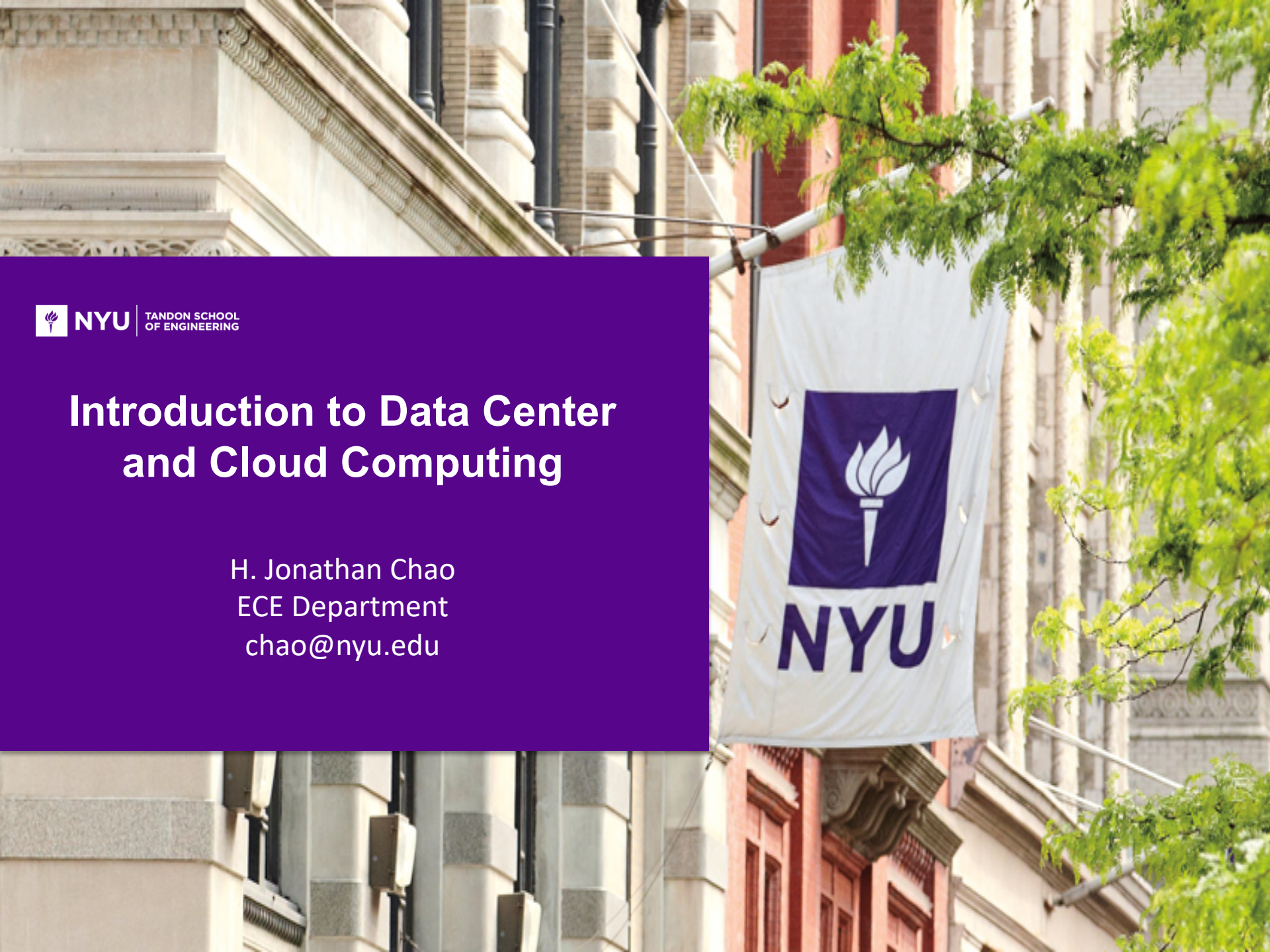


NYU TANDON SCHOOL OF ENGINEERING

Introduction to Data Center and Cloud Computing

H. Jonathan Chao
ECE Department
chao@nyu.edu

What will you learn from this course?

1. What is a data center and cloud computing?
2. How are 100,000 servers in a data center interconnected?
3. What are the interconnection network architectures?
4. How is a complicated job, e.g., data mining, executed simultaneously by hundreds of servers? With many jobs ruing at the same time in a data center, how is the computing resource of 100,000 servers shared efficiently and effectively?
5. How are the traffic flows of the many jobs scheduled to minimize the overall completion time?
6. How to isolate the operation from different enterprises that share the resource in a data center?
7. How to avoid network congestion to complete the flows in time?
8. How does the emerging edge computing play a role for low-latency services, such as, AR/VR, CAVs, haptic surgery, ….?

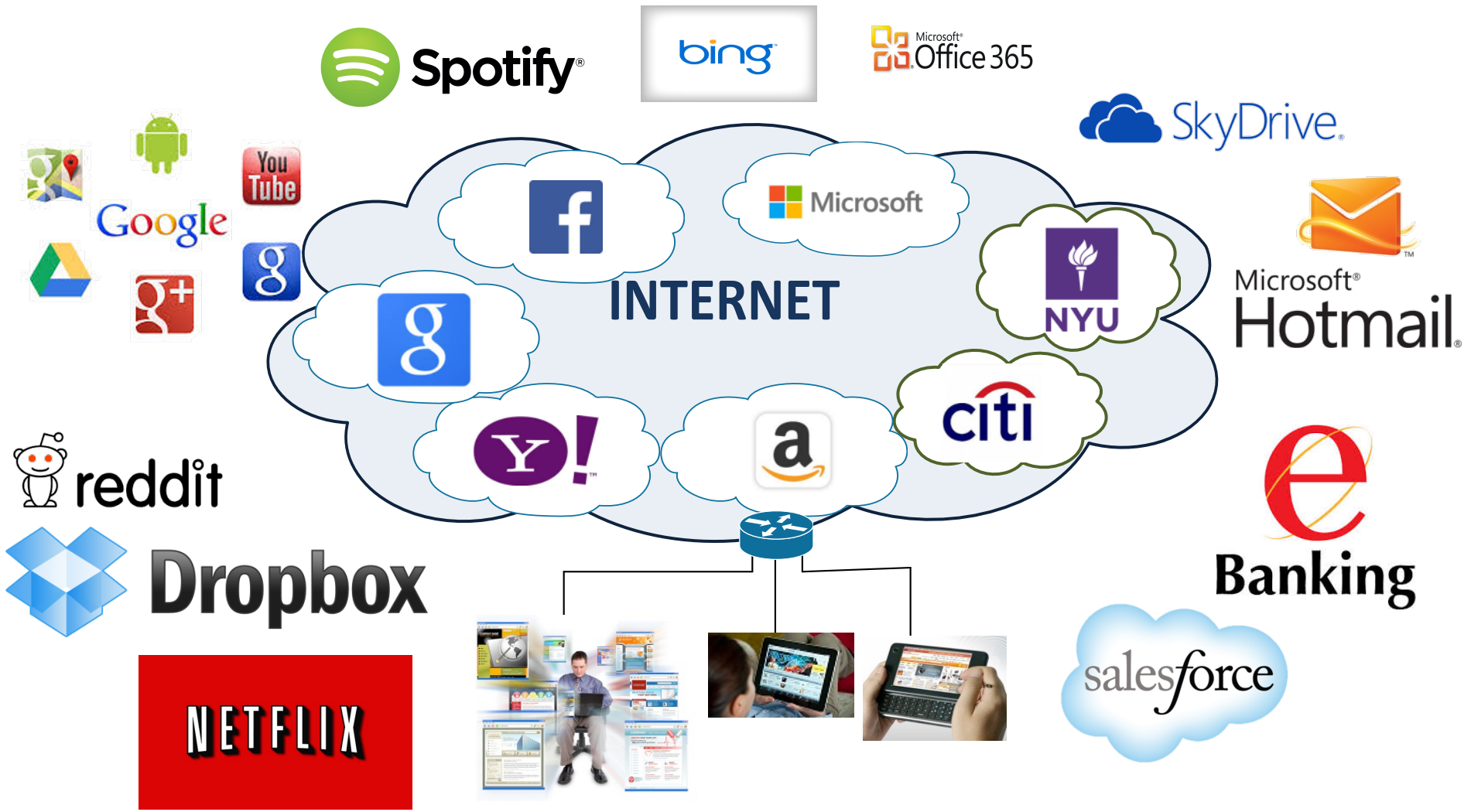
Outline

1. Introduction of cloud computing
2. Benefits of Cloud Computing
3. Success Stories of Building Business on Cloud Computing
4. Could Service Models
5. Could Deployment Models
6. Could Computing Resources
7. Introduction to Data Centers
8. Edge Computing

1. Introduction of Cloud Computing

Cloud Computing Facilitated by Computers, Communications, Control (3C)

Many Apps on Cloud Computing



What is Cloud Computing?

- Cloud computing is the delivery of computing services — servers, storage, databases, networking, tools, applications, analytics, and more—over the Internet ("the cloud").
- Companies offering these computing services typically charge for cloud computing services based on usage.
- Cloud computing provides rapid access to free, or low cost services such as:
  - Email, edit/store documents, phone calls, watch movies or TV, listen to music, play games, store/download pictures
  - Critical operations of business: create new apps and services; store, back up, and recover data; host websites; deliver software on demand; analyze data for patterns and make predictions (Big Data), and AI computations

[https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/]

powered by aws

Products of Amazon Web Services


Analytics


Application Integration


AR & VR


AWS Cost Management


Blockchain


Business Applications


Compute


Containers


Customer Engagement


Database


Developer Tools


End User Computing


Game Tech


Internet of Things


Machine Learning


Management & Governance


Media Services


Migration & Transfer


Mobile


Networking & Content Delivery


Quantum Technologies


Robotics


Satellite


Security, Identity & Compliance


Storage

Products and Services

Application development

- If people are developing web, mobile, or gaming apps, the cloud can help them quickly create cross-platform experiences that scale as their user base grows. Many cloud services include pre-coded tools—such as directory services, search, and security—that can speed and simplify their development.

Test and development

- The cloud can provide an environment to help save costs and bring the apps to market faster. Rather than securing budgets and spending valuable project time and resources setting up physical environments, the teams can quickly set up and dismantle test and development environments in the cloud. People can scale these dev-test environments up or down depending on need.

Big data analytics

- With cloud computing, people can tap into their organization's data to analyze it for patterns and insights, make predictions, improve forecasting, and make other business decisions. Cloud services can provide higher processing power and sophisticated tools for mining massive amounts of data, as well as the ability to quickly scale their environment as data grows.

https://azure.microsoft.com/en-us/overview/examples-of-cloud-computing/

Google Cloud Industry Solutions

Retail
Analytics and collaboration tools for the retail value chain.

Financial Services
Computing, data management, and analytics tools for financial services.

Healthcare and Life Sciences
Health-specific solutions to enhance the patient experience.

Media and Entertainment
Solutions for content production and distribution operations.

Telecommunications
Hybrid and multi-cloud services to deploy and monetize 5G.

Gaming
AI-driven solutions to build and scale games faster.

Manufacturing
Migration and AI tools to optimize the manufacturing value chain.

Energy
Multi-cloud and hybrid solutions for energy companies.

Government
Data storage, AI, and analytics solutions for government agencies.

Education
Teaching tools to provide more engaging learning experiences.

Small and Medium Business
Explore SMB solutions for web hosting, app development, AI, analytics, and more.

Cloud Natives
Resources and solutions for cloud-native organizations.

Alibaba Cloud | Worldwide Cloud Services Partner

Products and Services


Elastic Computing


Storage & CDN


Networking


Database Services


Security

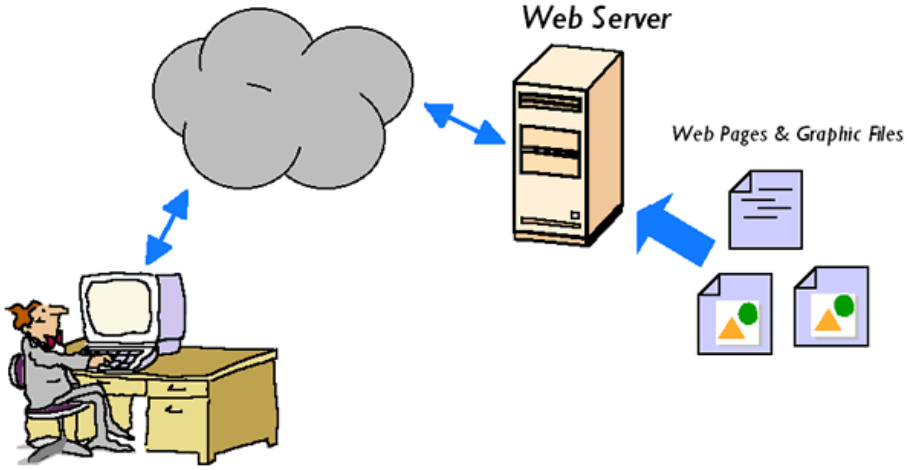
- Alibaba Cloud Elastic Compute Service (ECS) can provide hundreds of thousands of vCPUs in minutes for a single customer in a single region by using sophisticated smart placement algorithm, and dynamic and automatic planning.
- Alibaba Cloud Object Storage Service (OSS) is an encrypted, secure, cost-effective, and easy-to-use object storage service to store, back up, and archive large amounts of data in the cloud, with a guaranteed durability of 99.9999999999%.
- Alibaba Virtual Private Cloud (VPC) helps build an isolated network environment, including customizing the IP address range, network segment, route table, and gateway.

2. Benefits of Cloud Computing

How to build a website? The old way…

- Buy a server, wait for a few days.
- Install OS + software, half day.
- Find a hosting company, a few days.
- Configure networks, bind to a domain name, a few hours.
- Design and upload your website, hours to days.

- Wait… That's not all. What about upgrading and maintenance?!

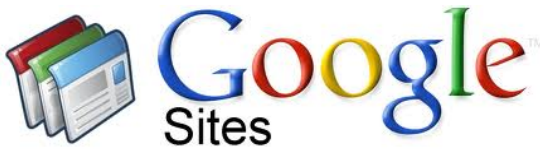


Why don’t we like the old way?

- Time to market is long.
- Upgrading the hardware is hard.
- Upgrading the software is non-trivial.
- Failures?
  - Let’s hope that it won’t happen.
  - But, we know it is impossible!

The easier way for a simple website

- How does Google site work?
- One can create a website on sites.google.com in a few minutes—from scratch.
- We don't really think about a single word about hardware, networking, OS, web software.
- And we don't care about upgrading, maintenance, failures.
- Why? Google does all the dirty jobs!



The key elements

- OS + web server software
- Networking, Domain Name Server (DNS)
- Web authoring tools
- Search engine
- Availability

Top benefits of Cloud Computing

1. Cost: Cloud computing eliminates the capital expense of buying hardware and software and setting up and running on-site datacenters — the racks of servers, the round-the-clock electricity for power and cooling, the IT experts for managing the infrastructure.
2. Speed: Most cloud computing services are provided self service and on demand, so vast amounts of computing resources can be provisioned in minutes, typically with just a few mouse clicks, giving businesses a lot of flexibility and taking the pressure off capacity planning.
3. Global scale: Scale elastically to deliver the right amount of IT resources (computing power, storage, bandwidth) right when they are needed, and from the right geographic location.

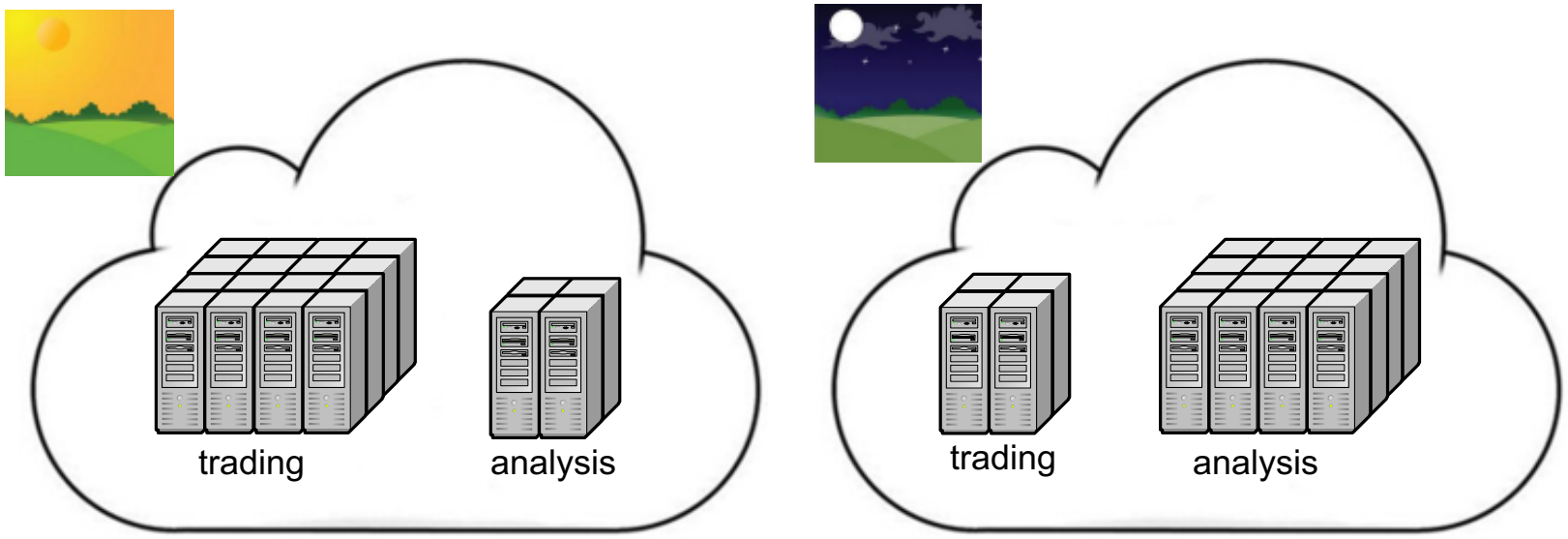
[https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/]

Top Benefits of Cloud Computing

4. Productivity: Datacenters require a lot of “racking and stacking”, hardware set up, software patching, and other time-consuming IT management chores. Cloud computing removes the need for many of these tasks, so IT teams can spend time on achieving more important business goals.
5. Performance: Cloud computing services run on a worldwide network of secure datacenters, which are regularly upgraded to the latest generation of fast and efficient computing hardware.
6. Reliability: Cloud computing makes data backup, disaster recovery, and business continuity easier and less expensive, because data can be mirrored at multiple redundant sites on the cloud provider’s network.

