evolved specifically for data centers to facilitate microsecond scale of operations for congestion control. In-band Network Telemetry, as a means to monitor the network of devices in a large scale in a better fashion, QUIC as a means to provide low latency HTTP services, which is now being standardized and what we see as HTTP 3.2 and on the architecture side, we will see how the clouds have emerged, and the new models of how the services over cloud are delivered. In this aspect, we will try to see the serverless computing, what it means, and what is the architectural paradigm that it brings, and the means of how the services that the user can use these cloud utilities to deploy and run the services being provided through the means of serverless computing.

(Refer Slide Time: 01:51)



But first, let us see and try to dig in, like why we really needed these data centers, what is the key background that led to the usage of these data centers, and try to understand what really led us towards this. And when I say data centers, in fact, they play a very fundamental role in our society, towards building what we call this digital economy, which is hinged on these data centers.

What that means is everything that happens online is in some way connected with data centers. And if we think of this, like whenever we are transacting online, we are always talking about computing, that is, network. So, there is a user on one hand, and there are services that are offered on the network. And both are trying to interact together to facilitate services to all the users. So, in this, if we see what really happened in the early 90s, was with the evolution of the personal computers and a lot of emerging applications, there were also need from academia as well as industry towards digitization. And if we think of like very large-scale, simulations, molecular dynamics, essentially studying them trying to see nuclear simulations, all of these led to the need for lots of computation power than what you could do on just a single computer.

And also, you might have heard about the SETI, which is a project that was led to find the extraterrestrial intelligence system trying to look at large amounts of data and utilize basically the personal computers of all the available resources in all the users in an available manner. And all of these demanded very large computation, and very large storage requirements.

And likewise, when we see from the user applications, like stock trading, and if you are trying to facilitate stock trading over the internet, you would need very low latency communication, very high processing for lots of data that you would host for different companies. And likewise, the e-commerce applications that typically facilitate shopping, banking all of these services that a user would want to do, they required specialized computing demands.

But this, in essence, was also not something new because in the early 90s, we saw what was the rise of cluster computing to facilitate and run single computation over clustered computer devices or a group of computers that would basically facilitate to provide the services and in the academic community and as the public community when you want to have the public utilities, public offices that would store their operational data and services, the means for grid computing also evolved. And on the other end, when we needed the supercomputers or to have very high-performance computing, where you would want to run services using these clusters in a purpose-built fashion, the use of HPC is also emerged.

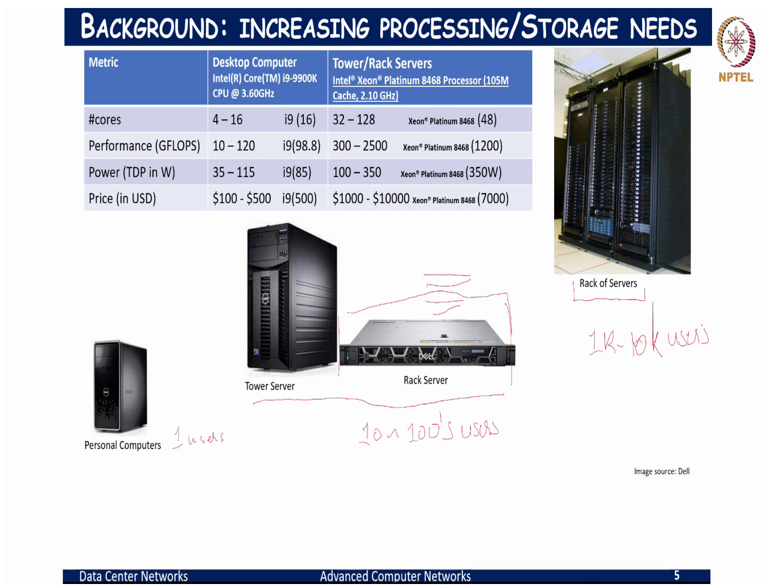
But then, when you had an enterprise or a campus that used to facilitate few users and as the scale grew and as the applications grew in space to support multiples of users, wherein the space of saying supporting internet scale that is millions to billions of users at a time, you need a lot more computation and a lot more storage as zettabytes to exabytes of data that would be collected like a web server that is running, serving millions of users who would end up providing several petabytes of logs per day.

And if you had to store them, analyze them, for that data, you definitely need a lot more computation and storage capabilities. And also, like when we see the emergence of social networks, including Facebook, Twitter, and now streaming video, all of these kinds of applications, which are being used by millions of users, at a time demanded the need for very high, very large amount of processing and storage capabilities.

And not everyone has the same expertise to build such a reliable cluster. And that is where the model to say how one could really leverage the expertise of the people who would build a reliable cluster and make them available to many also became a common practice. And also, we spoke earlier about the capital expenses. Think of trying to buy a PC affordable, trying to buy hundreds of servers, maybe for an enterprise may be affordable, but if you want to scale immediately to the internet scale, you want to have thousands of such devices that you will want to operate with. So, investing on them may not be feasible.

So, hence, why not utilize in a common shared pool of resources that where you can leverage onto capital expenses being put by someone else. And all of these economic aspects, as well as the changing needs of the computing aspects, led towards the need for common frameworks or platforms where you could have lots of compute and storage capabilities that anyone could use.

(Refer Slide Time: 07:11)



And in this sense, if we see how these increasing processing and storage needs and requirements are met, we start with on one end, the personal computers, which typically we as either the desktops or laptops that we use, in the early 90s, they were all single processor mode. But now what we see in the market is all multi-core devices, like at best, they may support around 16 cores.

If I see that from the most recent Intel 9 processor series, you get around 16 cores on your personal computer. And when we say of these cores, around 4 to 16 of them, each core has a peak operation or a maximum performance that it can deliver in terms of how many operations it can support. And if we measure them in terms of the floating-point operations, and quantify the performance, what we see is they may be able to support around ten to hundreds of gigaflops, that is 10^9 floating-point operations per second. And such of those in the number of ten to hundreds.

And for them to run with a thermal design power ratings. And around at the minimum, we have around 35 watts to now what we see on the orders of around 100 watts. So, if we can see that in the perspective of i9, it requires 85 watts of TDP. And if we see from the price perspective, they are relatively cheaper, and provide around 100 to 500 US dollars per processor.

And if we see the i9, most latest of the series from Intel, we end up, say, around 500 US dollars. And these were good to support. Say one to a few users. But as we said that, as an enterprise as a campus, we want to support more users, and we definitely need a lot more than this. And that is where the need for what we call as the server machines or the kind of hardware which would essentially support a lot more performance and be able to support a lot more I/O operations was required.

And that is where the Tower Servers or the Rack Servers that come into different variants in terms of how they are housed, the Tower Servers look exactly identical to what you would see as your personal desktops in terms of the make and the chassis that you would build. While the rack servers have different standards in terms of how their chassis are built you can see them in the picture here.

And what it really enables is to have a lot more computational support in terms of if you see the number of CPU cores, they may have in the orders of 32 to around 128 CPU cores that could run on these. So, what that also equally corresponds is the need in terms of having to run them; you need a lot more power, and their thermal design power rating is somewhere up around 100 to 350 bytes.

Essentially, if you have a lot more cores, you are able to deliver a lot more performance. And the performance numbers, again, range somewhere around 300 to around 2500 gigaflops ratings. And, as you can see, it is already more of a 10x 200x that you can get with these server machines. And likewise, when we see the price, they are a lot more expensive. And like if you see with respect to the most latest Intel Xeon platinum rack servers, you are able to see around 7000 US dollars, but like if you see very high-end ones, you may range around 1000 to 10,000

US dollars. So, likewise, these kinds of servers, which provide an expandable I/O and are able to add a lot more computation, bring in a lot more computation and facilitate to add a lot more memory.

