Introduction to Edge Data Center for IoT platform

Dr. Rajiv Misra, Professor
Dept. of Computer Science & Engg.
Indian Institute of Technology Patna
rajivm@iitp.ac.in

Preface

Content of this Lecture:

- Current demand of Data centers
- Why to move Data centers to Edge?
- In this lecture, we will discuss a brief introduction to Cloud Computing and also focus on the aspects *i.e*. Why Clouds, What is a Cloud, Whats new in todays Clouds and also distinguish Cloud Computing from the previous generation of distributed systems

Waves of Innovation: Cloud IoT Edge ML

Cloud

(the waves of innovation started with cloud)
Globally available, unlimited compute resources

IoT

(IoT-as-SaaS platform is key drivers of public cloud)
Harnessing signals from sensors and devices, managed centrally by the cloud

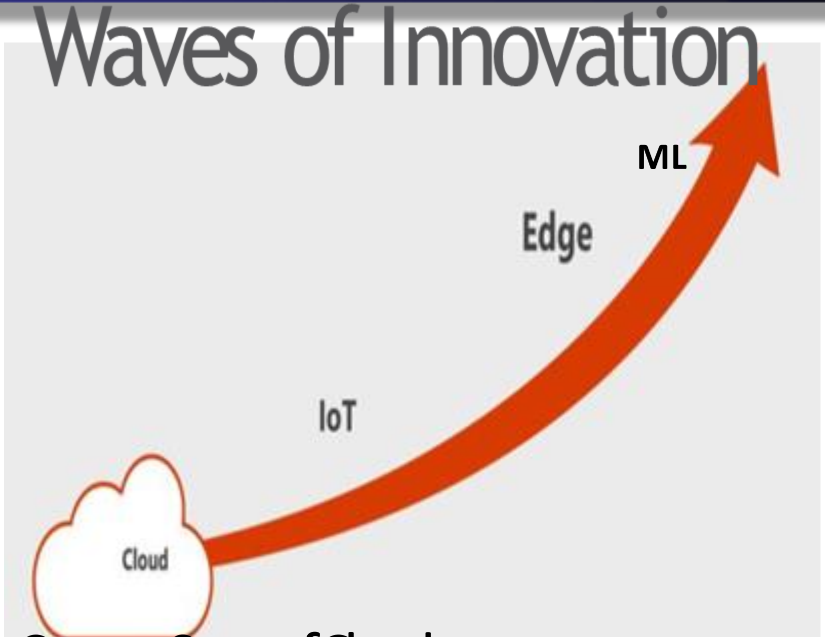
Edge

(IoT realize not everything needs to be in the cloud)
Intelligence offloaded from the cloud to IoT devices

ML

(rise of AI, ML models are trained in cloud are deployed at the edge to make inferencing for predictive analytics)

Breakthrough intelligence capabilities, in the cloud and on the edge



Current State of Cloud

- Highly centralized set of resources,
- Resembles Client/Server computing
- Compute is going beyond VMs as Containers becoming mainstream
- Storage is complemented by CDN is replicated and cached at edge locations
- Network stack is programmable SDN enabling hybrid scenarios

Edge Computing

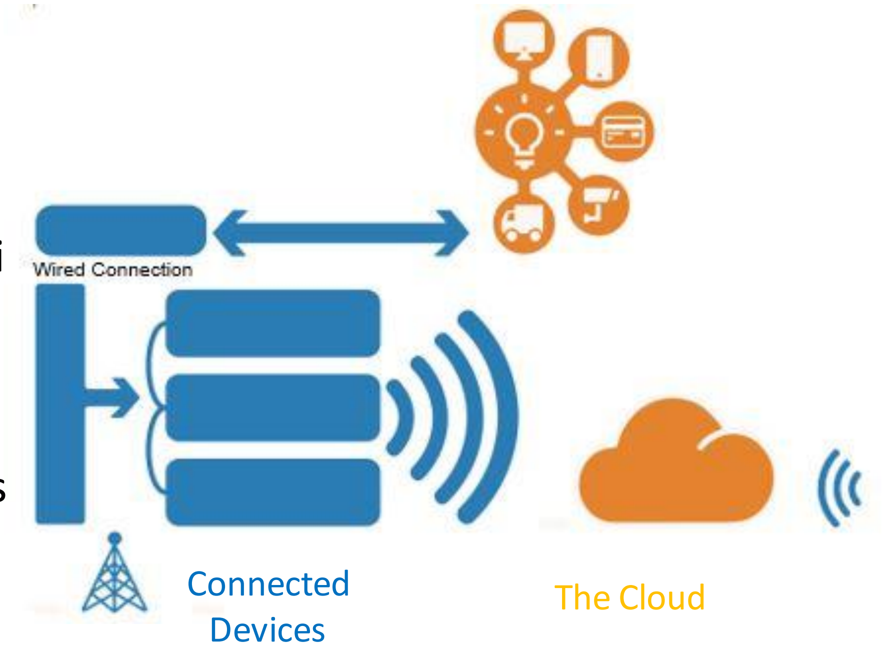
- Edge computing makes the cloud truly distributed
- Moves core cloud services closer to the origin of data
- Edge Mimics public cloud platform capabilities
- Delivers storage, compute, and network services locally.
- Reduces the latency by avoiding the roundtrip to the cloud
- Brings in data sovereignty by keeping data where it actually belongs, savings on cloud and bandwidth usages