[https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/]

Efficient Resource Usage



- Computing resources can be used in a time-interleaved way to improve efficiency and reduce cost.

Cases for Cloud Computing Economics

- When demand for a service varies with time (provisioning for peak load is a waste)
- When demand is unknown in advance (no ideas of how many customers in the future)
- When performing batch analytics by massive servers to finish computations faster
  - cost of using 1,000 EC2 (Amazon Elastic Cloud Compute) machines in 1 hr

    = one EC2 machine for 1,000 hrs

Gartner: IT Spending Will Jump More Than $1T

- More than $1 trillion in IT spending will be directly or indirectly affected by the cloud shift during the next five years from 2016. Gartner also says that the cloud will be one of the most impactful forces of IT spending since the early days of the digital age.
- Traditional IT spending is shifting over to cloud services, and the market for cloud services has grown to a notable percentage of total IT spending. The collective amount of cloud shift in 2016 alone is estimated to reach $111 billion, increasing to $216 billion in 2020.
- More and more data traffic flows in and out from data centers for different applications, causing the Internet traffic grows exponentially.
- Could you think of any applications not done in data centers?

https://www.sdxcentral.com/articles/news/gartner-it-spending-jump-more-1t/2016/07/

Global Mobile Data Traffic Growth / Top-Line

Global Mobile Data Traffic will Increase 7-Fold from 2016–2021

47% CAGR
2016–2021

Compound Annual Growth Rate

- **Exa=10^{18}**

Source: Cisco VNI Global Mobile Data Traffic Forecast, 2016–2021

https://disruptive.asia/5g-mobile-data-growth-2020-cisco-vni/

Global Mobile Data Traffic Growth

- By 2021, global mobile data traffic will reach 49 exabytes per month or 587 exabytes annually.
- Mobile data traffic will represent 20% of total IP traffic – up from just 8% of total IP traffic in 2016.
- Mobile network connection speeds will increase threefold from 6.8 Mbps in 2016 to 20.4 Mbps by 2021.
- Machine-to-machine (M2M) connections will represent 29% (3.3 billion) of total mobile connections – up from 5% (780 million) in 2016.
- M2M will be the fastest growing mobile connection type as global IoT (Internet of Things) applications continue to gain traction in consumer and business environments.
- The total number of smartphones (including phablets) will be over 50% of global devices and connections (6.2 billion)—up from 3.6 billion in 2016.
- Cisco and other industry experts anticipate large-scale deployments of 5G infrastructures to begin by 2020.
- Mobile carriers will need the innovative speed, low latency, and dynamic provisioning capabilities that 5G networks are expected to deliver to address not just increasing subscriber demands but also new services trends across mobile, residential, and business markets.

Global Mobile Data Traffic Drivers

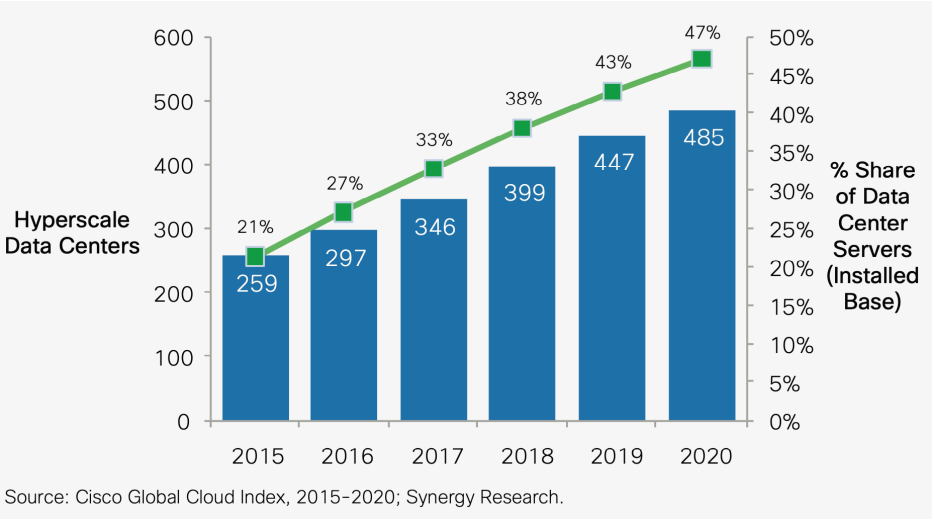
Mobile Momentum Metrics

		2016	2021
More Mobile Users		4.9 Billion	5.5 Billion
More Mobile Connections		8 Billion	12 Billion
Faster Mobile Speeds		6.8 Mbps	20.4 Mbps
More Mobile Video		60% of Traffic	78% of Traffic

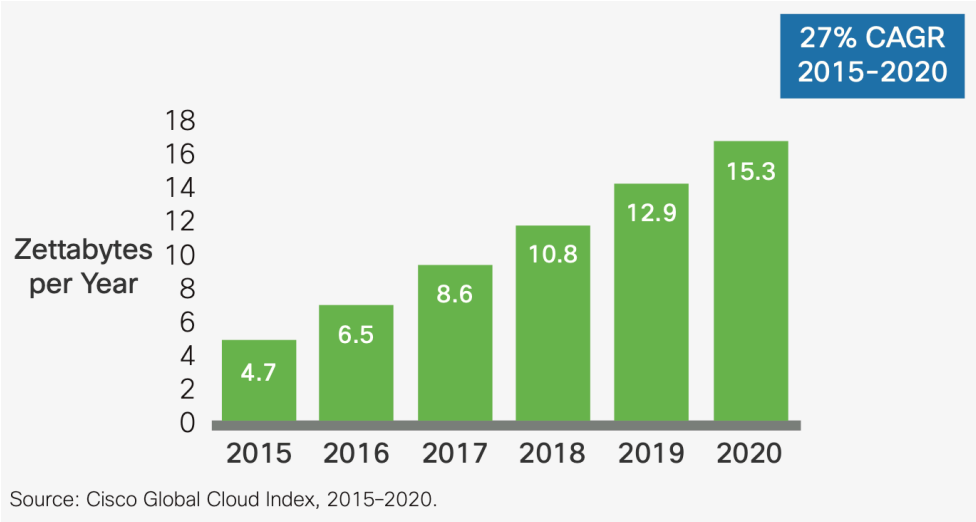
Source: Cisco VNI Global Mobile Data Traffic Forecast, 2016–2021

https://disruptive.asia/5g-mobile-data-growth-2020-cisco-vni/

Data Center Number and Traffic Growth



Source: Cisco Global Cloud Index, 2015–2020; Synergy Research.



Source: Cisco Global Cloud Index, 2015–2020.

- These hyperscale data centers will grow from 259 in number at the end of 2015 to 485 by 2020.
- They will represent 47 percent of all installed data center servers by 2020.
- The amount of annual global data center traffic in 2015 is already estimated to be 4.7 ZB (Zettabytes, 10^{21} bytes) and by 2020 will triple to reach 15.3 ZB per year.
- The amount of global traffic crossing the Internet and IP WAN networks is projected to reach 2.3 ZB per year by 20201.

https://www.iotjournaal.nl/wp-content/uploads/2017/02/white-paper-c11-738085.pdf

Global Internet Phenomena

Video is almost **58%** of the total downstream volume of traffic on the internet

NETFLIX is **15%** of the total downstream volume of traffic across the entire internet

More than **50%** of internet traffic is encrypted, and TLS 1.3 adoption is growing

BITTORRENT is almost **22%** of total upstream volume of traffic, and over **31%** in EMEA alone

GAMING is becoming a significant force in traffic volume as gaming downloads, Twitch streaming, and professional gaming go mainstream

40.2% of all connections in the APAC region connect to Google services

https://www.sandvine.com/hubfs/downloads/phenomena/2018-phenomena-report.pdf

3. Success Stories of Building Business on Cloud Computing

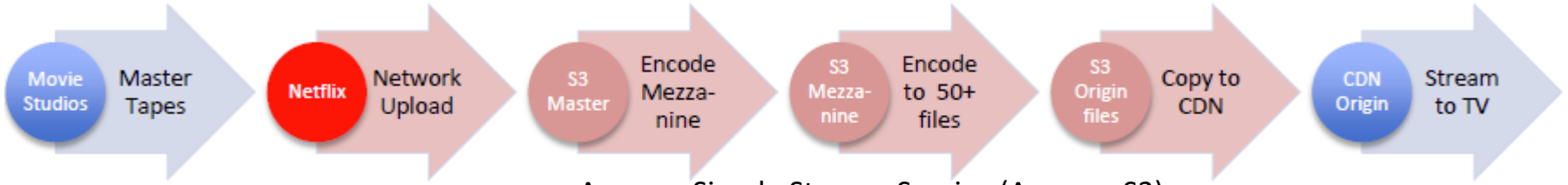
Brief history of Netflix

- 2007 Start of Internet streaming
- 2010 Migrate to the Amazon Cloud
- 2010 Streaming > DVD mailing
- 2010 Expansion to Canada
- 2012 > 30 million subscribers
- 2013 > 44 million subscribers

- In January 2017, > 93 million subscribers worldwide, including > 49 million in US
- As of Oct 2018, the streaming video giant consumes 15% of the total downstream volume of traffic globally. In the United States, that figure jumps to 19.1% of total traffic.

Netflix-Amazon-CDN

Netflix uses Amazon Web Services (AWS) for their storage and compute operations with S3 being their system of record.



Amazon Simple Storage Service (Amazon S3)

CDN: Content Delivery Network

S3: Simple Storage Service

Licensed content is provided to Netflix as high quality master tapes

Many formats are reduced to a single high quality mezzanine format on S3

Individual formats and speeds are encoded in over 50 combinations

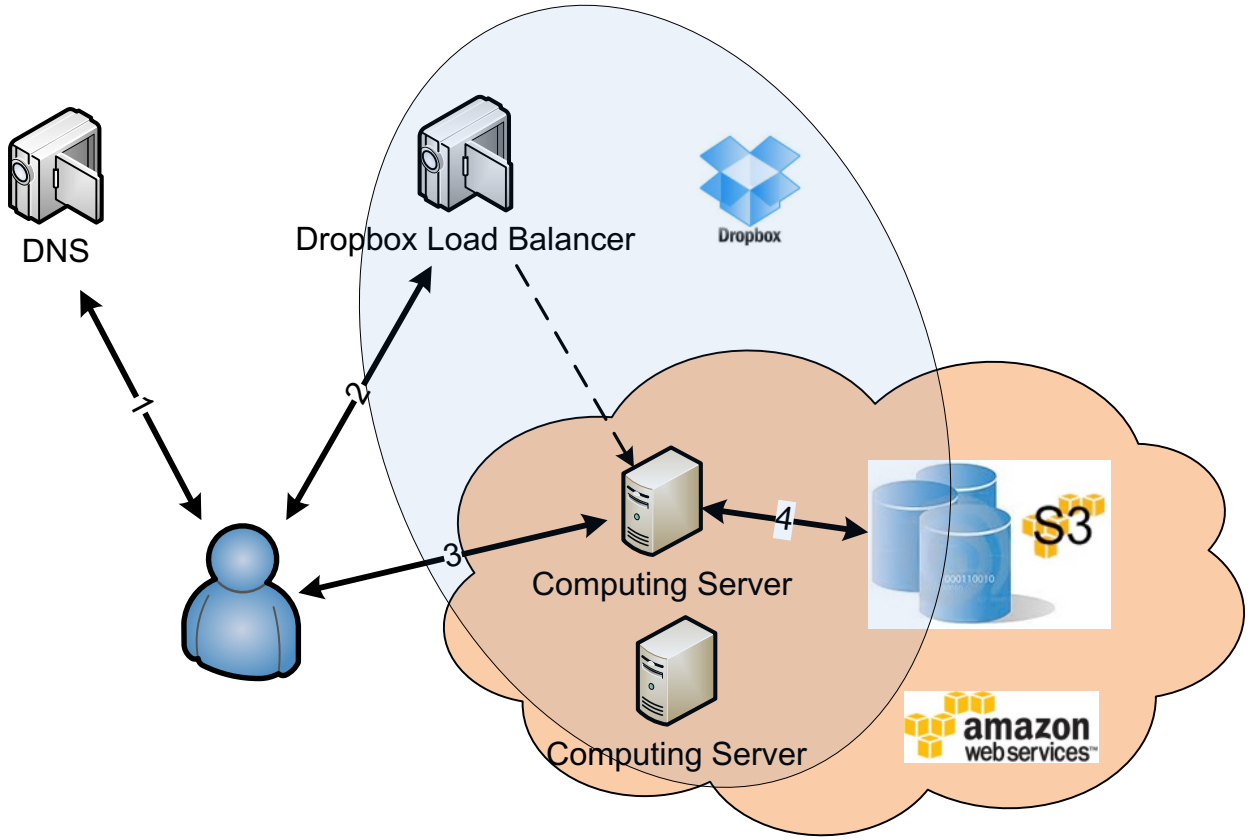
- Many formats for older and newer hardware and various game consoles
- Many speeds from mobile through standard and high definition

Static files are copied to each Content Delivery Network's "origin server"

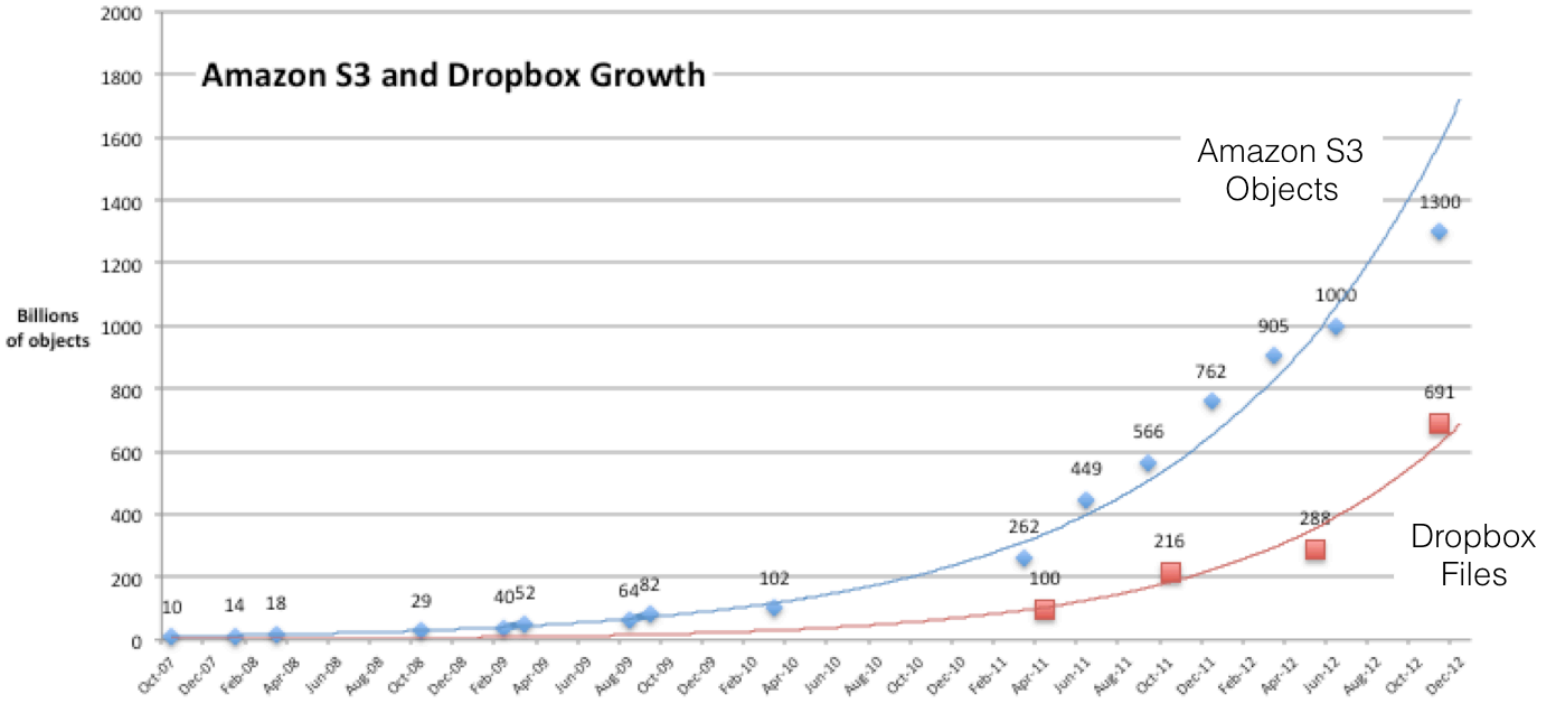
CDNs migrate files to "edge servers" near the end user