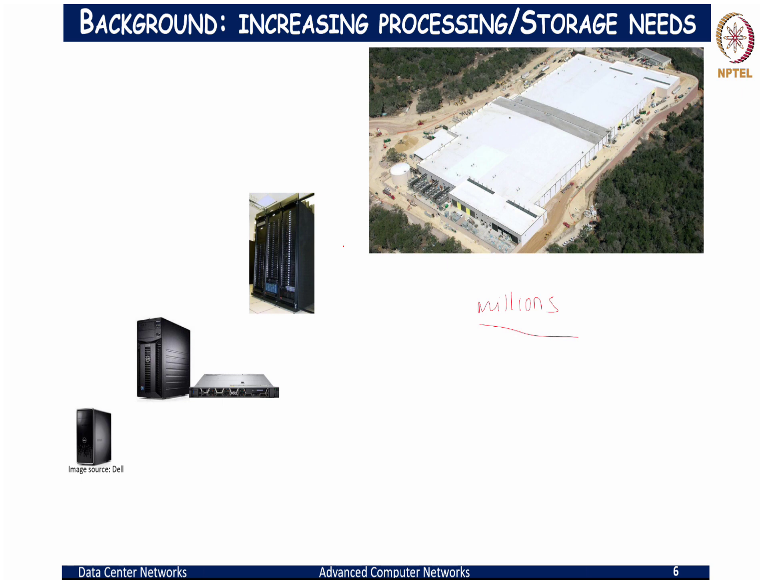
And post, a lot of storage started to emerge. And these can support around tens to hundreds of users very easily. And now if we are speaking of a company or a campus that supports thousands of users, what we would need is to scale these numbers of servers accordingly.

And that is where the standards of how to build a rack server what the unit started to end so that you are able to instead of building taking a lot of space for Tower servers, you could make them use very thin rack servers or servers which take less of a thing, and you can stack them one over the other. And that requirements and standards grew in terms of how you can build a rack of such servers.

So, we are speaking of the servers that are stacked one over the above and placed in one rack or one container. And this way, you are able to scale the computation and storage requirements that you will want to serve and be able to place them more efficiently as opposed to bringing these tower servers. And that is where these rack servers, with around hundreds of such rack servers, would now be able to facilitate around 1k to 10k users.

And this is how the scalability, you can start to see that can help in terms of meeting the needs of the processing and storage demands.

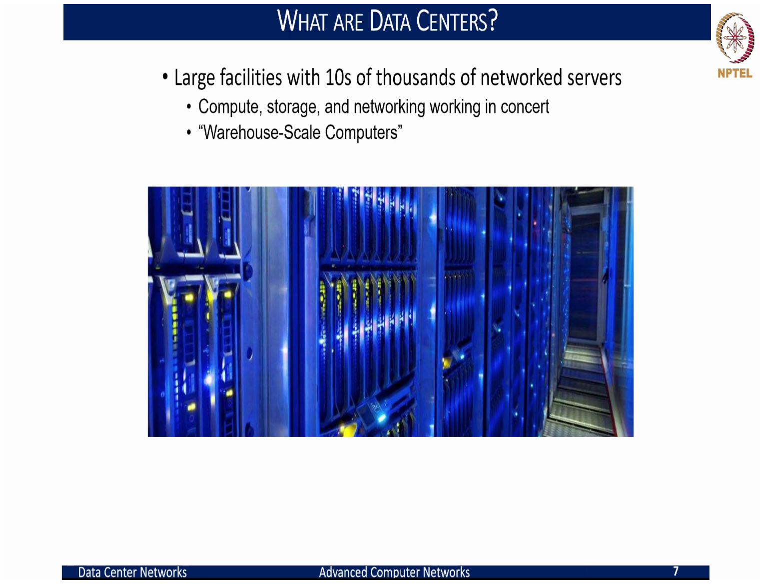
(Refer Slide Time: 12:42)



And when we speak today of the internet scale, we are speaking roughly around supporting millions of users. And in order to support such millions of users, we would need a lot more of these rack servers. And what we see in this picture on the right here is a big building that is hosting just these compute servers.

Or think of these as the godowns or the warehouses where you basically store large amounts of these computers that are all networked and able to deliver basically, run continuously to deliver the services over the internet. And this is exactly what we see as data centers. So, in essence, data centers are the supercomputers on which our day-to-day activities run.

(Refer Slide Time: 13:37)



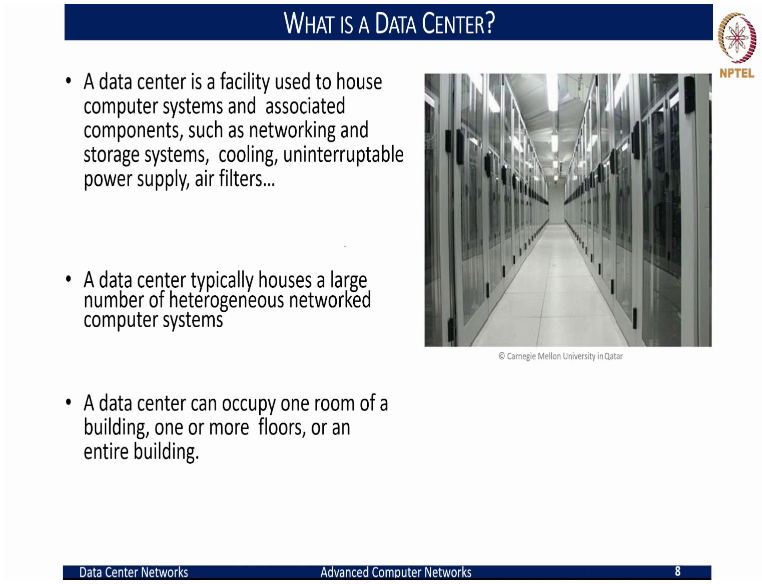
So, if we think of these data centers, they are in essense facilitating to host hundreds of thousands of these network servers, stacked in specific order run infrastructure being built to cater to this in terms of how they are powered, how they are managed. All of this has to be done in a specific order. But if we think of it on a high level, these data centers, they provide you with the compute and storage.

And if they have to provide to a large-scale to compute and storage requirements, they all have to be networked. And when they are internetworked to work together, we can see it is as to warehouse-scale computers serving for several users at the internet scale. So, what these data centers are essentially, we see as the integral parts of the enterprise designed to support several of the business applications.

And if we see these data center networks, you would have to invest a lot. Like if I have to build a home or construct and pay for the space first. Even without the computers and then you have to host the computers in them in thousands of much numbers and each we said costs you somewhere around 1000 dollars. So, that automatically means that you will need millions of dollars of investment to even create such large-scale data centers.

And if this requires such an infrastructure, also the maintenance operations make it even more complex. So, all of these aspects are to be addressed when we think of data centers.

(Refer Slide Time: 15:25)



So, to define formally, what a data center would mean, is that think of a warehouse that is able to facilitate and run lots of computer systems or components that are networked. And these require essentially an uninterrupted power supply. And you think of your computers, they generate a lot of heat. And the server machines again, will generate a lot of heat, and they need to be cooled.