Functionality of Edge Computing for IOT

- Data Ingestion and M2M Brokers
- Object Storage
- Functions as a Service
- Containers
- Distributed Computing
- NoSQL/Time-Series Database
- Stream Processing
- ML Models

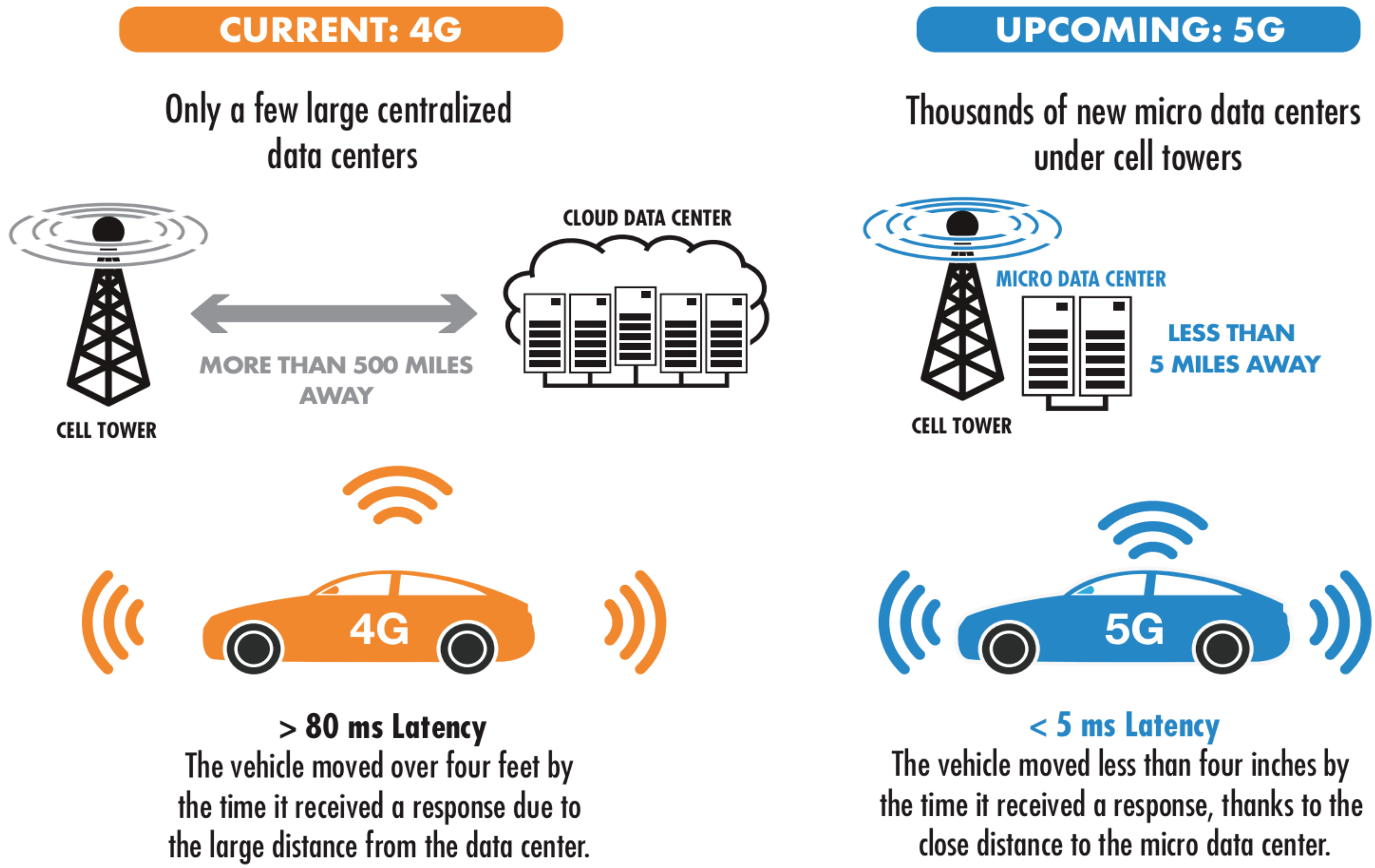
Cloud Data Center: Current Demand

- In the next decade, we will continue to see skyrocketing growth in the number of IP-connected mobile and machine-to-machine (M2M) devices, which will handle significant amounts of IP traffic.
- Tomorrow's consumers will demand faster Wi-Fi service and application delivery from online providers. Also, some M2M devices, such as autonomous vehicles, will require real-time communications with local processing resources to guarantee safety.