Files stream to PC/Mac/iPad or TV over HTTP using "range get" to move chunks

Dropbox - Amazon



Dropbox & Amazon

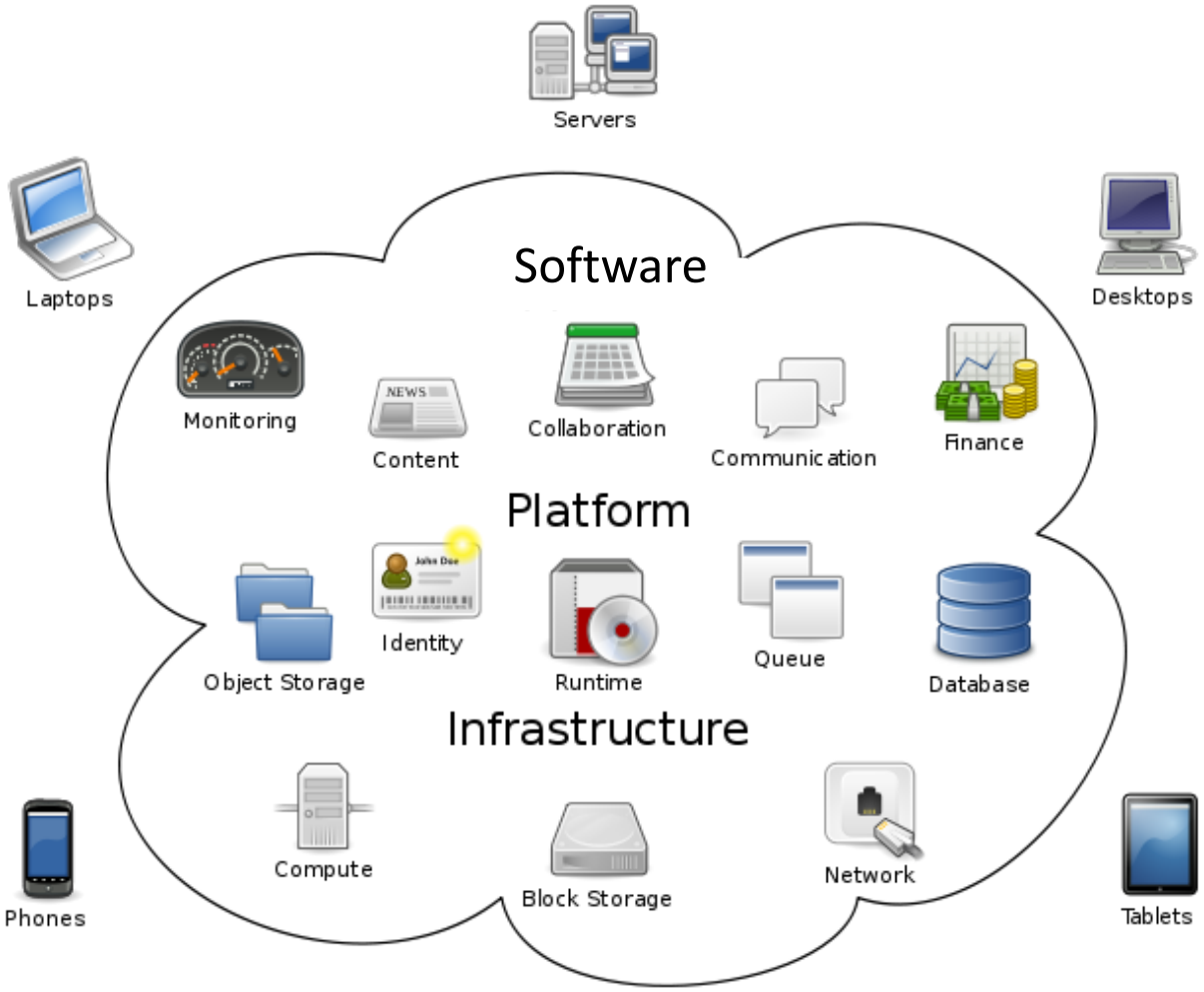


[blog.oxygencloud.com]

- Dec 2012: Amazon S3 stores 1.3 Trillion objects.
- July 2012: Dropbox has over 50M users, adding 1 billion files every day.
- Nov 2013: Dropbox hit 200M users

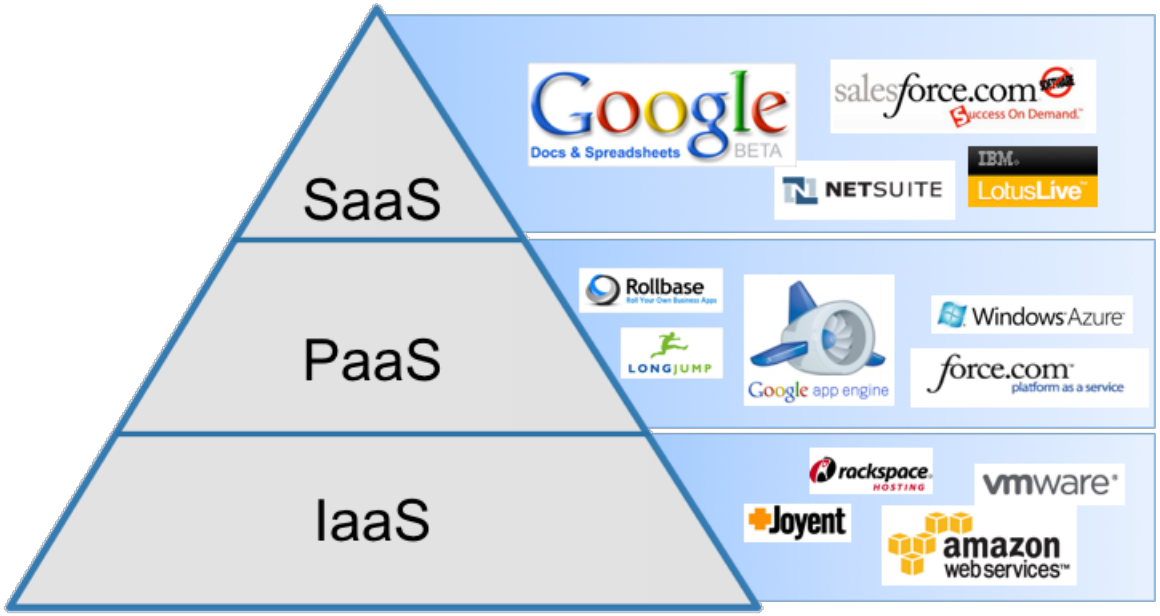
4. Could Service Models

Cloud service models



Cloud service models

- Software as a Service (SaaS): email, customer relationship management (CRM)
- Platform as a Service (PaaS): App Dev
- Infrastructure as a Service (IaaS): networking, security, system management



Use Model:

IaaS/Cloud

PaaS/Cloud, PaaS/IaaS/Cloud

Saas/Cloud, SaaS/PaaS/Cloud

SaaS/PaaS/IaaS/Cloud

Cloud service models

SalesForce CRM

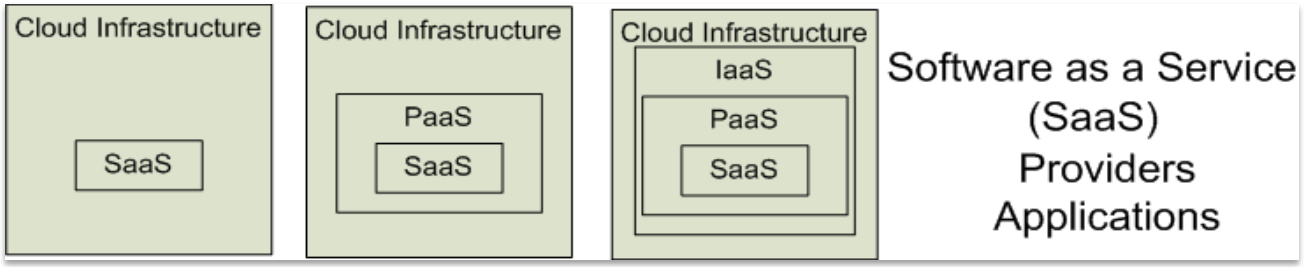
LotusLive

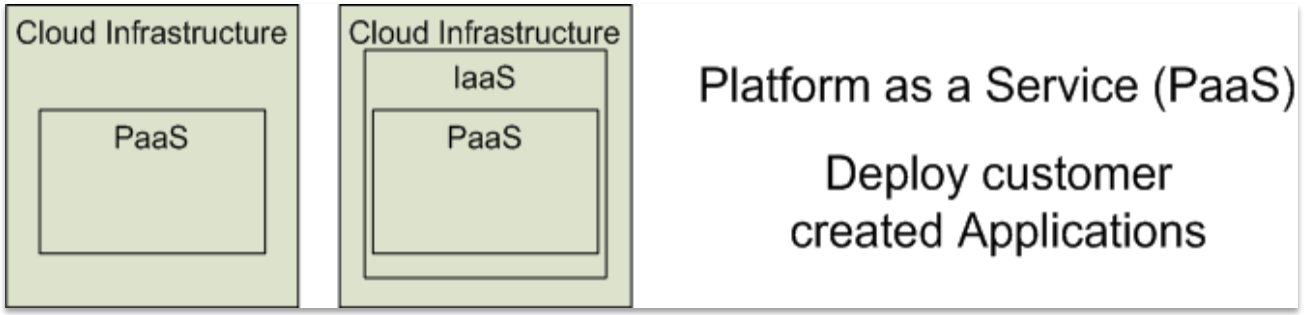
Google App

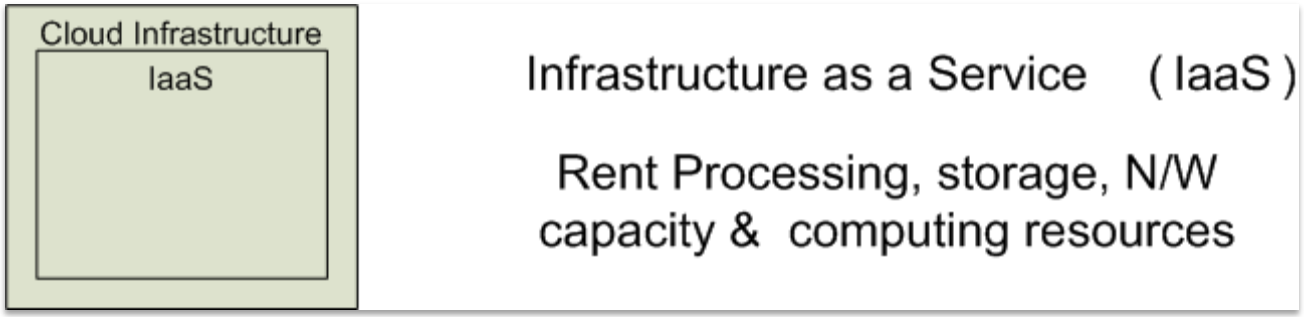
Windows Azure
The Future Made Familiar

amazon web services™

rackspace HOSTING







Adopted from: Effectively and Securely Using the Cloud Computing Paradigm by peter Mell, Tim Grance

Software as a Service (SaaS)

- SaaS is a method for delivering software applications over the Internet, on demand and typically on a subscription basis.
- With SaaS, cloud providers host and manage the software application and underlying infrastructure, and handle any maintenance, like software upgrades and security patching.
- SaaS allows an enterprise to get quickly up and running with an app at minimal upfront cost.
- Users can run most SaaS apps directly from their web browser on their phone, tablet, or PC without needing to download and install any software, although some apps require plugins.
- Users do not have control over the infrastructure and software, except limited configurations.

Software as a Service (SaaS)

- To provide SaaS apps to users, the enterprise doesn't need to purchase, install, update, or maintain any hardware, middleware, or software.
- Common examples are email, calendaring, and office tools (such as Microsoft Office 365), Google Docs, Dropbox, Zoho (including office, CRM, project management, invoice…)

Customer relationship management (**CRM**) is a model for managing a company's interactions with current and future customers. It involves using technology to organize, automate, and synchronize sales, marketing, customer service, and technical support [Wikipedia].

Platform as a Service (PaaS)

- PaaS supplies an on-demand environment for developing, testing, delivering, and managing software applications, E.g., Google App Engine
- PaaS is designed to make it easier for developers to quickly create web or mobile apps, without worrying about setting up or managing the underlying infrastructure of servers, storage, network, and databases.
- Users can cut the time it takes to code new apps with pre-coded application components built into the platform, such as workflow, directory services, security features, search, and so on.
- Enterprises are moving from proprietary platforms like Oracle Forms and Microsoft .NET to open Java platforms. For instance, WaveMaker migrates proprietary applications to use the latest Web 2.0 and Java technologies.

WaveMaker

force.com
platform as a service

Infrastructure as a Service (IaaS)

- The most basic category of cloud computing services. With IaaS, you rent IT infrastructure—servers and virtual machines (VMs), storage, networks, operating systems—from a cloud provider on a pay-as-you-go basis with a quick scale up or down.
- With IaaS, the user is able to deploy and run arbitrary software, which can include operating systems and applications.
- The user does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls).
- E.g., Give me 100 VMs interconnected as a LAN.
- An example is Amazon Elastic Compute Cloud (Amazon EC2)

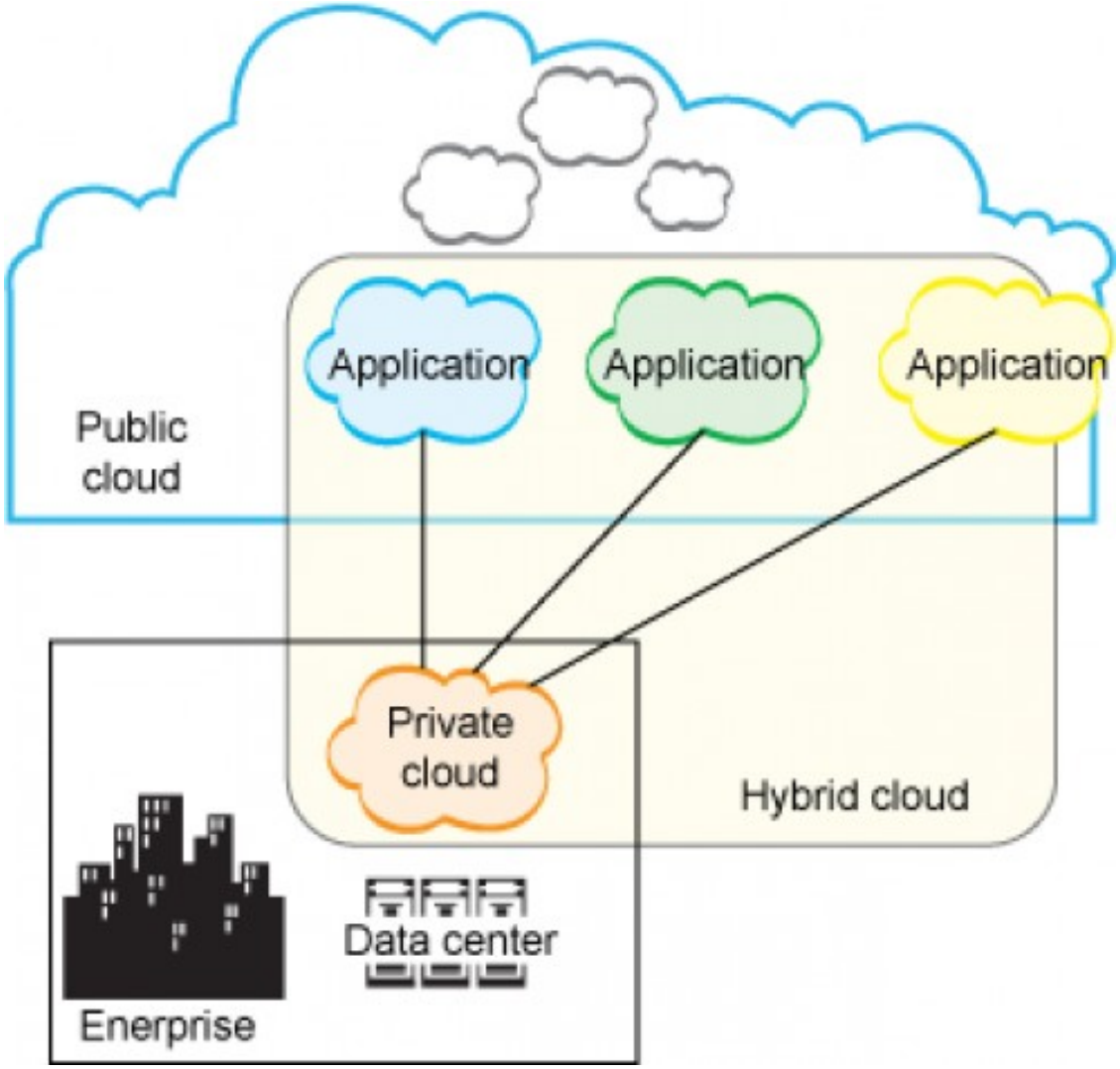
Some major IaaS providers

Provider	Pricing	Average Price / Month (US$)	SLA	Datacenters	Scale Up	Scale Out	OSs	Instance Types	Data Transfer out (/GB)	Data Transfer in (/GB)
Amazon EC2	Pay-as-you-go or Year + Discount	80.81	99.95%	7	No	Yes	9	12	0.12	0
BitRefinery	Monthly	137	100%	1	Yes	Yes	3	Configurable	0	0
GoDaddy	Monthly	39.99	99.90%	8	No	Yes	4	5	0	0
GoGrid	Pay-as-you-go or Monthly	273.6	100%	2	Yes	Yes	4	1	0.29	0
Hosting.com	Monthly	270	100%	4	Yes	Yes	3	Configurable	0	0
NephoScale	Pay-as-you-go or Year + Discount	146	99.95%	1	Yes	Yes	4	6	0.13	0
OpSource	Pay-as-you-go or Monthly	87.6	100%	4	Yes	Yes	4	Configurable	0.15	0
Rackspace	Pay-as-you-go	51.1	100%	9	Yes	Yes	8	8	0.18	0
ReliaCloud	Monthly	135.05	100%	2	No	Yes	5	5	0.12	0
Softlayer	Pay-as-you-go or Monthly	135.05	?	7	No	Yes	6	13	0.1	0
Terremark*	Pay-as-you-go	133.39	100%	9	Yes	Yes	5	Configurable	0.17	0.17

http://www.techrepublic.com/blog/datacenter/11-cloud-iaas-providers-compared/5285

5. Could Deployment Models

Deployment models



Deployment models

- Public cloud: owned and operated by a third-party cloud service provider, e.g., Microsoft Azure. All hardware, software, and other supporting infrastructure is owned and managed by the cloud provider. Users access these services and manage your account using a web browser.
- Private cloud: used exclusively by a single business or organization. It can be physically located on the company's on-site datacenter. Some companies also pay third-party service providers to host their private cloud. A private cloud is one in which the services and infrastructure are maintained on a private network.
- Hybrid cloud: combine public and private clouds, bound together by technology that allows data and applications to be shared between them. By allowing data and applications to move between private and public clouds, hybrid cloud gives businesses greater flexibility and more deployment options.
- Community cloud: provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations).