So, you also need infrastructure to cool these devices enabling the servers are maintaining a clean space. So, it encompasses basically a large facility, where all of these operations with respect to how to manage these computers of this scale becomes essential. So, it is not necessarily that all the devices have the same kind; we may have a variety of networked computer systems, meaning I may have some of the servers that are windows, some of them may be Linux, some of them may be Mac.

And in fact, not just for the software, the entire hardware make would completely differ. So, I may have variety of vendors who may be providing these servers, maybe Dell, maybe HP, any other servers, even the custom built servers that could be put up. And these data centers occupy a large space like you need them to manage an entire room or a building for building with lots of floors that you would have to support.

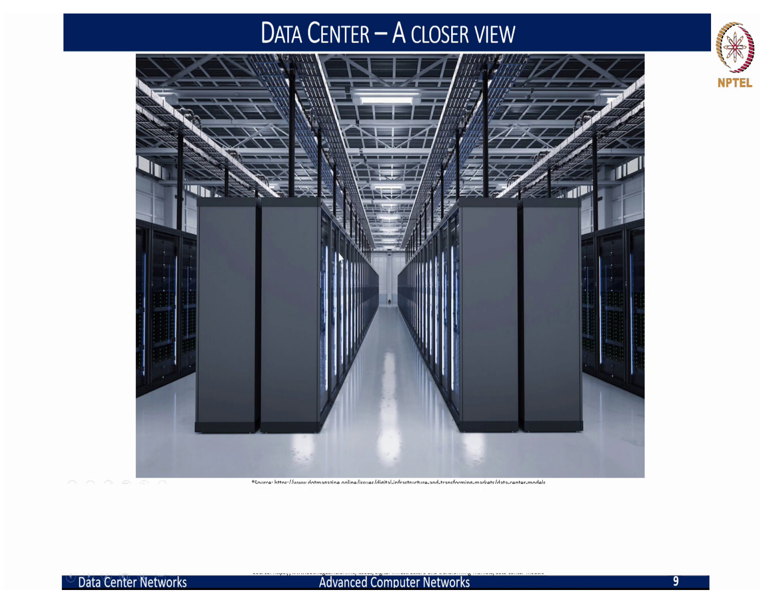
So the power management, the ventilation management, the air cooling, all of these aspects have to be addressed when we think of these data centers. And these data centers are essentially an integral part of any enterprise that are designed to support several of the business applications.

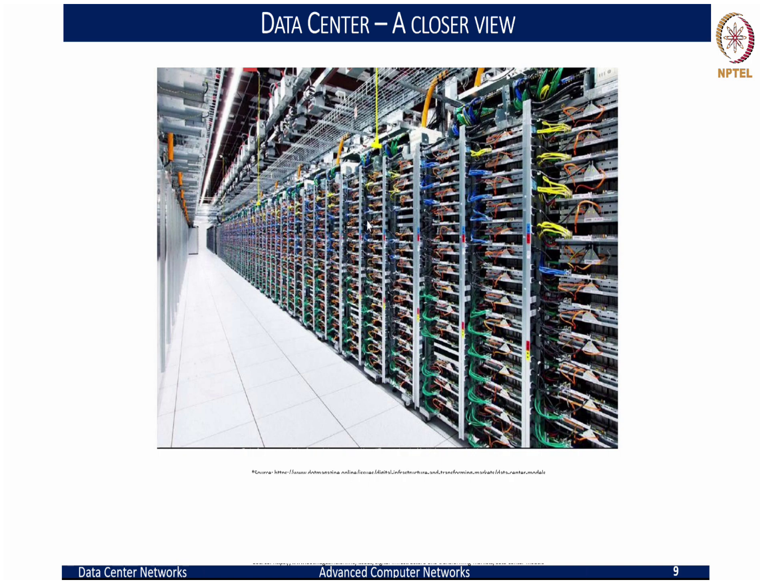
And if you think of like what kind of business applications, then you may think of data storage, like what we do with Google Drive, or there are specific storage services that you may think of where you want to keep lots of data, like OneDrive, Google Drive, and many other services.

And, in fact, when you have this data storage, how do you replicate them? How do you manage them, what is the means to recover them, all of these go behind the scenes with respect to all the services of data storage are provided. They may be used to facilitate several of the productivity applications, for example, emails that we use in our day-to-day life. And when Gmail servers are serving so many of the users, they have to host lots of data and provide the services on a large scale.

Next, if you think of the e-commerce transactions that we do, for Amazon or eBay, all of these which support a very high volume of transactions, you would need somewhere to be provided facilitated through these data centers. And even the online gaming that we can see where millions of users are hooked. And if not, on the scale of billions, at least thousands of users are playing together on a given server. And this means that you need to support the online gaming communities that have to scale and also have to operate with lots of computers; you need these data centers.

(Refer Slide Time: 19:31)





And today as we speak, there are more than around 7 to 10 million data centers worldwide. And practically every business government entity, even where you have all of the digital services that are being provided, like think of Aadhaar. If we have Aadhaar database, it requires a large data center and replications of the Aadhaar information so that you have reliable service to be provided requires these data centers and all of these digital transformations that we are building, think of any server that you would want to host which supports multiple users as a part of such data centers.

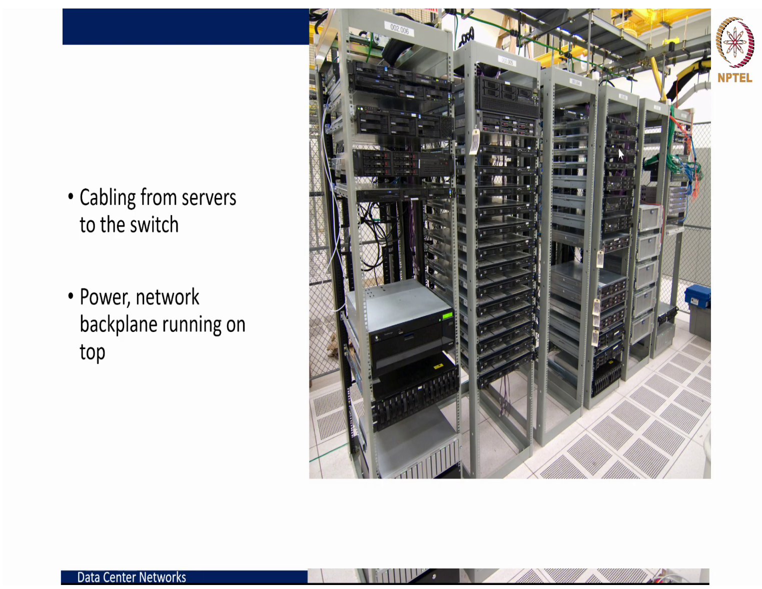
And if we now look at the data centers in a more closed view on like, what we saw as a building, that building would consist of what we can see here in this diagram as a scale of a lot of these rack servers that are being placed together. And what we are seeing here is a bunch of the racks that are hosting those rack servers being put together and powered together and managed within the single building.

And if we peel in into this and see what goes in into them, you can see like, these are all the rack servers, which are now being interconnected using these ethernet, the copper cables, making sure that you are able to connect to each of these servers. And once you see that there are so many of these devices, so many of these wires management, you can think is a painstaking job in terms of even trying to locate one of the servers within such hundreds of thousands of servers becomes a major concern.

And if we see much more closer, all of these devices that we are hosting together, they are network that means they have to have their own addresses. So, how do we manage such address space, especially on networks? Like how many IPs can you provide? How many of these private IPs, how do I pinpoint these devices to be what is their layer 2 address what is their layer 3 address? All of these become a major challenge.