- Today's IP networks cannot handle the high-speed data transmissions that tomorrow's connected devices will require. In a traditional IP architecture, data must often travel hundreds of miles over a network between end users or devices and cloud resources. This results in latency, or slow delivery of time-sensitive data.

Cloud Data Center: Current Demand



CURRENT: 4G

Only a few large centralized data centers

> 80 ms Latency

The vehicle moved over four feet by the time it received a response due to the large distance from the data center.

UPCOMING: 5G

Thousands of new micro data centers under cell towers

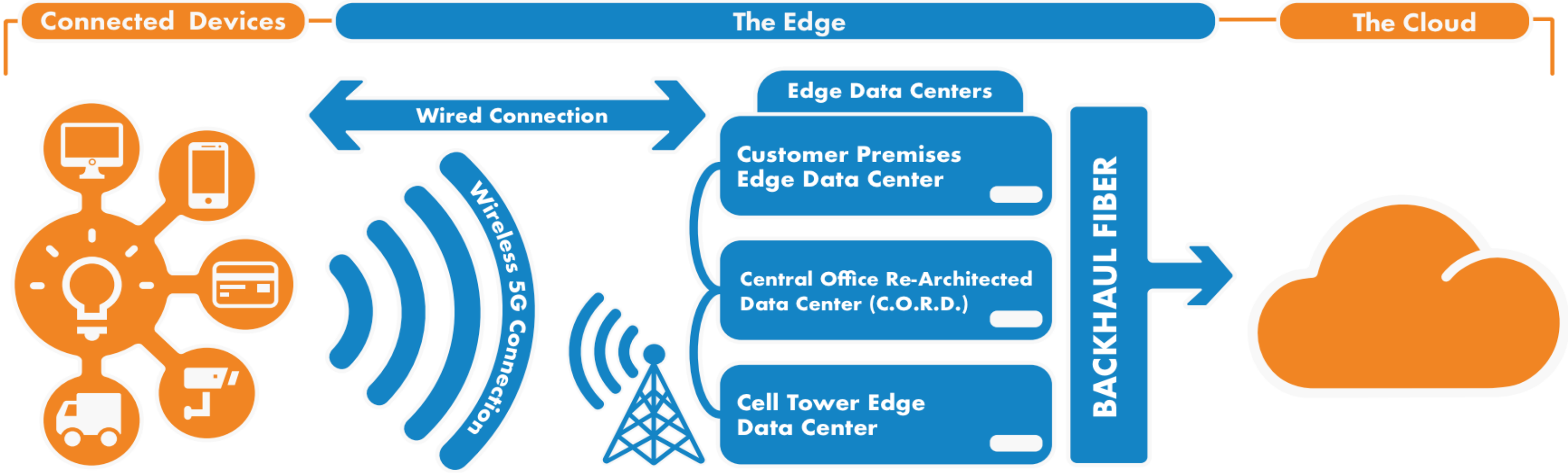
< 5 ms Latency

The vehicle moved less than four inches by the time it received a response, thanks to the close distance to the micro data center.