Enterprise Data Center (Private Cloud)

rtment of Education Data Center

NYC Department of Education

> 1,700 Schools

$24 Million Annual Budget

A

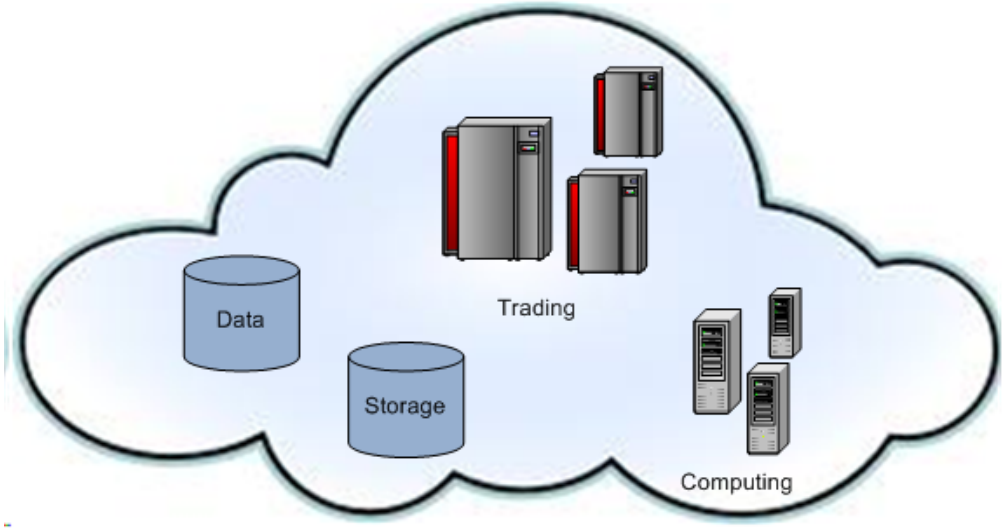
Site B

> 2K Servers

Links have been studied for 4 months

More than 500GB of data have been captured

NYSE and the Cloud



NYSE Euronext

- In Jun 2011, NYSE (New York Stock Exchange) rolled out its cloud system: Capital Markets Community Platform. It is the first financial services cloud delivering platform services to the capital markets community.
- The cloud provides data, storage, computing, trading, and other services on a unified platform.
- Key benefits include fast time to market, high scalability, simplified service access, low cost, and security.

NYSE Community Cloud

- Co-location is greatly facilitated
  - Customers rent VMs in the cloud and focus on their trading strategy, instead of worrying about physical infrastructure.
- New service enabled
  - Before: users download historical market data at midnight, analyze overnight, and return the results the next morning
  - After: data provisioning, computing, and trading are provided inside the cloud as a package. Enabling new trading strategies and services.
- Testing, risk control, and responding to government regulations are greatly facilitated.

6. Could Computing Resources

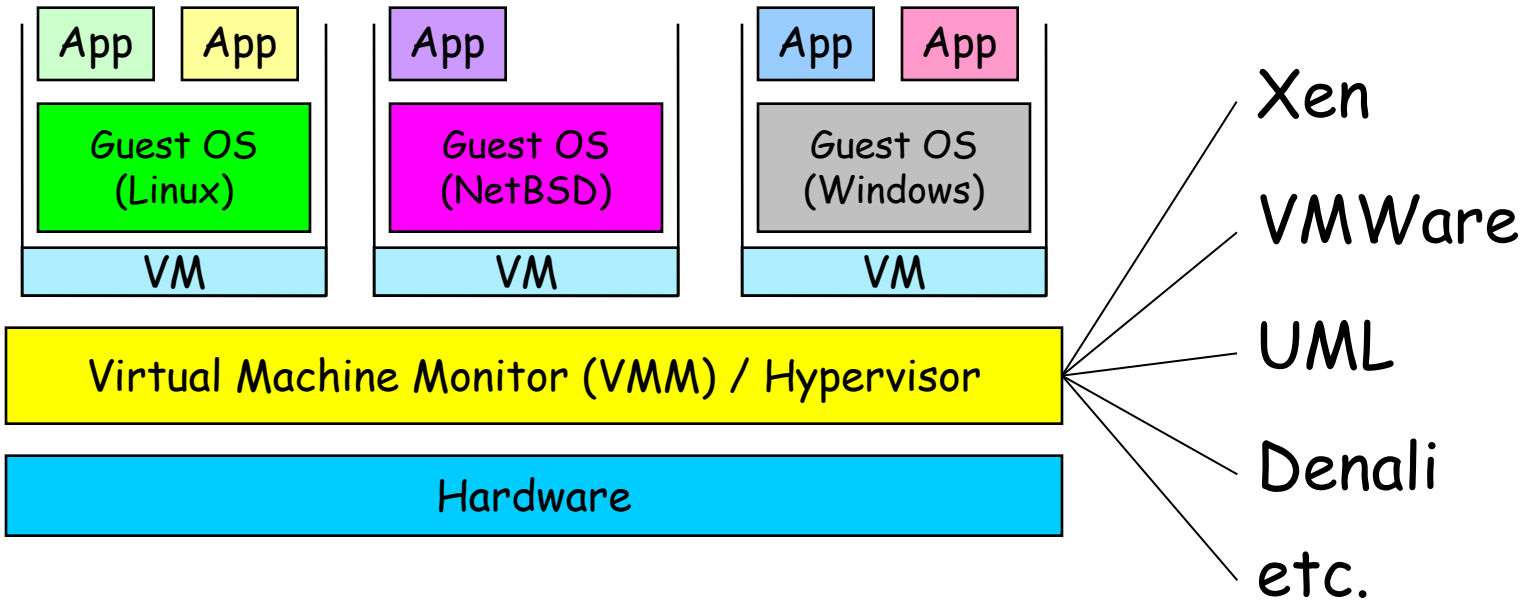
Virtualization Definition

- Virtualization is the ability to run multiple operating systems on a single physical system and share the underlying hardware resources*
- It is the process by which one computer appears as many computers.
- Virtualization is used to improve IT throughput and costs by using physical resources as a pool from which virtual resources can be allocated.

VMWare white paper, Virtualization Overview*

Virtual Machines (VMs)

- A VM is an isolated runtime environment (guest OS and apps)
- Hypervisor allows multiple VMs to run on a single physical machine.
- Each VM has IP/MAC address.

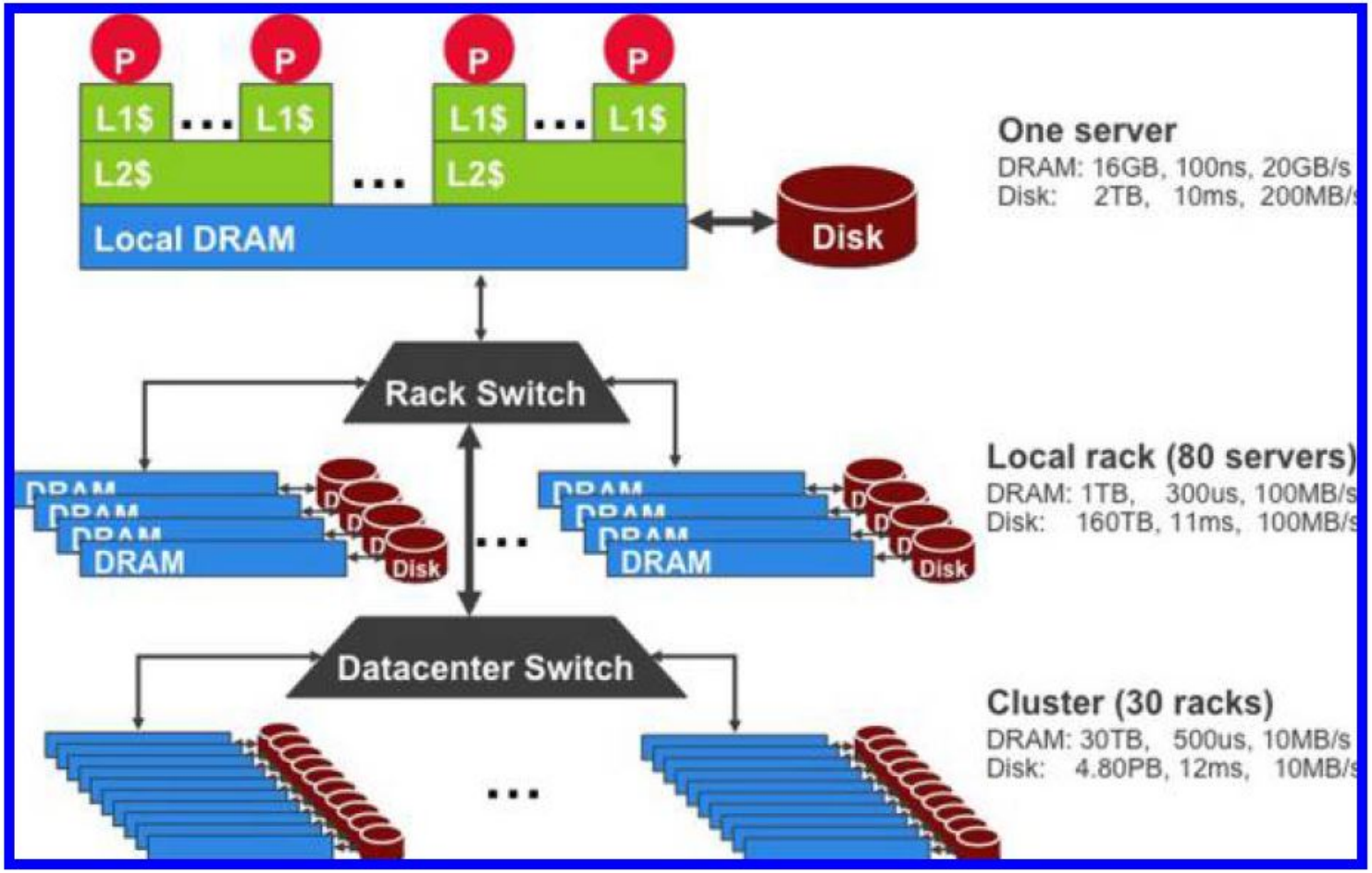


https://www.cse.unr.edu/~mgunes/cpe401/cpe401sp12/lect15_cloud.ppt

Benefits of Virtualization

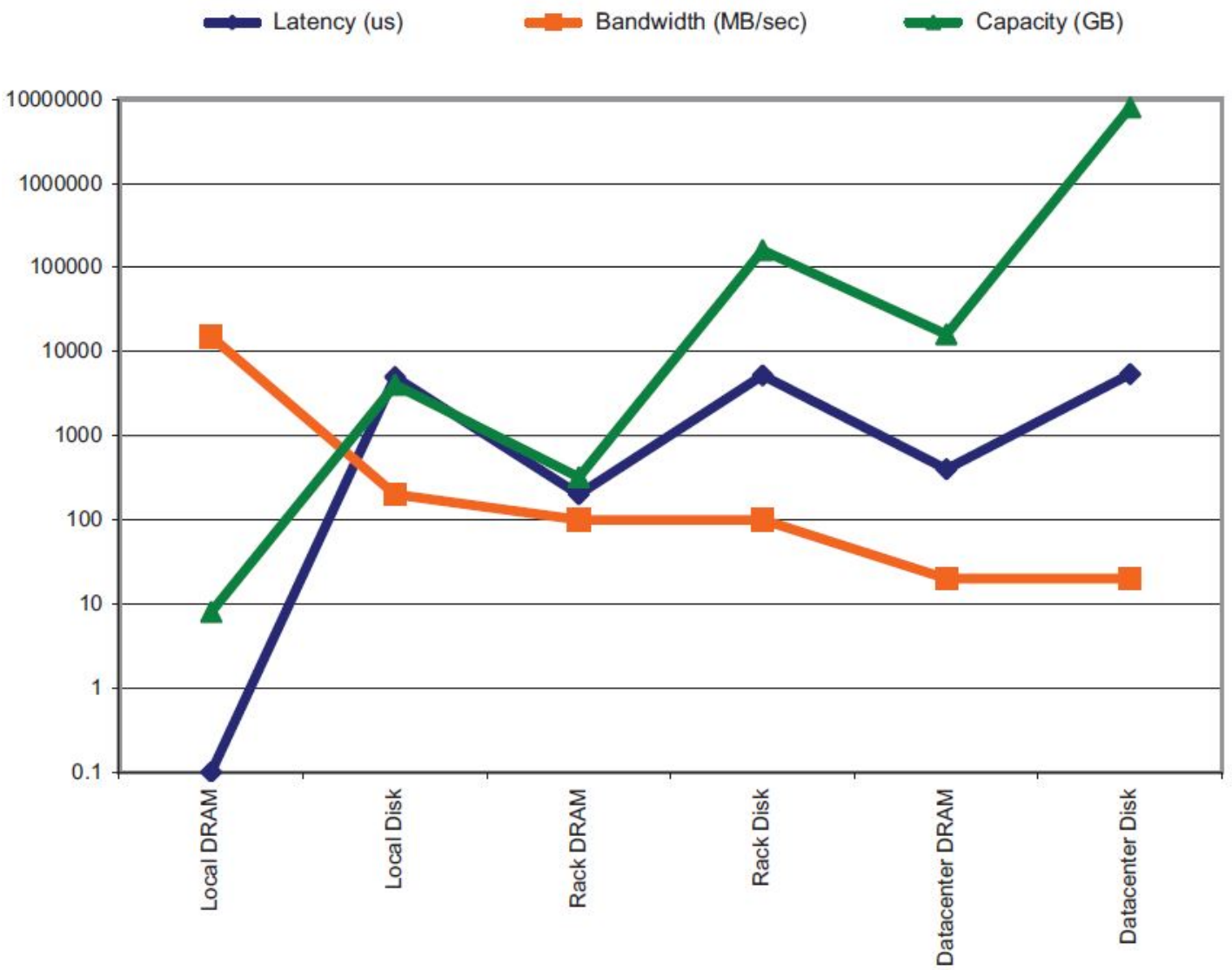
- Sharing of resources helps cost reduction
- Isolation: Virtual machines are isolated from each other as if they are physically separated
- Encapsulation: Virtual machines encapsulate a complete computing environment
- Hardware Independence: Virtual machines run independently of underlying hardware
- Portability: Virtual machines can be migrated between different hosts.

Storage Hierarchy in a Data Center



[Google]

Latency/bandwidth/capacity in a Data Center



[Google]

7. Introduction to Data Centers

Building a Data Center

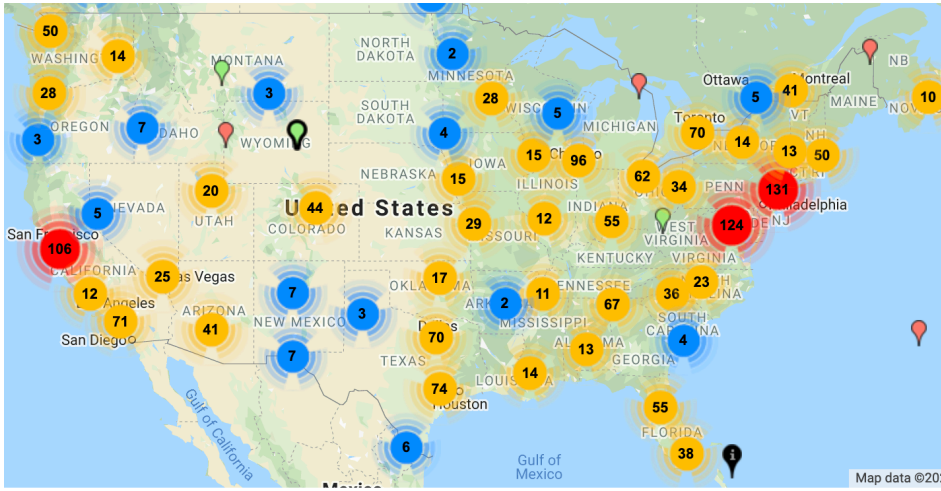
Microsoft video

Google gallery

Data Centers Everywhere

- Data centers are the infrastructure supporting cloud computing.
- A data center is a facility used to house computer systems and associated components, such as telecommunications and storage systems. It generally includes redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and various security devices. (Wikipedia)

Datacenter map in the world in 2020



Datacenter map in US in 2020

http://www.datacentermap.com/

World's Largest Data Center in Space

China Telecom Data Center, China

- Located in Hohhot (呼和浩特), Inner Mongolia Information Park
- Spans a grand total of 10.7M ft^2
- The most expensive data center in the world, cost over $3B
- Several factors make Hohhot an attractive location
  - An average annual rainfall of over 12 inches results vast reserves of hydroelectric power
  - At an average altitude of 1050 meters, the average temperature is 6°C (42.8°F) makes for "free air cooling for up to eight months a year".

https://www.technologymagazine.com/top10/top-10-biggest-data-centres-world/china-telecom-data-centre-china

The 3rd Largest Data Center in Space

The Citadel, United States

- The Citadel, still awaiting completion, lies near Reno in the North of Nevada.
- The facility covers 7.2M ft^2
- Will consume 650 Megawatts of power, 100% of which comes from renewable sources.
- Deliver 9-ms latency to Los Angeles and San Diego, with a 7-ms connection to the company's core facility
- The facility is also among the most innovative in the world, with over 260 patented innovations included in its construction and operation.