So, that means there is a need for a scientific approach to how we build these mechanisms. Also, when we have all of these servers that are being placed together, they will generate a lot of heat. And they will also need a lot of power to run. So, the means to supply power to each of these racks, each of the servers within the rack continuous or uninterrupted power supply to each of these what are the models because now you may need several kilowatts, to megawatts of power to run just these data centers.

(Refer Slide Time: 21:59)

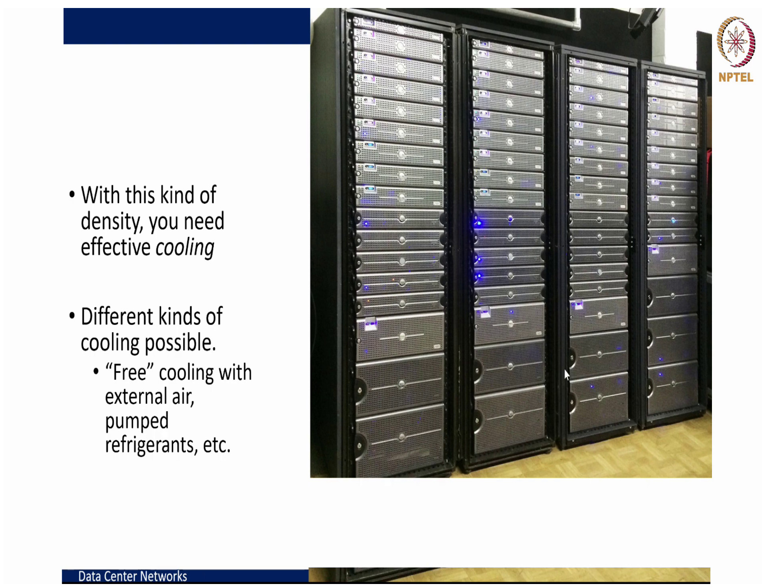


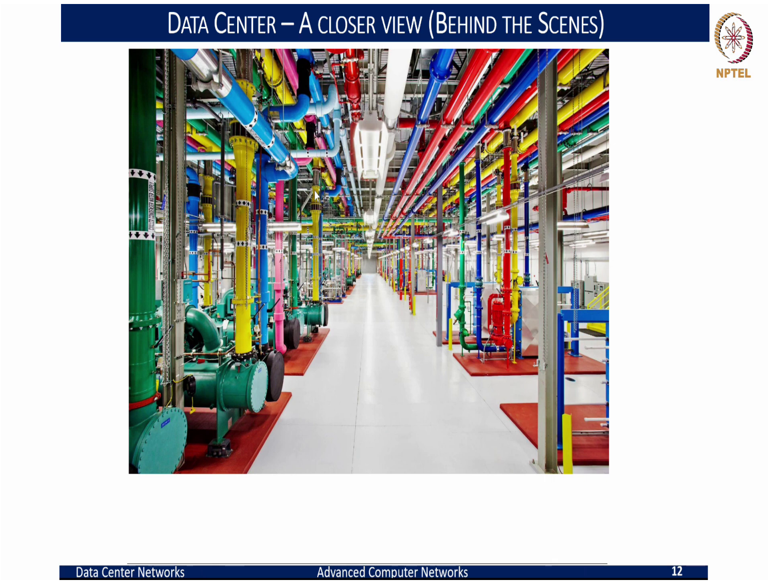
So, to give a very clear picture here, what we are showing here is how typical campus networks would look with a lot of racks that are being stacked. So, you can see that each rack somewhat comes in a standard size. And it hosts several of these servers, and some servers, you may find them to be somewhat thicker; some are very thin servers that you can see. But this is how typically you mount a server onto the racks and run them.

And what this requires is that these servers need to be connected. So, if you have to connect, you may need cabling from the servers to the switch. And what that means is each of these servers needs to be connected to one of these switches where to which they can all be interconnected. So, you need a network backplane.

And also, each of these servers may be powered differently. And with the power source may come from different entities altogether. So, you also need a backplane for how to power these servers within these racks in a way. So, all of these require us to say, we need some way to make these operationalizing aspects of data centers to be realized.

(Refer Slide Time: 23:19)



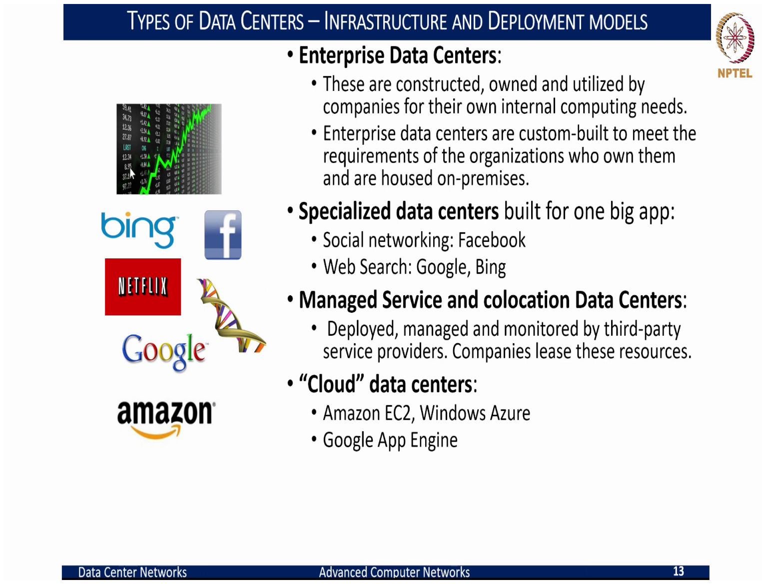


And if we see the same in when we completely enclosed manner, this entire rack is now filled with different kinds of servers. And all of these are basically facilitating a common pool. And this is typically just like our HPC on a campus network would look like. And this would also mean that they are going to generate a lot of heat. And we have to have the need for effective cooling mechanisms to dissipate the heat out and ensure that these devices can be cooled to operate reliably.

And hence, the need for the infrastructure also to cool these devices becomes important. And what you often see is what we just saw earlier in terms of the racks and series of racks but what goes behind this is basically the kind of cooling system, and this, in fact, is a lot bigger space requirement in terms of hosting these cooling systems for the rack of servers.

So, what I am showing here is one of the cooling systems that is being built for the data centers, like Microsoft data center that is that we saw earlier in the diagram, requires such a cooling space to ensure that you are able to recirculate the cool air where those data center racks are being hosted. And often now, like what we see is the amount of space required for such cooling systems has been increasing.

(Refer Slide Time: 24:50)



And if we have to look now into the types of data centers that we have in terms of the applications the needs and what kinds of data centers we see, we can think of most enterprises have their own local data centers, which are basically constructed, owned and utilized by the companies for their own computational needs.

And think of these enterprise data centers as custom-built to meet the specific requirements of the organizations who own them. Like if I am speaking of academic institutions, we may have entered the custom on-campus data centers, that would typically host either the ERP services where the student records hosting of student records, and other data library services that you will see the Digital Library, all of these being run as a part of these enterprise data centers.

And if you think of a very small companies, or business companies, they may have their own ERP systems that they want to run the employee database, all of that needs to be hosted, they will typically be using these enterprise data centers. And then there are also specialized data centers that are built for specific applications, like when we use the Facebook on social networks, or when we do Bing, or Google search, or YouTube for videos that we stream and look up. So, all of these require specific kinds of devices, specific kinds of computation, and storage requirements.