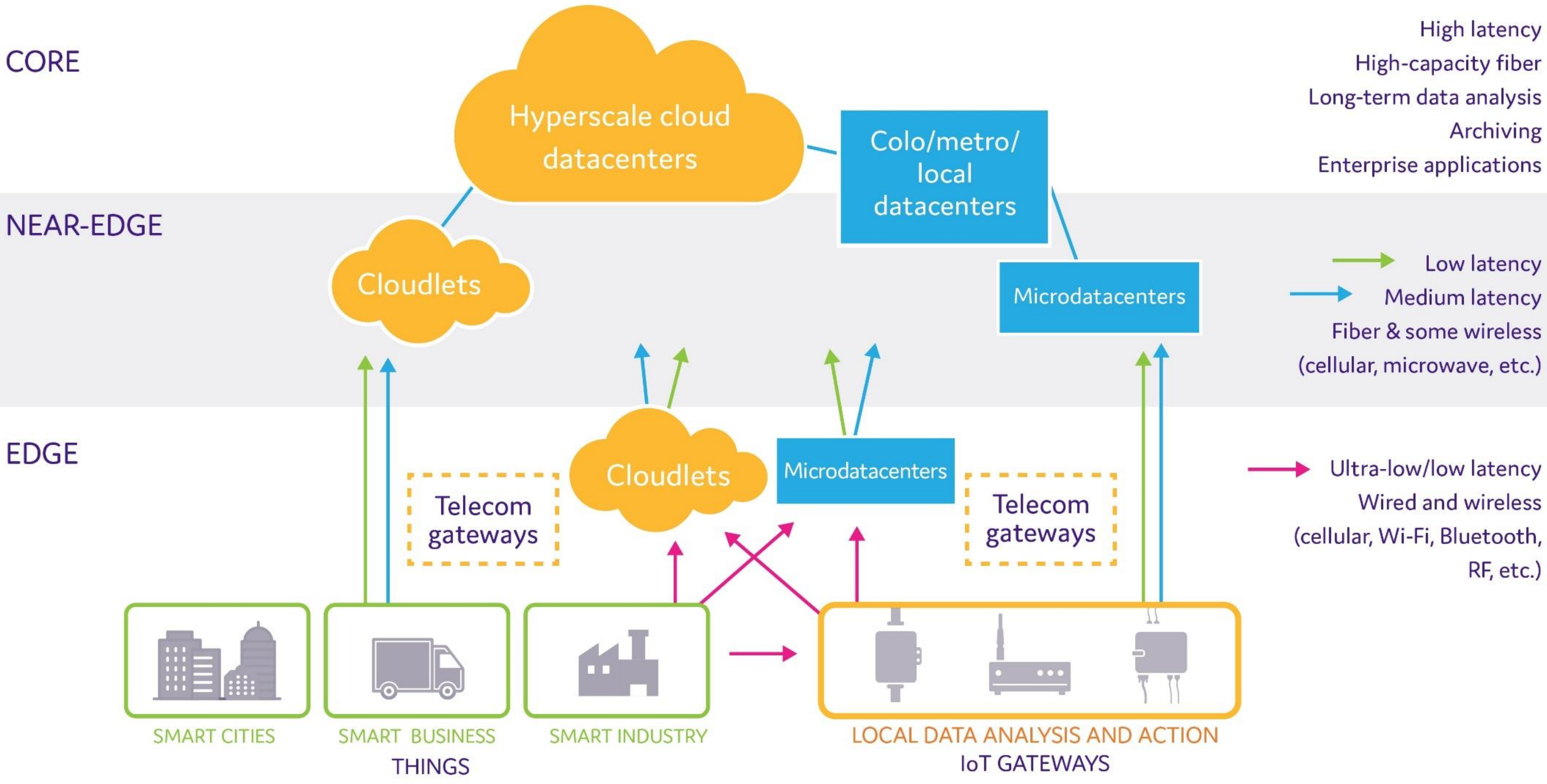
Edge Data Center: Solution

The solution to reducing latency lies in edge computing. By establishing IT deployments for cloud-based services in edge data centers in localized areas, we effectively bring IT resources closer to end users and devices. This helps us achieve efficient, high-speed delivery of applications and data. Edge data centers are typically located on the edge of a network, with connections back to a centralized cloud core.