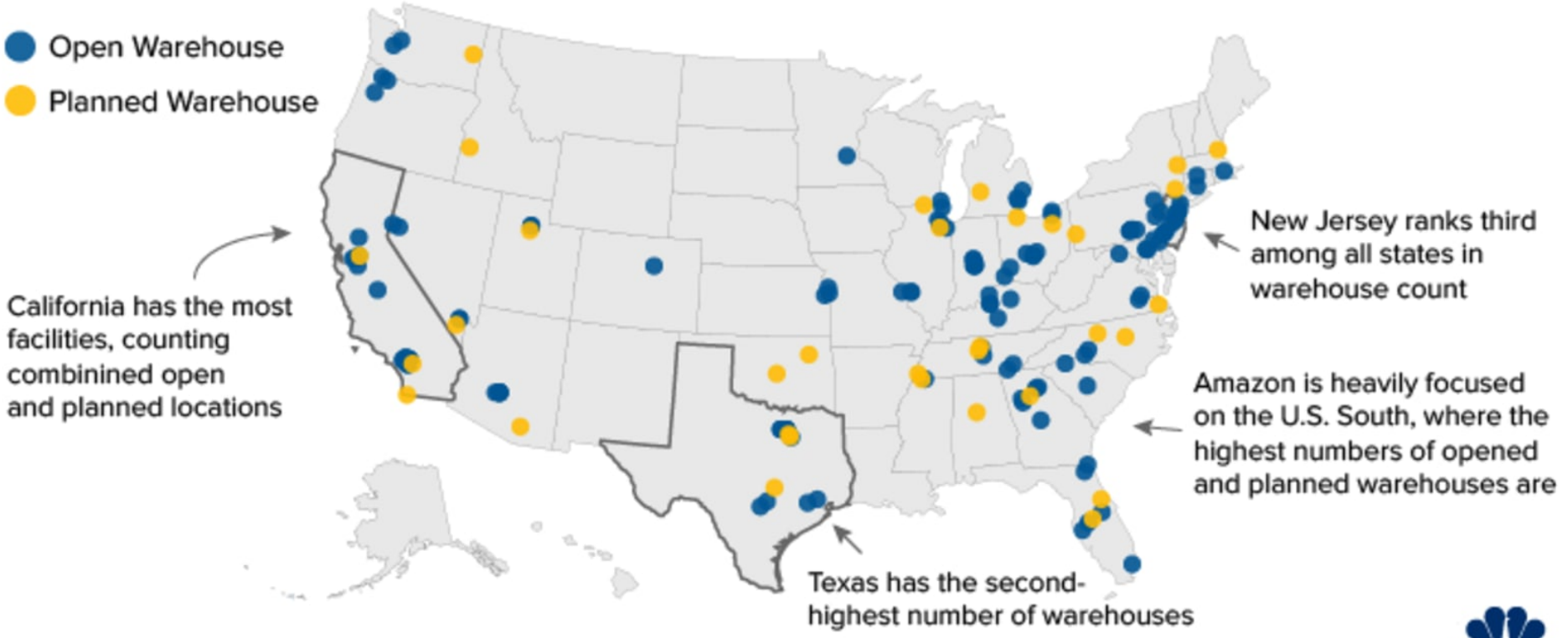
https://www.technologymagazine.com/top10/top-10-biggest-data-centres-world/china-telecom-data-centre-china

Google Cloud AI

- An eight-rack pod of Google's liquid-cooled TPU (Tenssor Processing Unit) version 3 servers for AI/ML.
- ML has produced business and research breakthroughs ranging from network security to medical diagnoses.
- Cloud TPU is the custom-designed machine learning ASIC that powers Google products like Translate, Photos, Search, Assistant, and Gmail.
- Users can put the TPU and machine learning to work accelerating their company's business, especially at scale.

Amazon fulfillment centers across the US

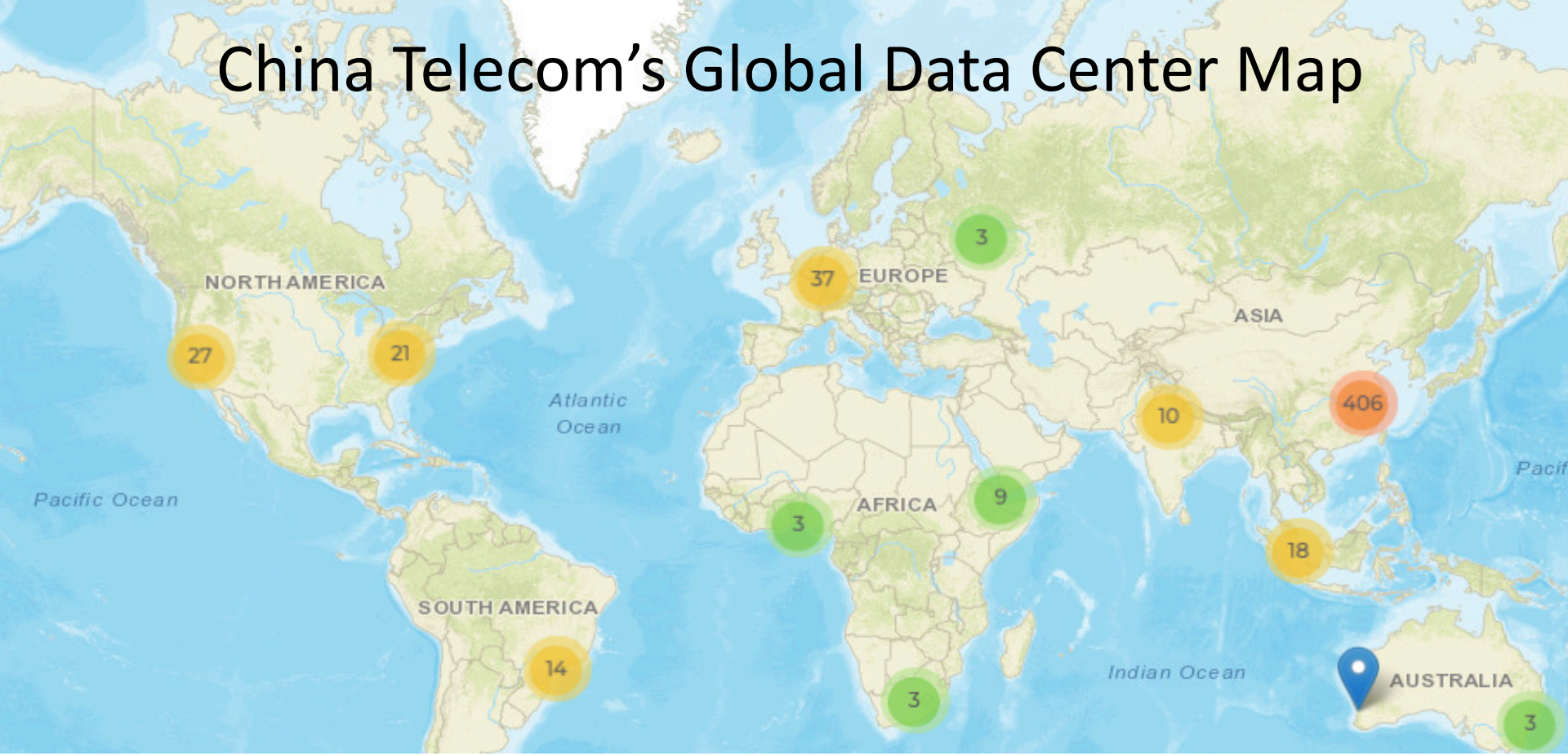
The company has over 100 active warehouses and dozens of planned locations



SOURCE: MWPVL International Inc., Supply Chain and Logistics Consultants. Includes warehouses of 500,000 square feet or more.

CNBC

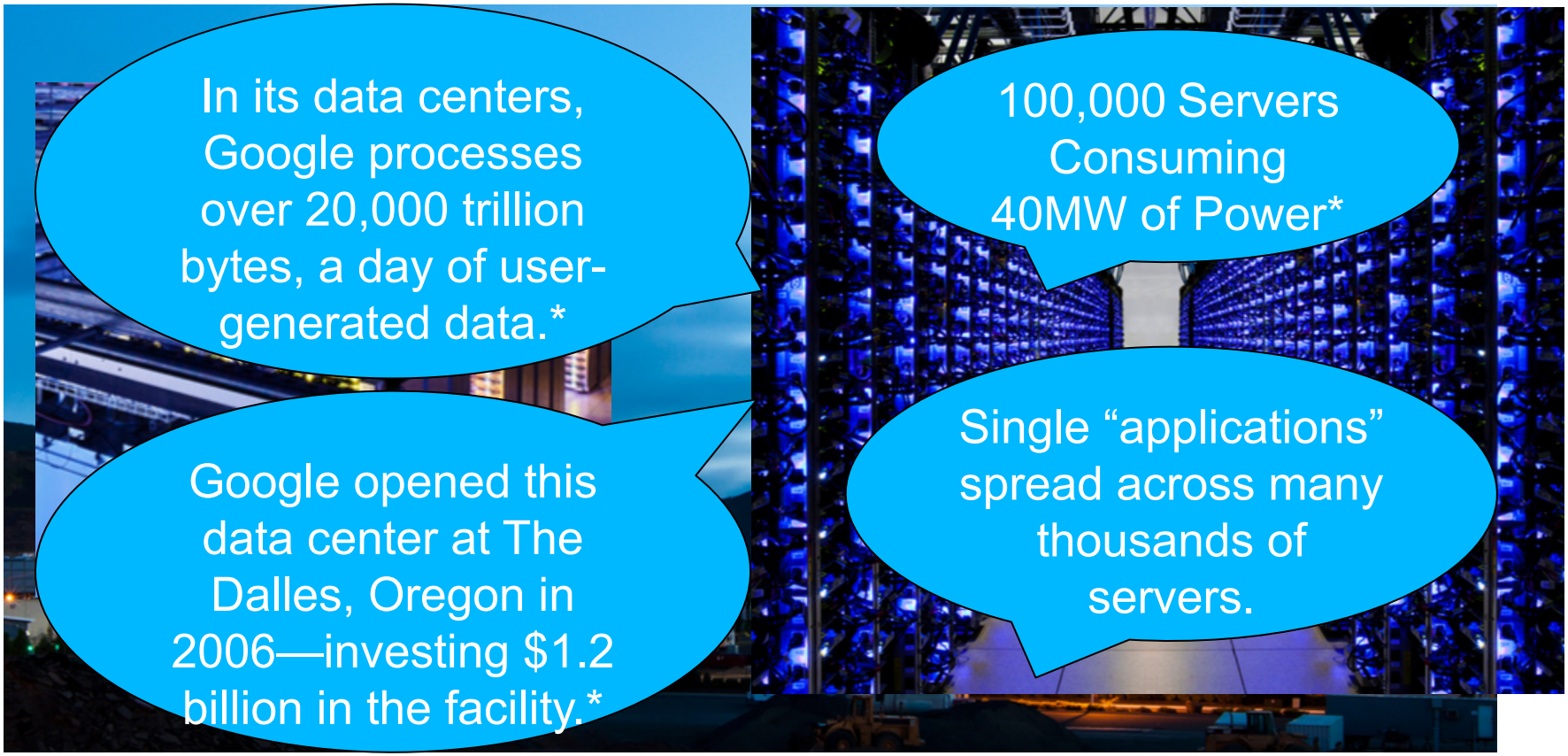
China Telecom's Global Data Center Map



- China Telecom operates 456 on-net data centers in Mainland China and 187 data centers across 71 key metro hubs globally.
- From 100+ global PoPs (point of presences) providing a variety of Layer 2 and 3 access options to China Telecom's backbone networks.
- Colocation services are available at over 100 off-net sites including equipment management, website hosting, hybrid/private cloud deployments, back-up and disaster recovery services.

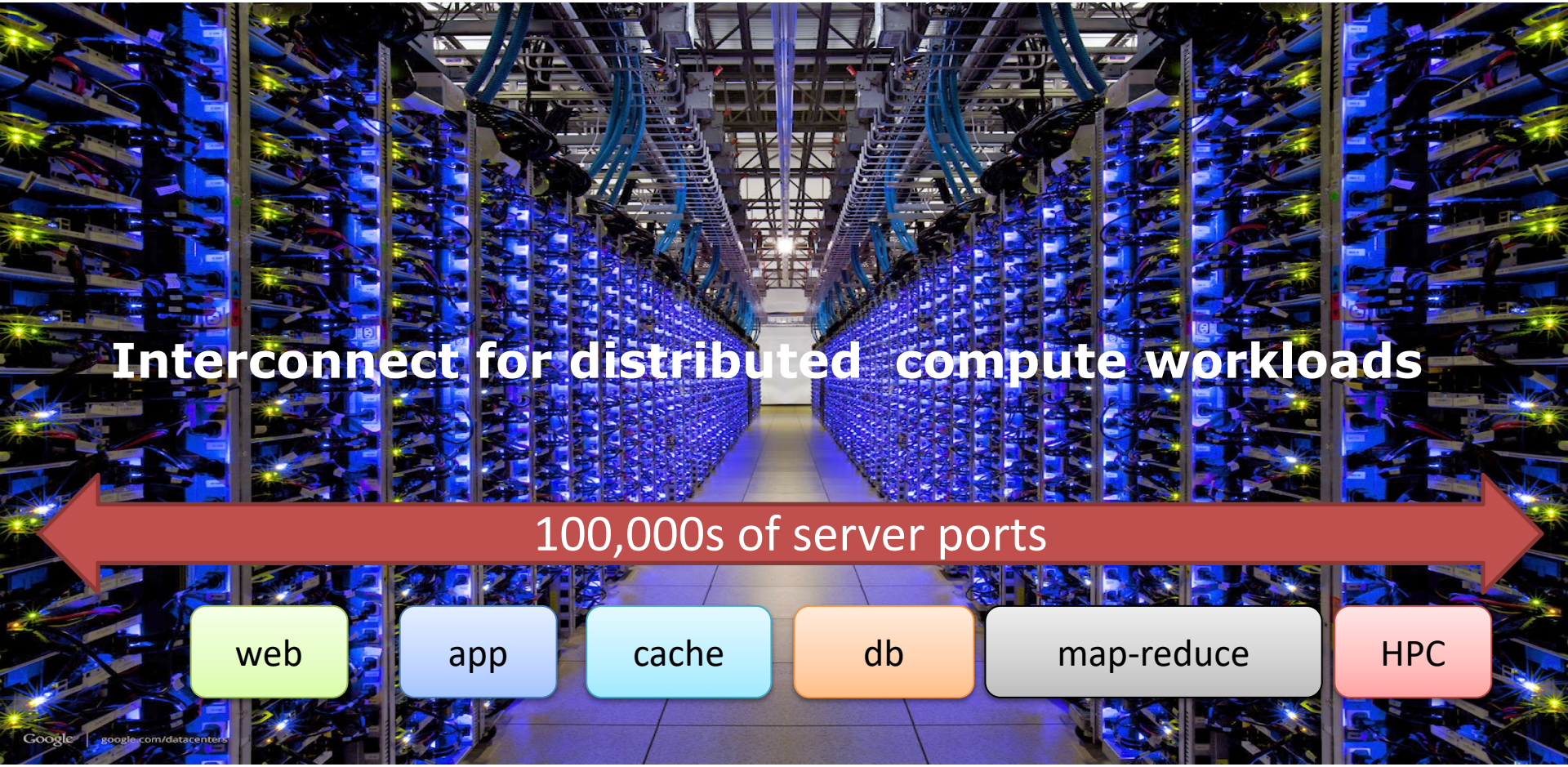
What does a data center look like?

A Google Data Center

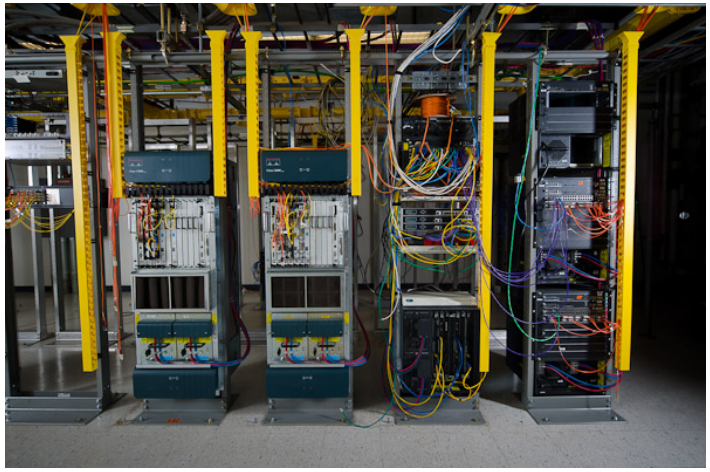


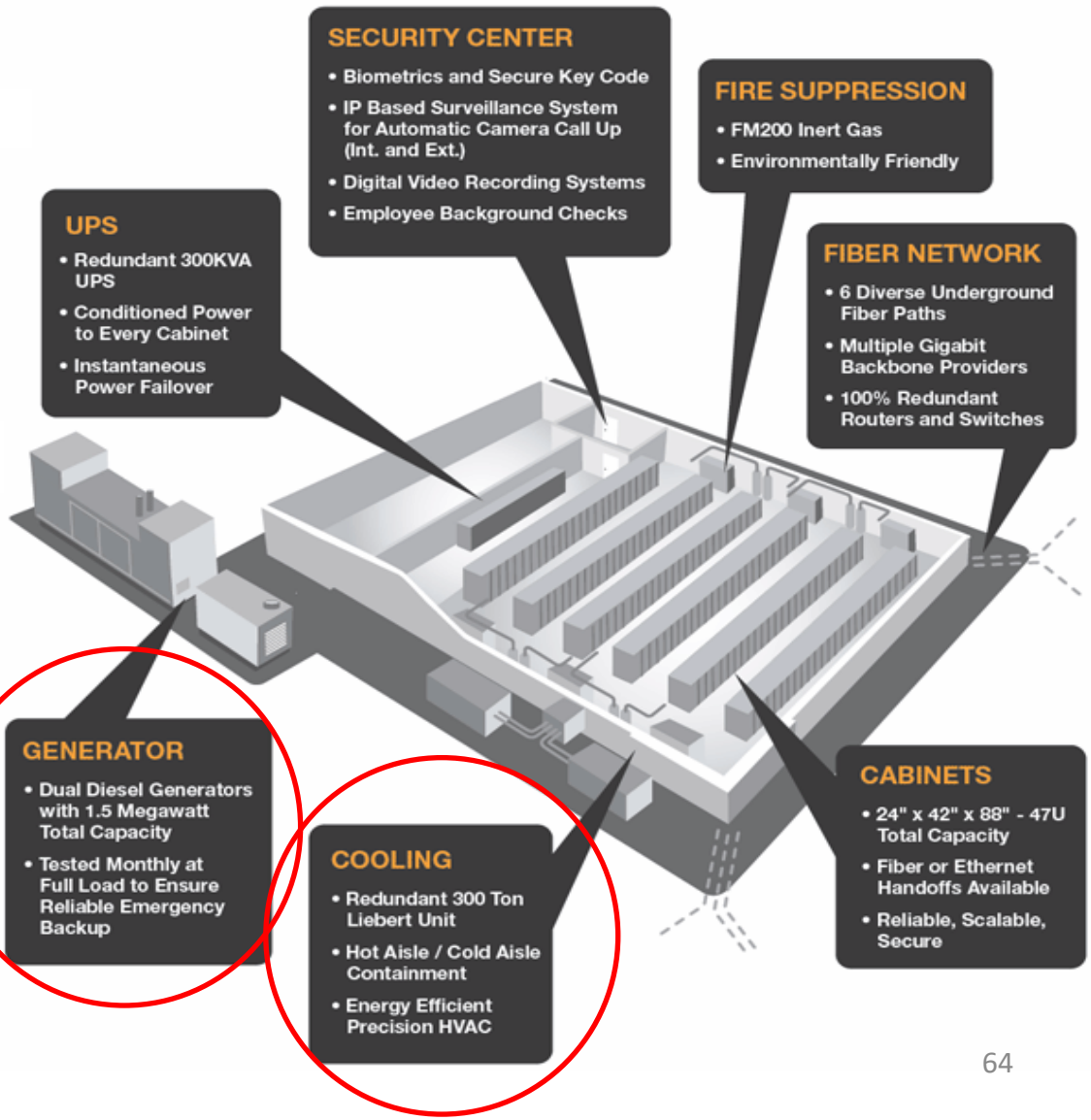
* All figures are taken from http://www.google.com/about/datacenters/inside/locations/the-dalles/index.html