And that means the companies could also build specialized data centers. Now, if you think of Facebook, Facebook as an enterprise would have its own enterprise data center. But Facebook has a service that is being provided to the internet users and would have a separate dedicated data centers for running their applications.

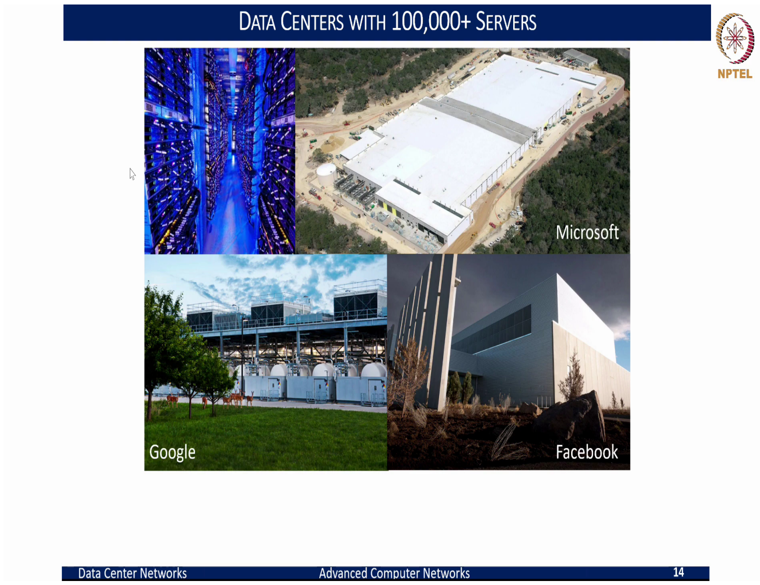
Likewise, with Amazon, Google and Bing. So, these aspects are where these specialized data centers are being built like when you think of Netflix, Hulu, any of the Amazon Prime, any of these services, or online gaming, all of these require specialized data centers. And the alternative model of this like when we look at the typical government infrastructures that we build and utilize, you would see a lot of managed services that are being run by the Central or the State government or even in a public-private model, what we typically use are managed service or colocation data centers are also termed the colos, where somebody hosts and manages the data centers, while the third parties would basically be involved in serving these data centers, and provide services to various users.

And typically, if we think of a very small startup companies, you will typically not have enough to build your own data centers. So, we would lease out the resources from these third-party service providers to facilitate that we can host our requirements on these managed services. So, someone else takes care of the management, someone else takes care of the infrastructure but provides the services for other people to use and share them, and hence, the colocation where multiple of two users are able to utilize the services on the same premise.

So, think of like what the Amazon builds it on a very large scale, the Amazon AWS facilities. So, these are the cloud data centers, which are on a very large scale. And a miniature that would be run would be the colocation services. So, also, you have typically these data centers that are being built now as platforms in the cloud like you think of Amazon storage, Amazon elastic compute, or the Microsoft, The Azure platform, or the Google App Engine as a platform service, or a Google Drive that you will use, all of these are typically what we see as the large end of these data centers that may be spread geographically at different locations. And this is where the cloud scale, what we termed as the cloud-scale data centers, are also being used.

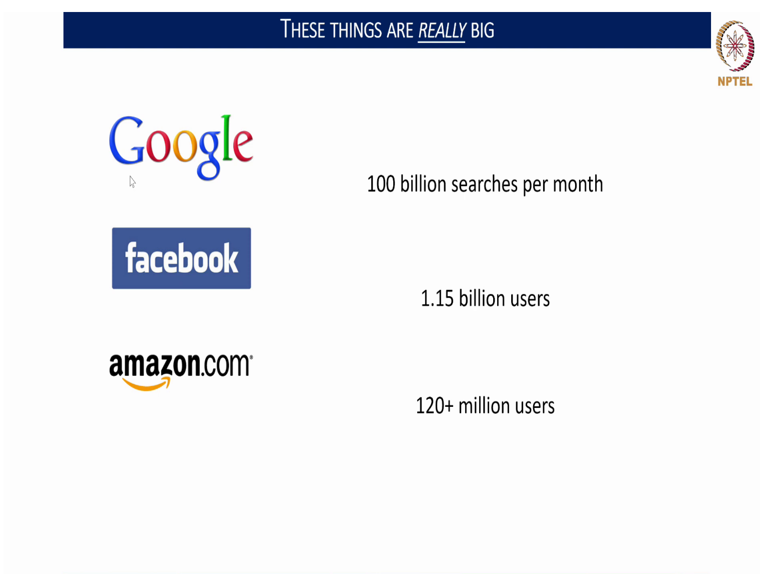
And in essence, what we have tried to see here is there are different kinds of data centers based and each has its own needs that it caters to. And also, in terms of the infrastructure, the requirements will be completely different. And the way that you would deploy the servers in them and run them would also vary. So, either I could have a privately-owned system, or it is owned by someone else, but rented or leased for others to use, or they could even be the public-owned data centers wherein the government-headed organizations can run their services.

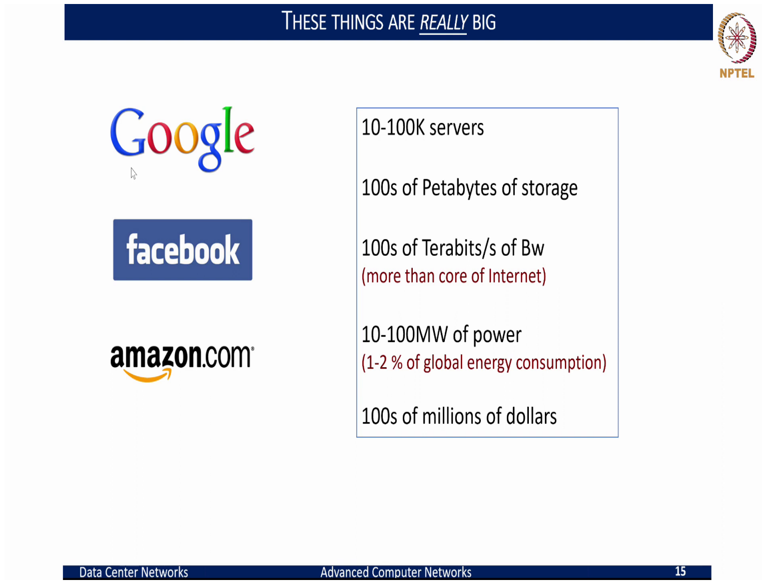
(Refer Slide Time: 29:44)



So, to put this in perspective, like we may have the data centers that are being built for Google, Facebook, Microsoft, all of them.