Datacenter Networks

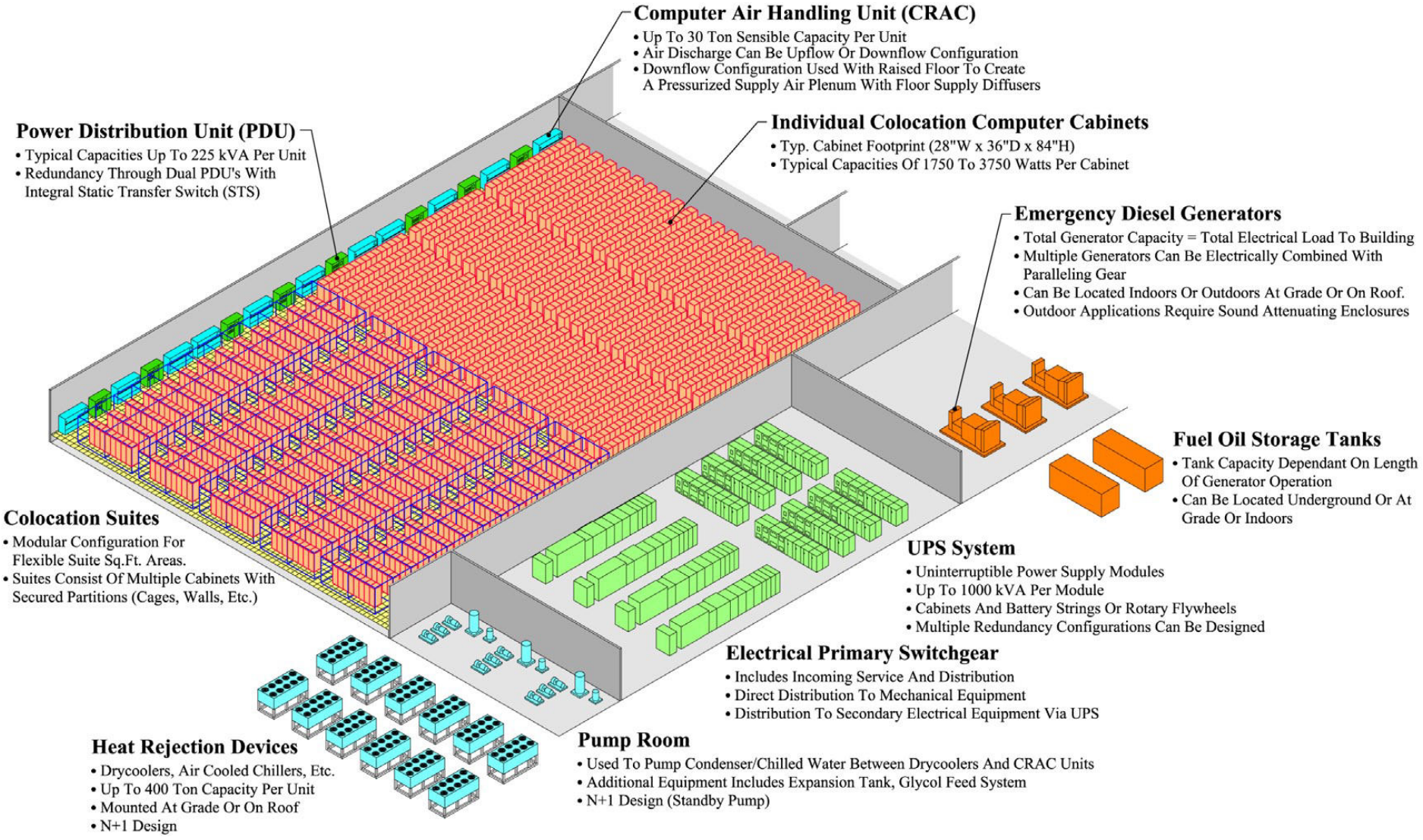


Inside a Data Center





A Datacenter Floor Plan



Energy Issue for Datacenters

- Data centers require tremendous power for daily operations
  - From 2005 to 2010, the electricity demand of worldwide datacenters increased by about **56%**.[1]
  - Total power requirement is around **38 GW** in 2012. A **17%** projected increase in 2013.[2]
  - The energy consumption of datacenters costs approximately **$27 billion** a year, and it is anticipated to double by 2014. [3]

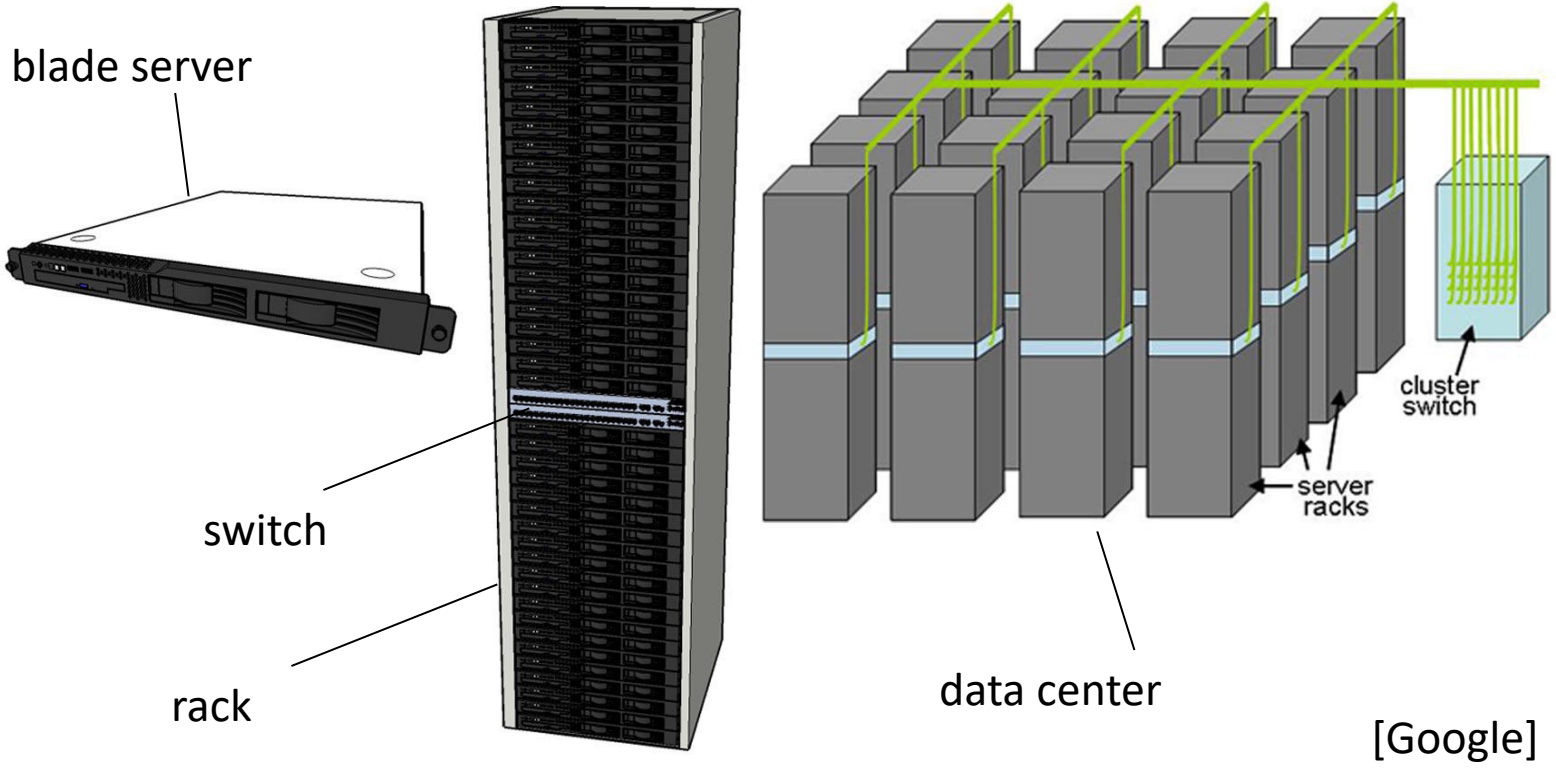
[1] J. Koomey, Growth in data center electricity use 2005 to 2010, Oakland, CA: Analytics Press, 2011.

[2] DCD 2012 Global Census

[3] N. Knupffer, The efficient datacenter, 2011. http://blogs.intel.com/technology/2011/10/the efficient datacenter - not/

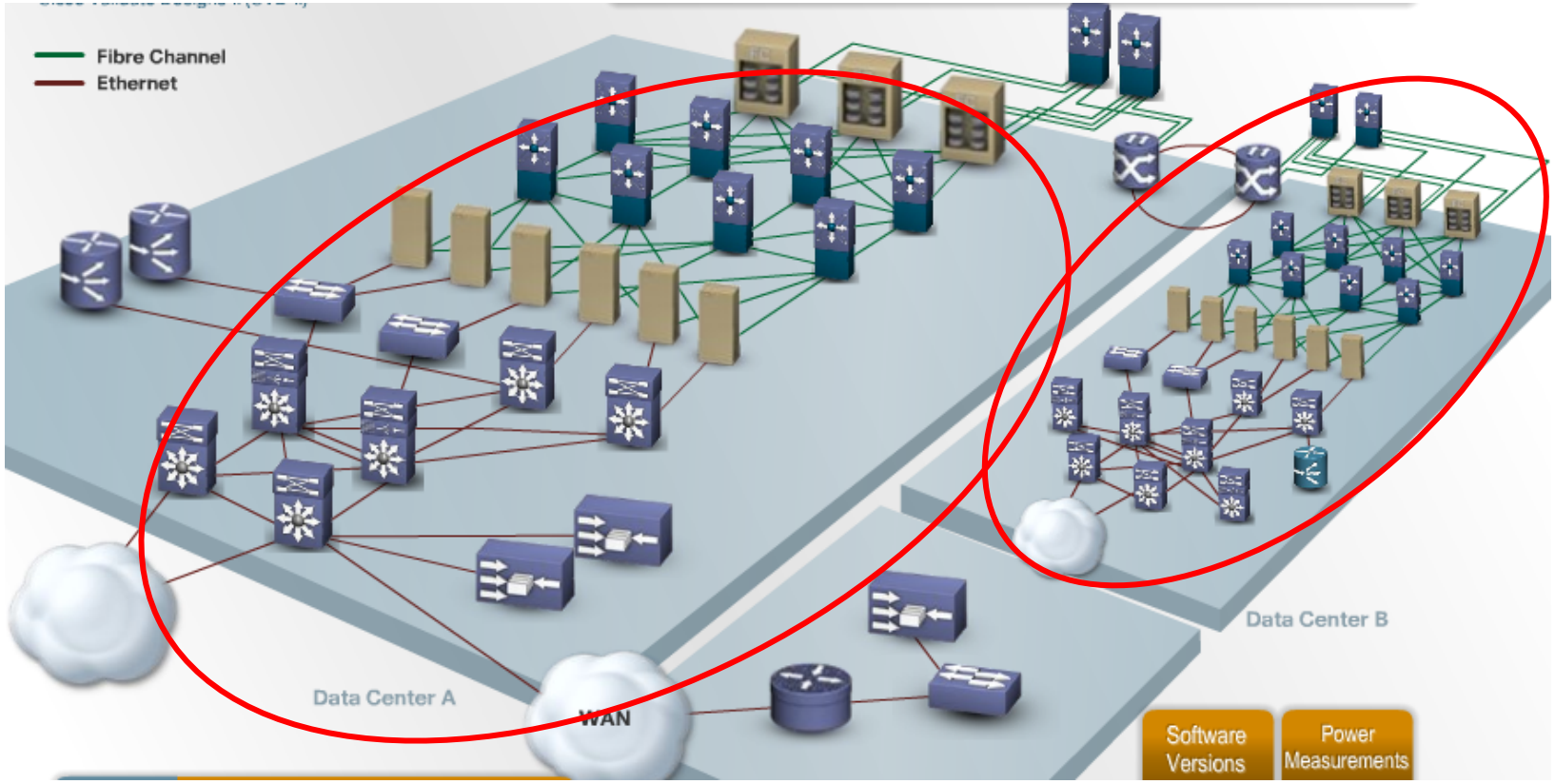
James Glanz of The New York Times and Ken Brill, an expert in the field, tour a data center and show what is required to keep it going even when power goes out. (https://www.nytimes.com/video/technology/100000001766676/what-keeps-a-data-center-going.html?action=click&contentCollection=technology&module=embedded&region=caption&pgtype=article)

Data Center Elements



[Google]

A Typical Design



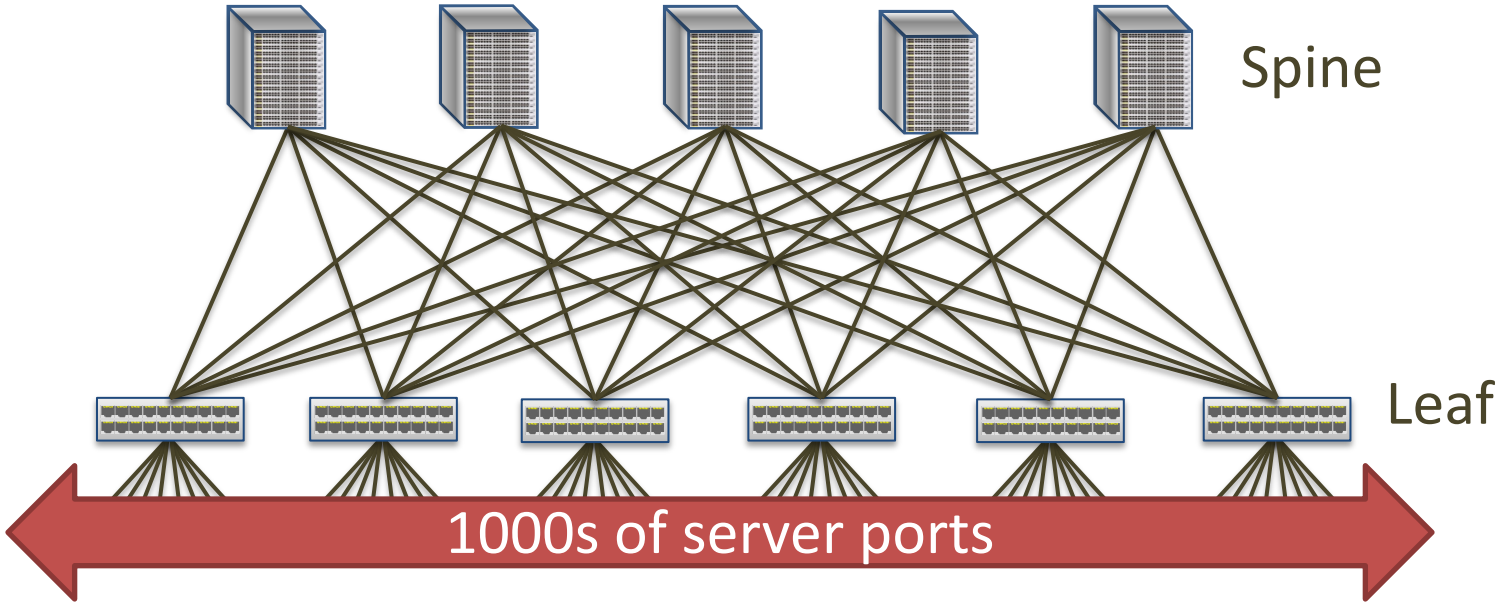
[Cisco]

Datacenter (DC) Network

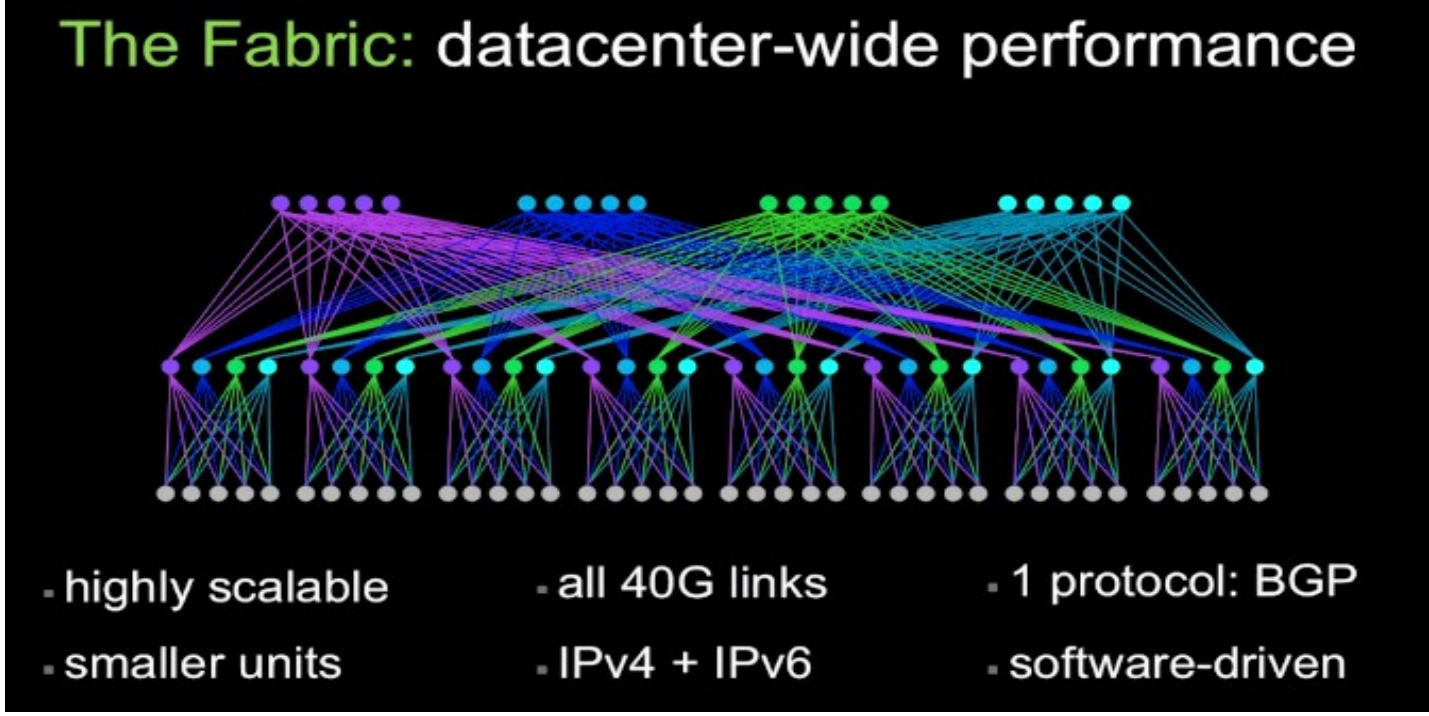
DC networks need large bisection bandwidth for distributed apps (big data, HPC, web services, etc)

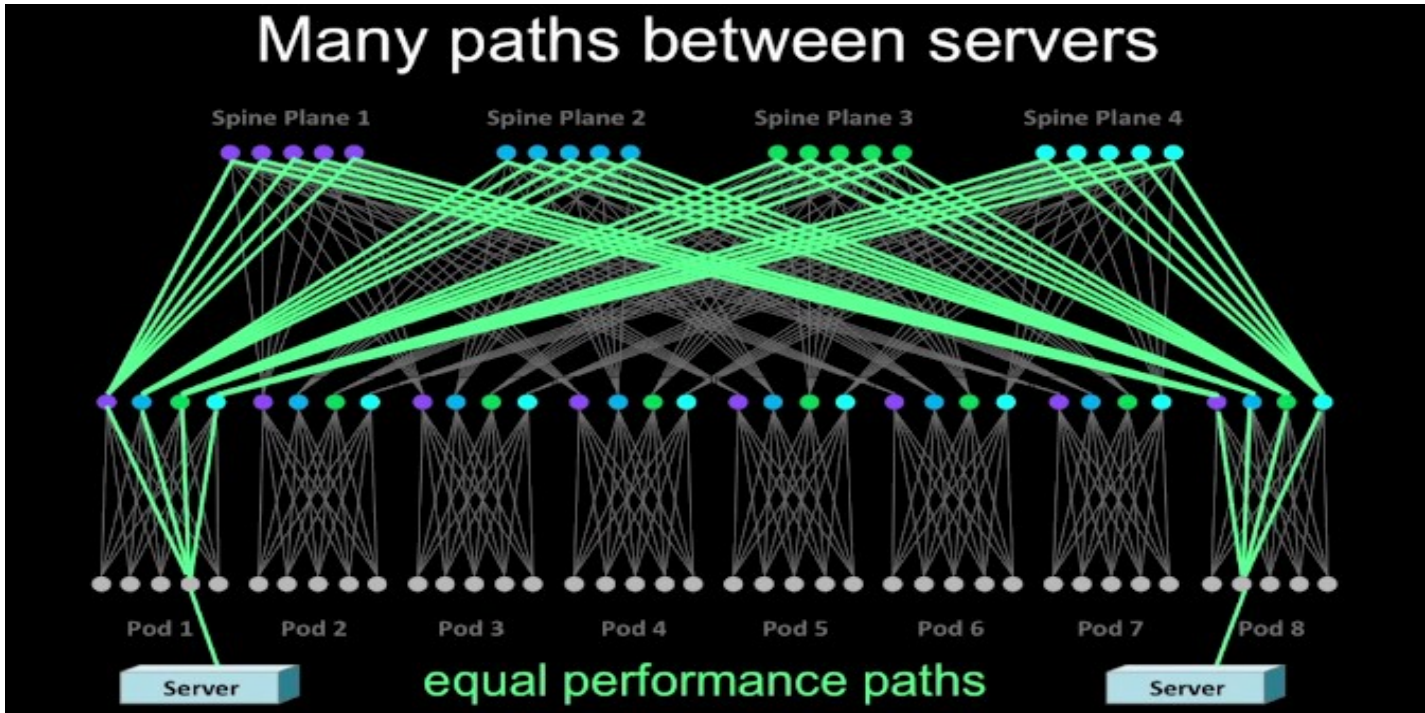
Multi-rooted tree [Fat-tree, Leaf-Spine, …]

- Full bisection bandwidth, achieved via multipathing



Facebook's Datacenter





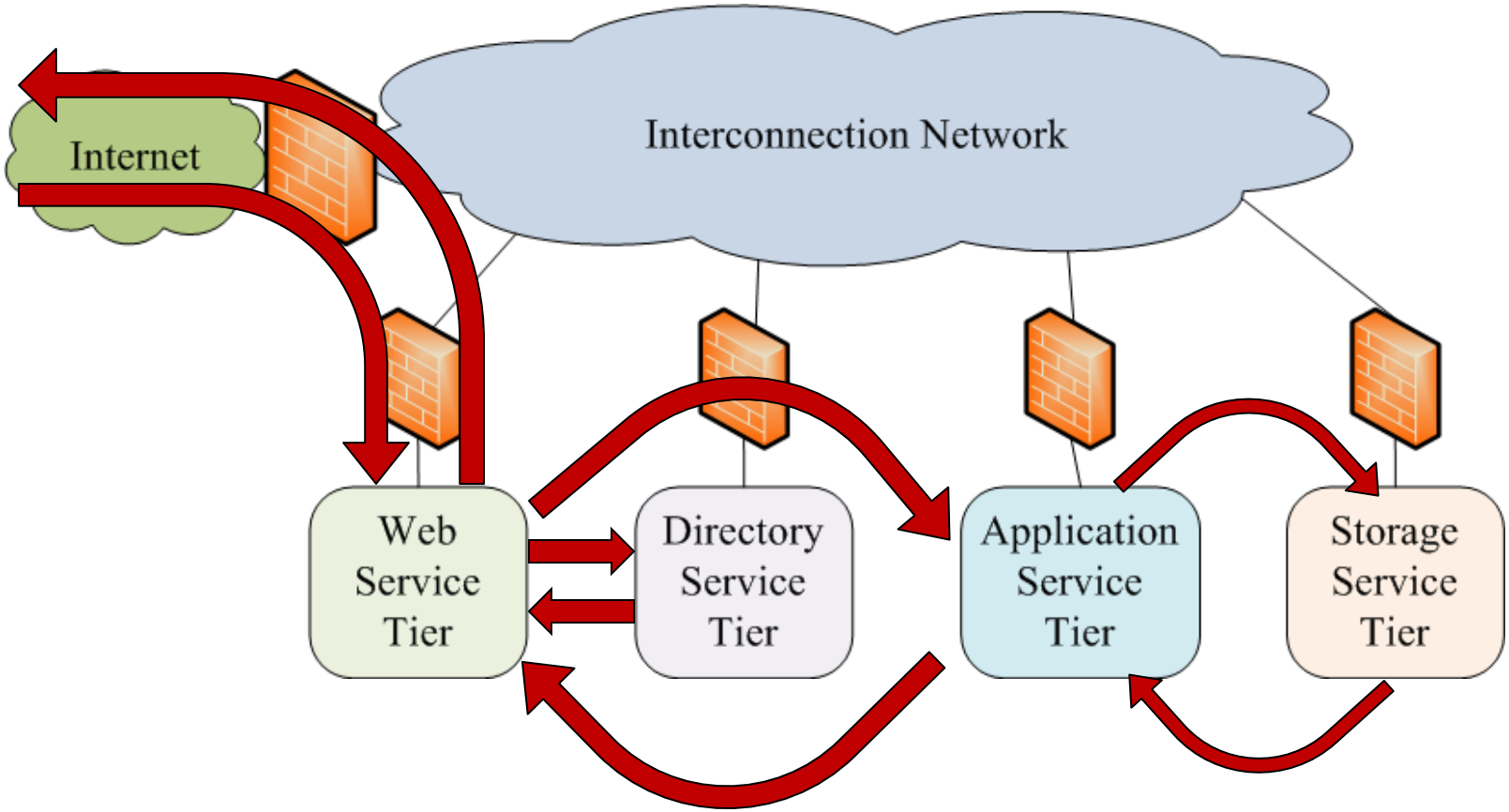
Modular Data Center

- A shipping container
- Up to 8 racks

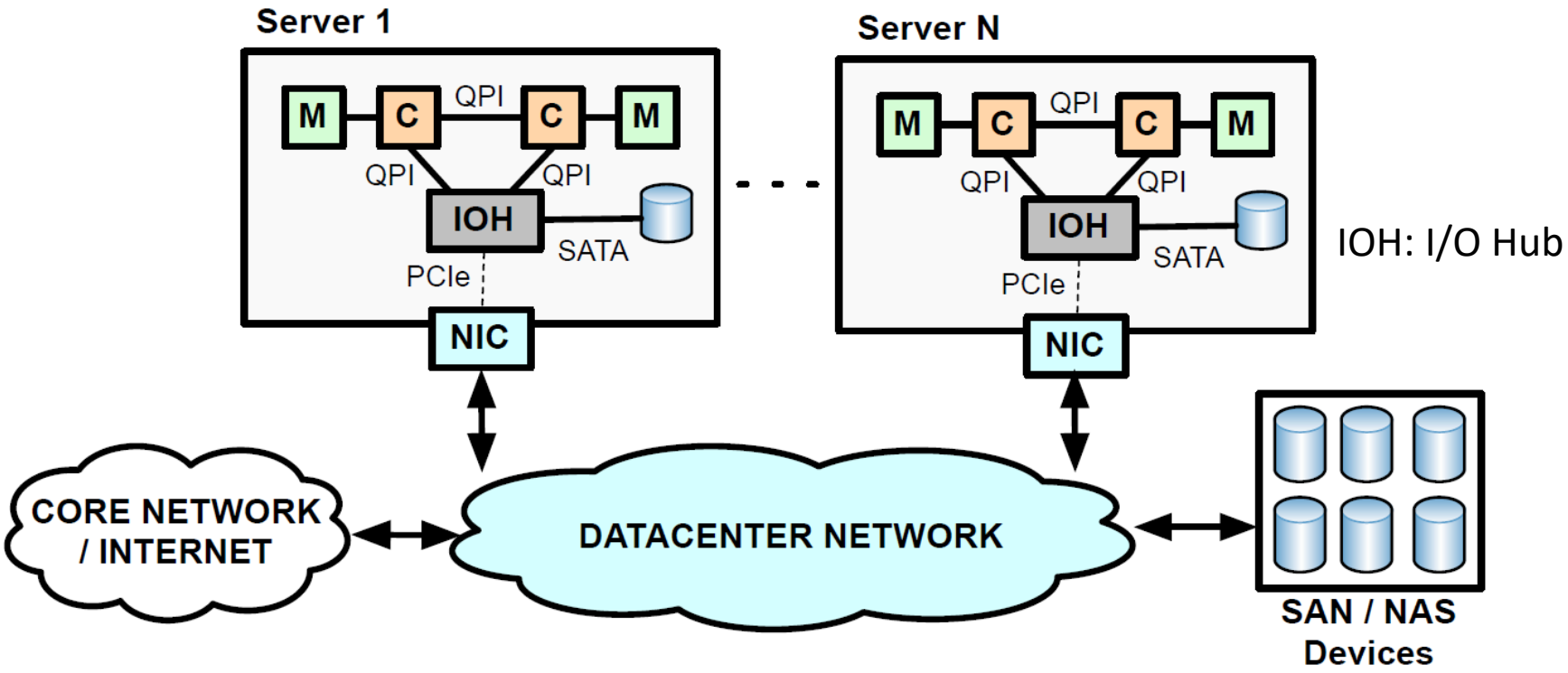
- Up to 280 servers
- Max power 200kW

Container-based Data Center

Logical View of A Data Center

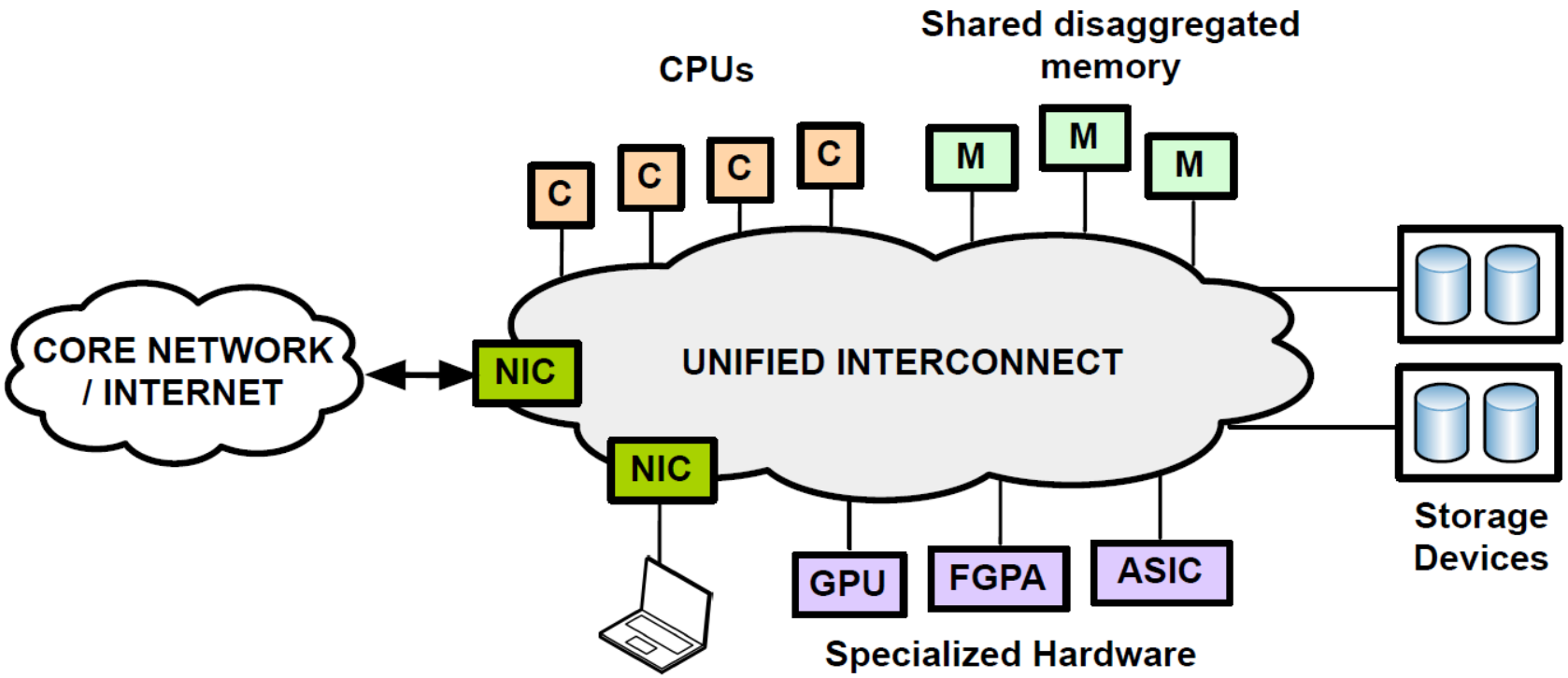


Server-Centric Datacenter



[Hotnets2013_ddc_slides]

Proposed Disaggregated Datacenters



[Hotnets2013_ddc_slides]

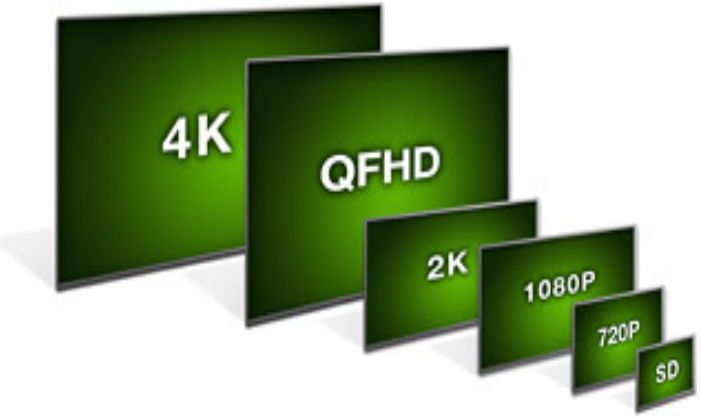
Challenging Issues in Building Data Centers

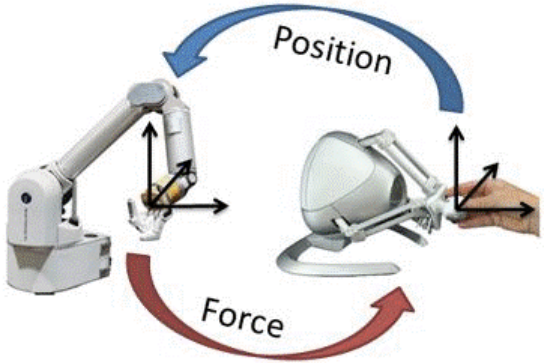
- Scale
  - Google: 0 to 1B users in ~15 years
  - Facebook: 0 to 1B users in ~10 years
  - *Must operate at the scale of O(1M+) users*
- Cost:
  - To build: Google ($3B/year), MSFT ($15B/total)
  - To operate: 1-2% of global energy consumption*
  - *Must deliver apps using efficient HW/SW footprint*
- Achieve high throughput and low latency for tasks by scheduling flows intelligently without causing data center network congestion
- Efficiently and effectively manage massive resources (compute/network), e.g., 100,000 servers or 1M virtual machines and communicating with each other at 100 Gbit/s, to accommodate frequent requests per unit time without human intervention and without any mistakes

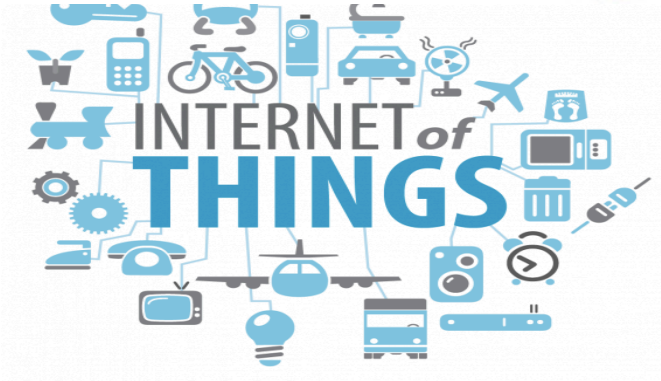
* LBNL, 2013.