(Refer Slide Time: 29:56)



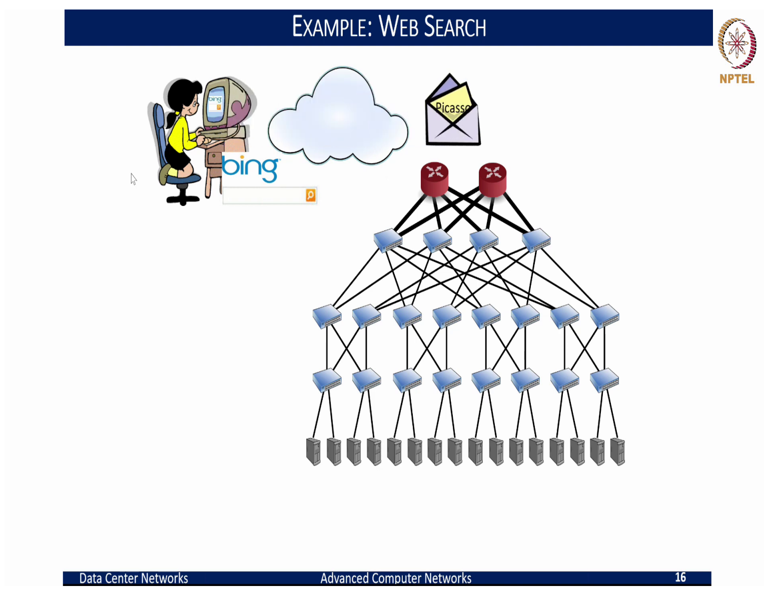


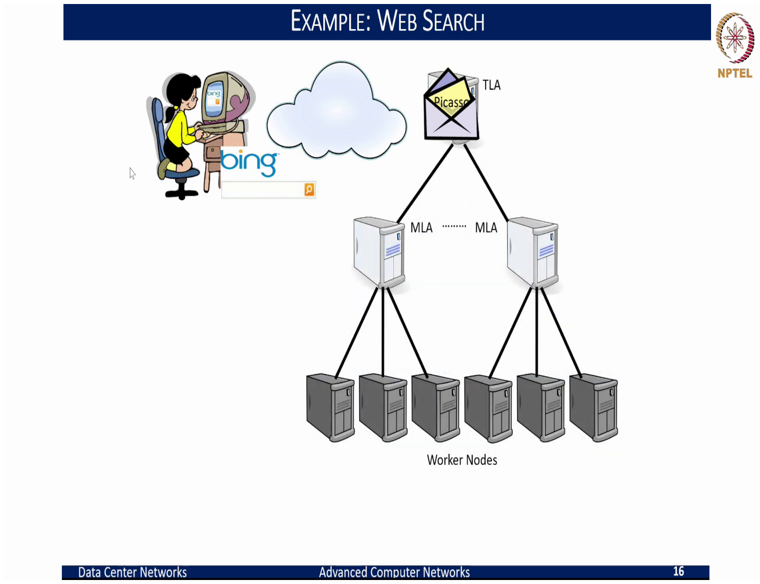
So, if we think of these as really on a scale, you can see that on Google, you will have, in fact, 100 billion of searches and much more of these in a given month. And on Facebook, as we speak today, we have roughly around 3 billion users that are connected. And also, like when we see Amazon being used, there are more than 150 million users that are being plugged in and running these services.

So, in essence, what we have seen is that these data centers are used to facilitate hundreds to thousands of servers that are going to be run, and each having petabytes to exabytes of storage facilities. And in terms of networking, they are able to facilitate hundreds to thousands of terabytes of bandwidth. In terms of bandwidth, now, if you see, you can compare to what you would have learned in the earlier internet aspect, the core networks are a few terabytes at max 2 to 4. But now we are speaking of a much more massive scale, and being able to support hundreds of terabytes of bandwidth for these operations to take place. And again, when we speak of the power requirements, again, it ranges in the order of hundreds of megawatts of power to run these infrastructures.

And in fact, data centers have been the major share in the global power consumption. And it is been much more now as of today with respect to the percentage of share that they have over the power. And it is been increasing rapidly as well. And the investment also is likewise hundreds to billions of dollars that are being spent for these data center infrastructure.

(Refer Slide Time: 31:40)



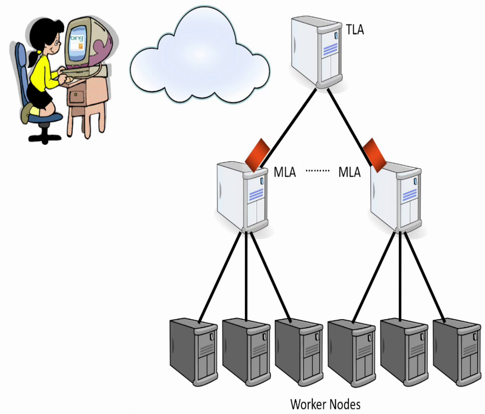


So, if we see that and try to put a perspective from a user point how these data centers would really operate, and see if we had, because as a user, I am trying to do some search on a Bing search engine. What that would really translate is, as an end user, when you type something on the Bing search, and the request would then go to one of the data centers where the Bing search engines are being run. And this typically needs to provide you with lots of resources and information back.

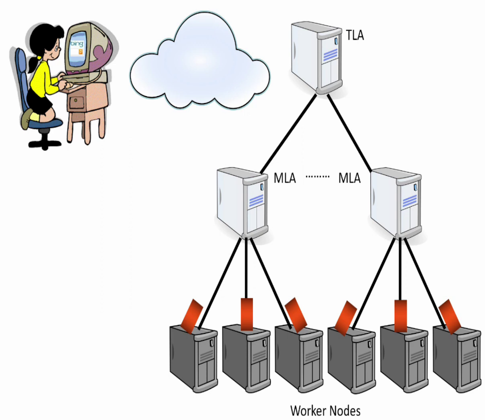
And how this was managed is what we just saw as a set of computers and interconnections that would allow your requests from millions of users to be processed. And to do that, typically, you would have an infrastructure somewhere within a data center. And as your request comes to a data center, you would first aggregate the kind of requests that you would have received. And in this case, a user made a request for a Picasso as a search, that is being aggregated at the top level, what we call as TLAs or the top-level aggregators.

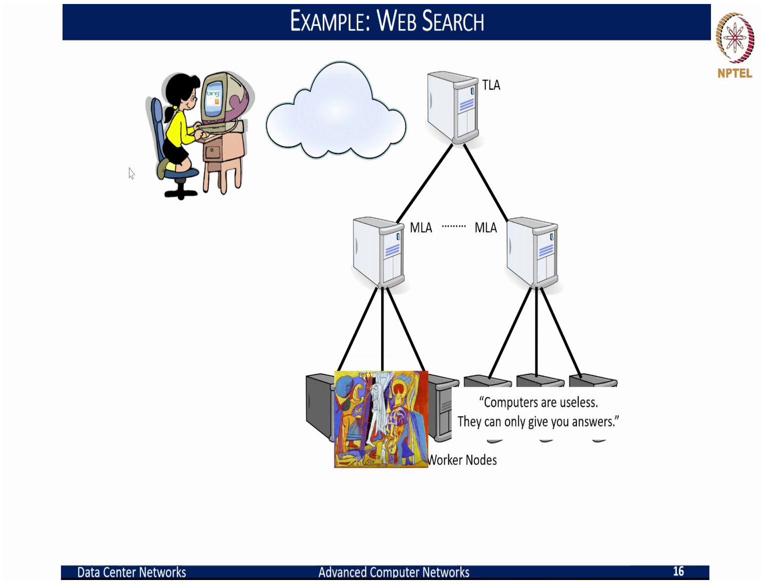
(Refer Slide Time: 32:51)

EXAMPLE: WEB SEARCH



EXAMPLE: WEB SEARCH



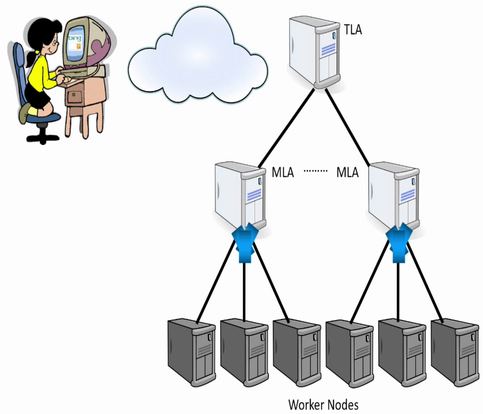


And then this request could be forwarded to the next level, that is the mid or medium-level aggregators. And basically, you are now scaling out the request so that it can be searched. And you can fetch a lot of the records back and display them back to the user. And these mid-level aggregators would, again, make the request reach out to a lot of the server nodes or the worker nodes, which would eventually gather the real data relevant to the search.