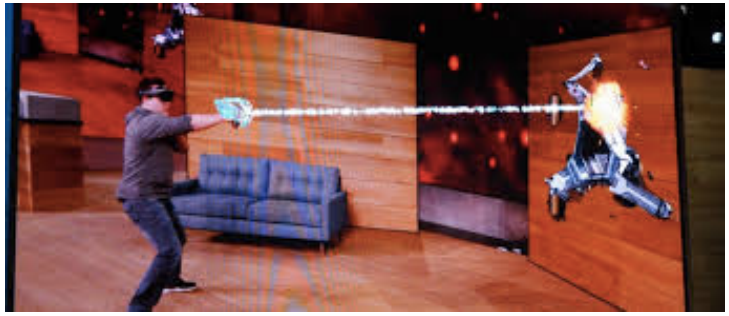
8. Edge Computing

New Applications Emerge









- Higher data capacity
- Stringent end to end latency
- Massive wireless device connections

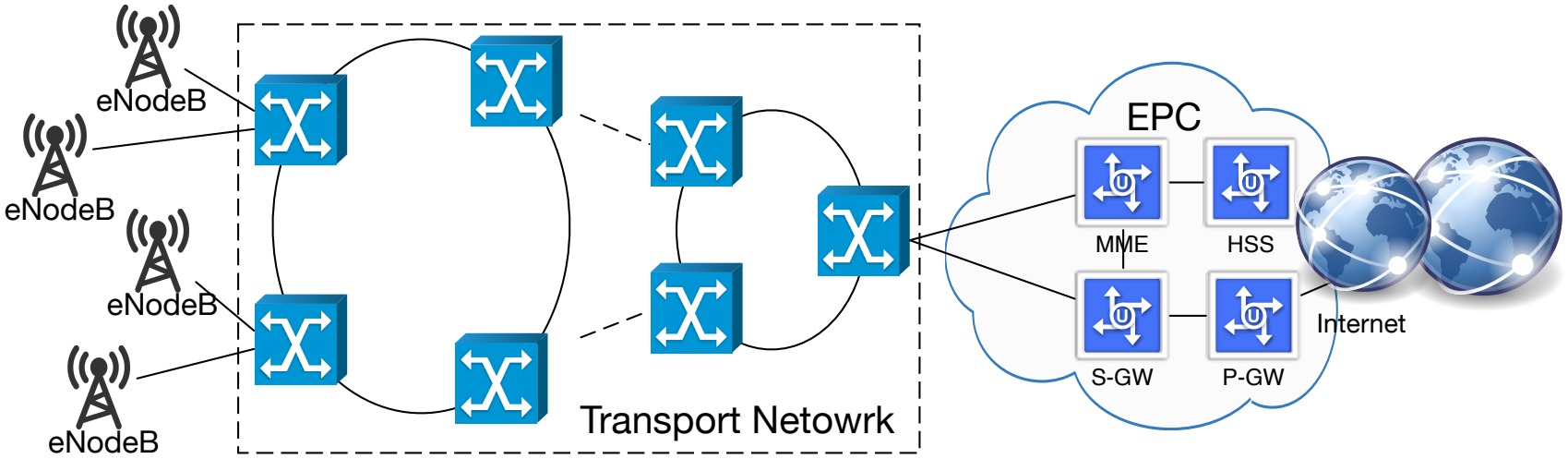
5G Network Requirements

- **Massive system capacity**
  - Support for connecting 20 million UE (user equipment) and 1 billion IoT (Internet of Things) devices and M2M (machine-to-machine) devices, and future V2V (Vehicle to Vehicle)
  - Lower energy per bit delivered
- **High data rates**
  - 1+ Gbps to UE
- **Low Latency**
  - End to end latency a few ms to support real time services
- **Ultra high reliability and availability**
  - Loss of connectivity and deviation for some applications must be extremely low

4G Cellular Network Architecture

EPC: Evolved Packet Core
MME: Mobility Management Entity
HSS: Home Subscriber Server

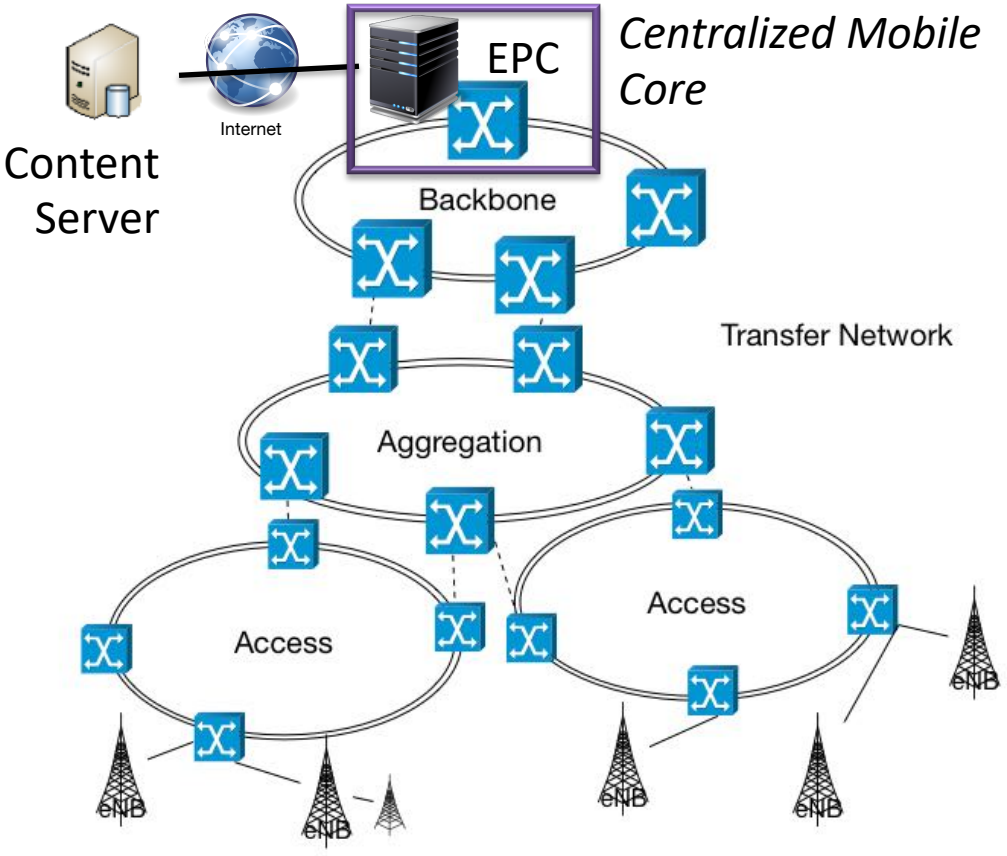
S-GW: Serving Gateway
P-GW: PDN Gateway
PDN: Packet Data Network



All Traffic Aggregated at EPC

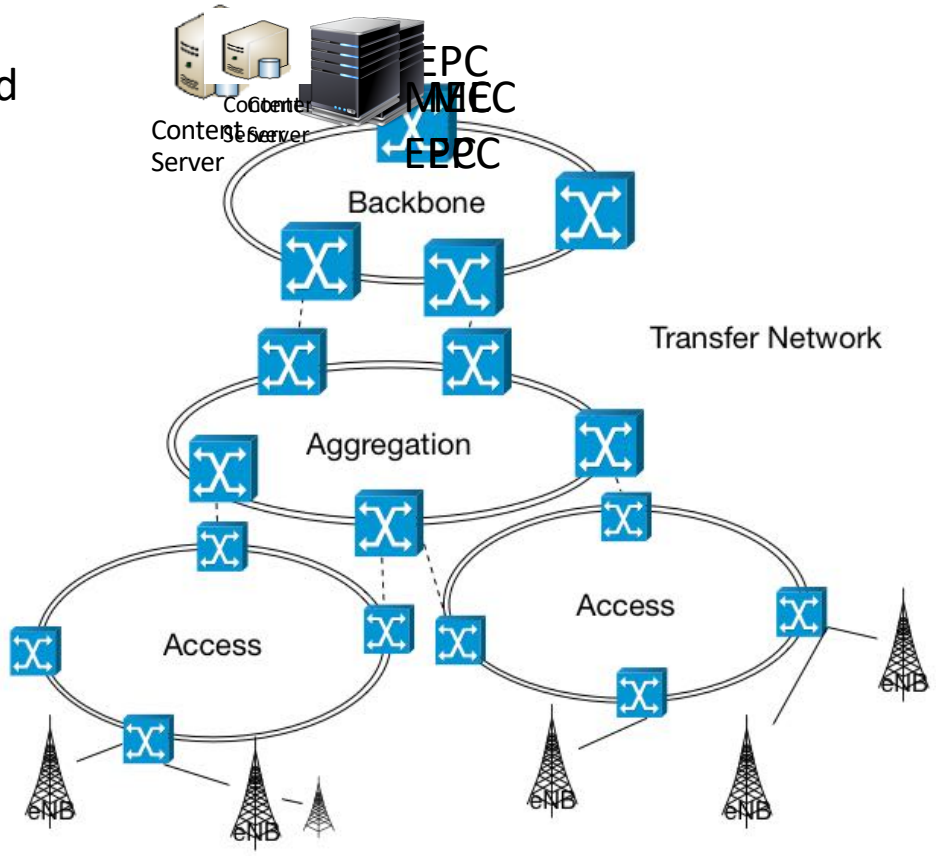
High Latency Low Data Rate

Mobile Network Architecture

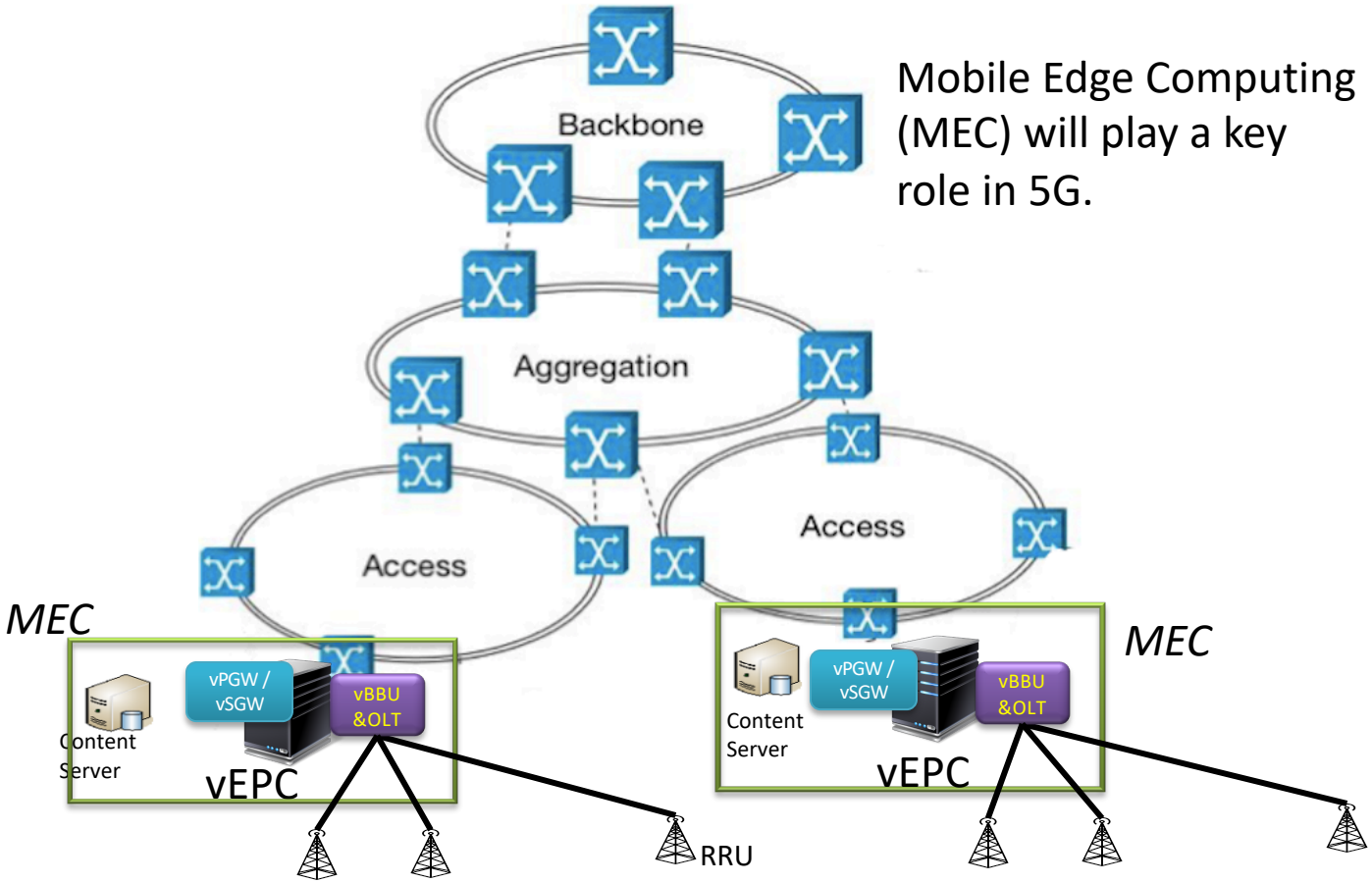


Future Mobile Network Architecture

Distribute EPC and content service towards the edge

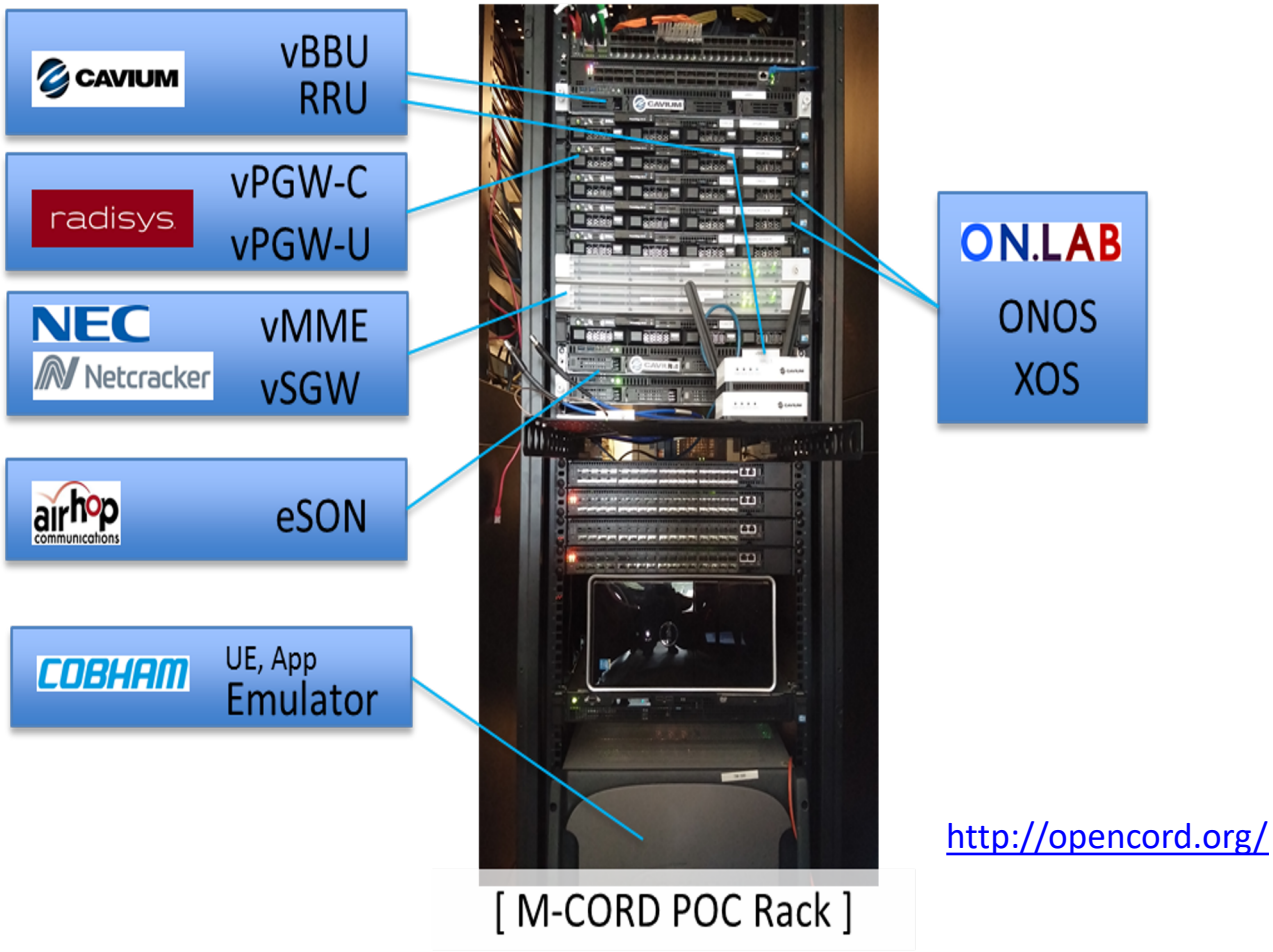


Future Mobile Network Architecture



Mobile Edge Computing (MEC) will play a key role in 5G.

Mobile Edge Computing Facility



http://opencord.org/

Issues of Distributed MECs

1. Many distributed MECs (e.g. 500 in Manhattan)
2. Coordinated resource management
3. Handover between distributed PGW and between content servers
4. Performance acceleration of vEPC and many other vNF
5. Reliability of distributed MECs
6. Security vulnerability to attacks