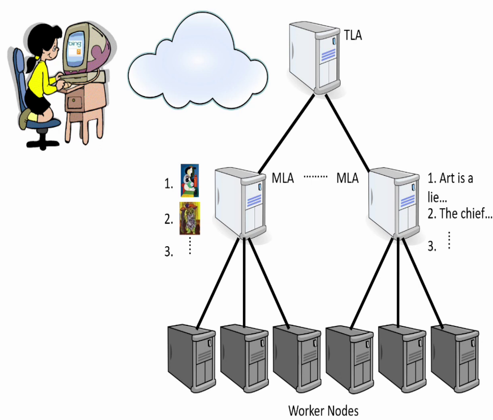
And these different server nodes would come up with different kinds of results. And now, each of these results that you would have taken from all of these different machines where they could be different storages that could be aligned based on the key-value pair. So, you may be doing lots of different stories that are sharded and also the computation in terms of what data that you will want to collect for each of them.

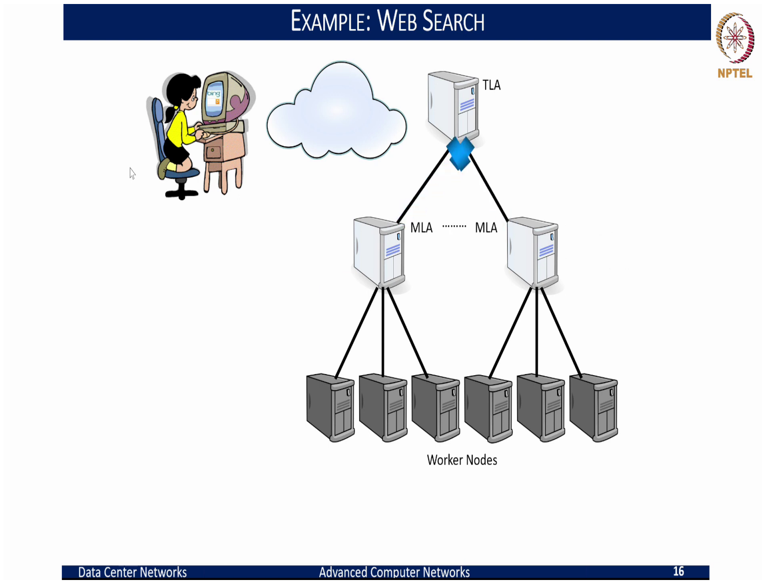
(Refer Slide Time: 33:46)

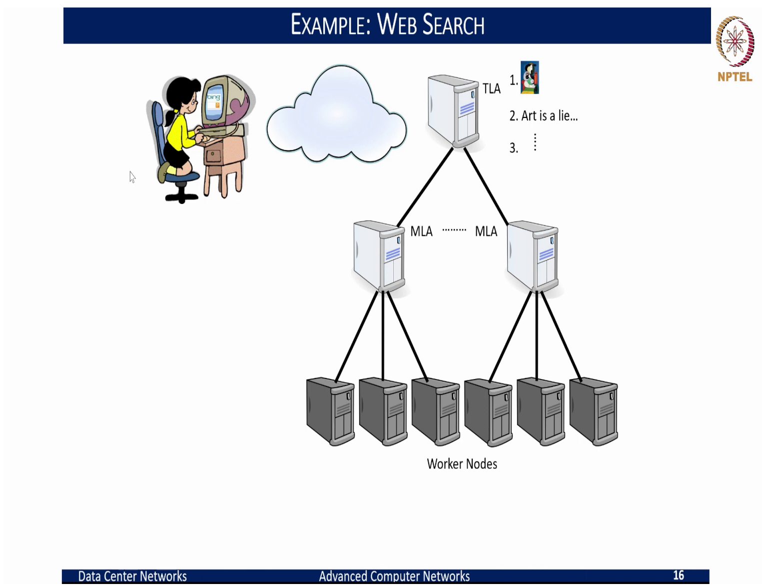
Example: Web Search



Example: Web Search

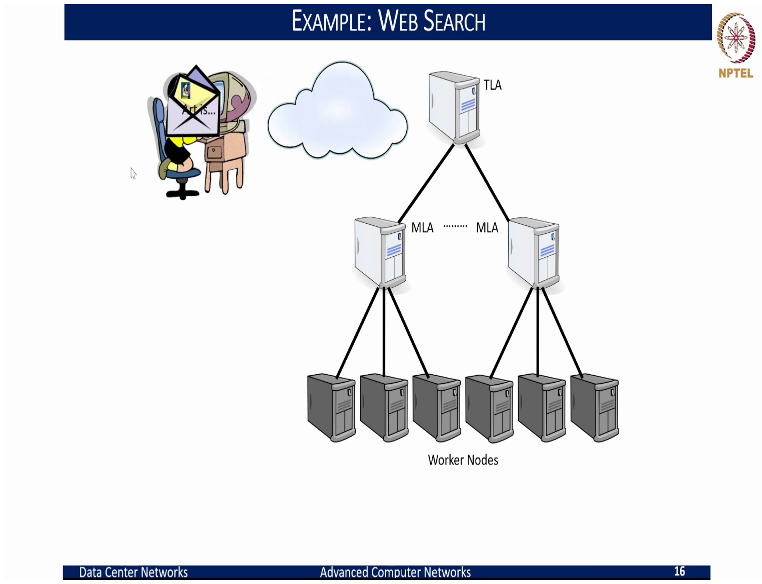


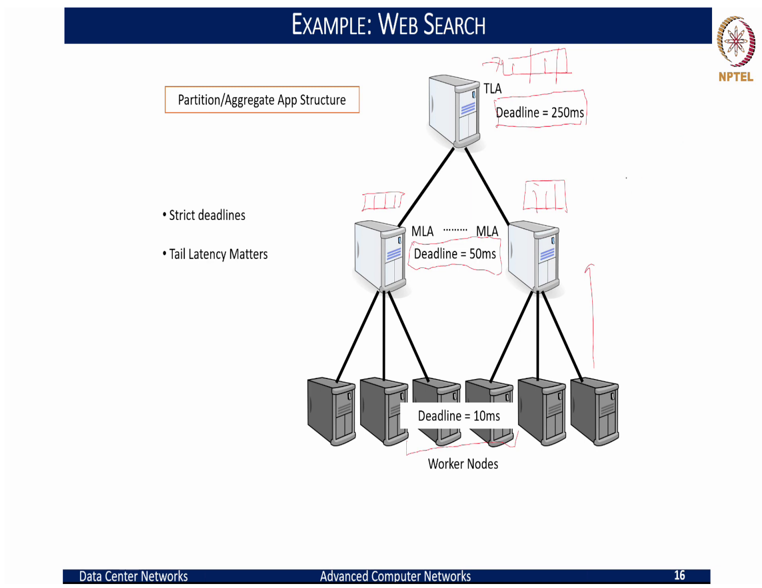




And then push this information back to the mid-level aggregators. And again, at this stage, they would align all the results responses that they would have gotten from several of these worker nodes, and then push them back to the top-level aggregator. And this top-level aggregator would now encompass all of the results that it obtained from the mid-level aggregators, and essentially downstream from all of the worker nodes.

(Refer Slide Time: 34:13)





And then, pass it back to the user's web browser so that now he is able to aggregate all of this information and view it in one way. So, what this web search kind of model shows us is that as we send one request, it is partitioned and run by multiple of these worker nodes, a single request could be run by hundreds of worker nodes in a data center. And all of these results are then again coalesced and then aggregated before providing it back to the user.

And all of this has to happen in a given timeline, that means there has to be some sense of deadline that within which these operations have to complete. So, if we consider a typical search, you would not want to wait for more than a second. You expect the response time to be within half a second the best so that you get a good user experience.

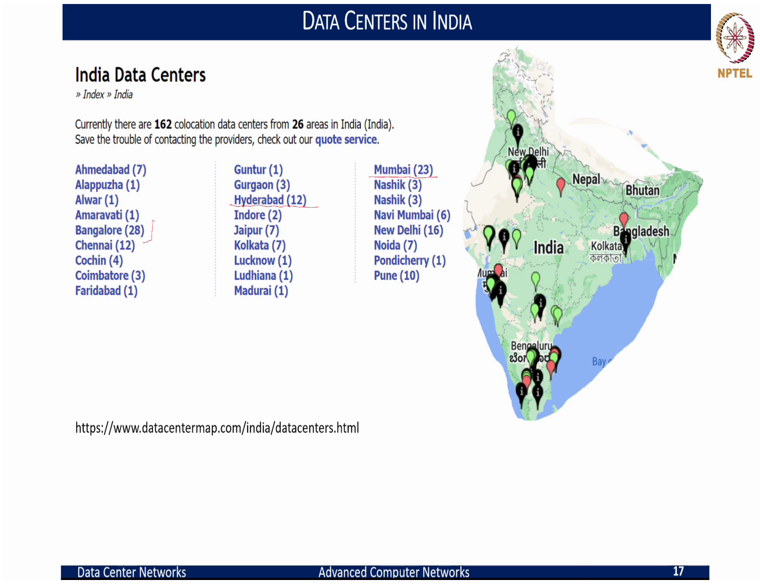
And if we see that what this really means is that you may have different deadlines for different levels, at the very least lower end, all of these worker nodes, we may have a deadline where they respond back within tens of milliseconds, tend to be around 50 milliseconds so that you are able to respond back to the mid-level aggregators in a given time.

Likewise, the deadlines for these mid-level aggregators could be around 50 to 100 milliseconds so that you are able to collect the information from many of the worker nodes and give the response back to the top-level aggregator, which will have its own deadline times, maybe around 250 milliseconds to half a millisecond, so that you are able to collect the information and send it back.

But in all of this, when there are multiple of the requests that are going to come, requests are also going to be queued at these top-level aggregators. And likewise, at the mid-level aggregators, and so on. This would mean that as the requests increase the latency, those requests will also increase. And hence, what also matters is how it affects two requests, the most recent coming requests, and what is the latency that they are going to be served with.

And this is what we refer to as the tail latency, or the latency at the peak. Like if I have 100 requests, what is the time within which the 99 percentile of the requests much what is the time it takes to serve the last 1% of the requests that are being queued.

(Refer Slide Time: 36:44)

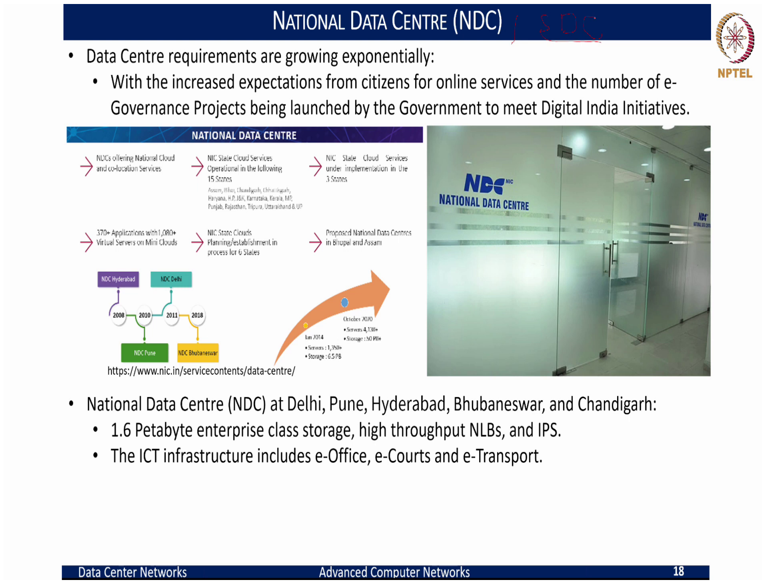


And these aspects really make us to think how to design data centers, and if I put into perspective for data centers that we see in India, and this is the information available at the data center map, which shows what are the primary locations in which data centers are hosted in India. It shows around 162 Colocation Data Centers being set up at around 26 areas in India. And these numbers, although seem very small, the numbers have been increasing over the years very rapidly.

And you might have heard many of the new Amazon and the Google data centers are being planned and being put. And these are basically the Colocation Data Centers that we have speaking, where several of the large-scale enterprises will then host and build these. But there are also other private enterprise data centers, which are not part of these numbers.

So, if you see Mumbai, Hyderabad, and Bangalore seem to be hosting the maximum of these Colocation Data Centers. And also, like we said, these data centers require good internet connectivity. And that is where also the connectivity aspects play a role in why you would see these data centers being set up more at a typical place like Mumbai, or Hyderabad, or Bangalore, which hosts the majority.

(Refer Slide Time: 38:13)



And from the initiatives from the Indian government, we will also see what we call as the National Data Center. So, what we call as the NDCs. And these data center requirements are, in fact, growing exponentially wherein, like these national data centers that host most of the government and Digital India initiative services that are being launched by government projects, where we have the expectations from many of the online citizens, citizens to use these online services like Aadhaar being one of the very simple use case that we see.

And also like these e-Offices, like e-Courts, all of these services that are being built, e-transport, any of the government services that you see in digitalized services, they need the data center infrastructure to be run. And that is where these national data centers started to build.

And in 2008, the first national data center at Hyderabad was built, and thereon, several of these National Data Centers are being built and the recent one being in Chandigarh, so you can see it NDCs built at Delhi, Pune, Hyderabad, Bhubaneshwar, and Chandigarh, so all of these basically facilitate several petabytes of storage support and also have the high-end infrastructure, including the load balancers and the intrusion prevention and detection systems, all of these infrastructure support have also been built around these data centers.

And what you also see is this initiative also led to NDC as well as State Data Centers for ensuring that you have National Data Centers as well as State Data Centers to facilitate all of the e-Services that are being provided by the State Government, as well as the Central Government.

So, these data centers are also on the rise and a lot of SDCs have been added over the last few years.

And if you see like in the early 2014, there were around 1300 storage servers with the storage capability of 6.5 petabytes. But as of October 2020, there were around 4000 odd servers supporting more than 50 petabytes of storage, and these numbers have been increasing as well. So, we can see that the need for data centers is for one and all, like you have the enterprises that need their own data centers, any public service that you would host requires the data centers and even the organizations require their own data center like academic institutions, or any other kinds of institutions. And all the government agencies or government infrastructures would also need these data centers. So, data centers have been pivotal in providing these infrastructures over the internet and ensuring that we run a digital economy in a smooth fashion.