Instead of bringing the users and devices to the data center, we bring the power of the data center to the users and devices. Edge computing relies on a distributed data center architecture, in which IT cloud servers housed in edge data centers are deployed on the outer edges of a network. By bringing IT resources closer to the end users and/or devices they serve, we can achieve high-speed, low-latency processing of applications and data.


Connected Devices
The Edge
The Cloud
Wired Connection
Wireless 5G Connection
Edge Data Centers
Customer Premises Edge Data Center
Central Office Re-Architected Data Center (C.O.R.D.)
Cell Tower Edge Data Center
BACKHAUL FIBER

Edge Data Center: Solution



CORE

NEAR-EDGE

EDGE

Hyperscale cloud datacenters

Colo/metro/ local datacenters

Cloudlets

Microdatacenters

Cloudlets

Microdatacenters

Telecom gateways

Telecom gateways

SMART CITIES

SMART BUSINESS

THINGS

SMART INDUSTRY

LOCAL DATA ANALYSIS AND ACTION

IoT GATEWAYS

High latency
High-capacity fiber
Long-term data analysis
Archiving
Enterprise applications

→ Low latency
→ Medium latency
Fiber & some wireless
(cellular, microwave, etc.)

→ Ultra-low/low latency
Wired and wireless
(cellular, Wi-Fi, Bluetooth, RF, etc.)

Why Move Data Centers to the Edge?

There are four main benefits of moving data centers to the edge, which involve improvements to latency, bandwidth, operating costs, and security:

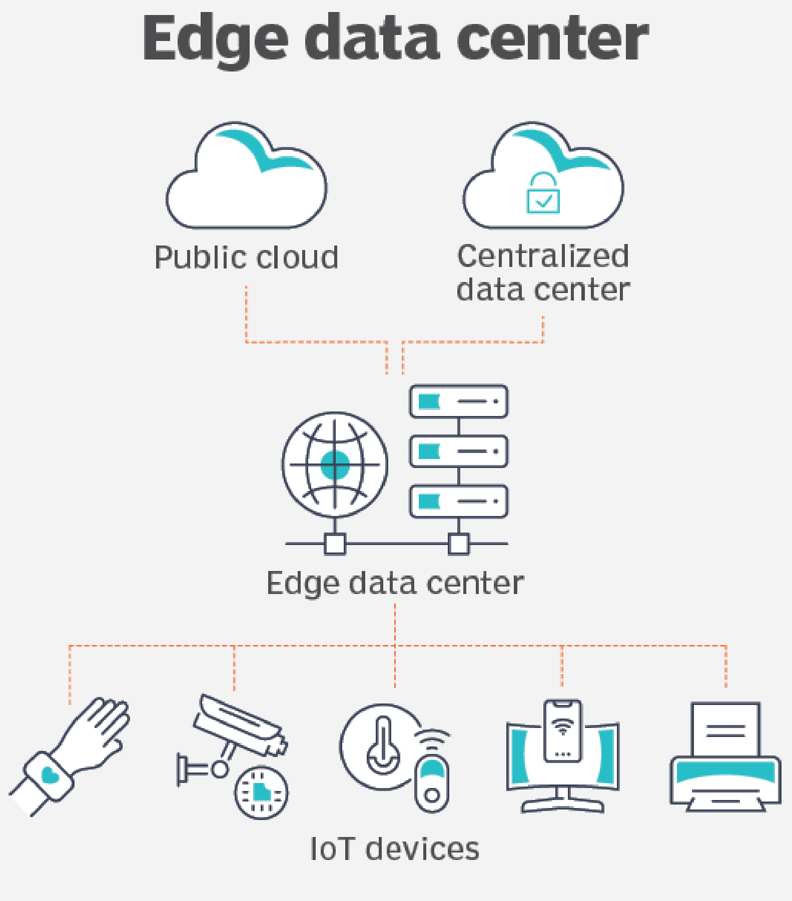
1. **Latency:** edge data centers facilitate lower latency, meaning much faster response times. Locating compute and storage functions closer to end users reduces the physical distance that data packets need to traverse, as well as the number of network "hops" involved, which lowers the probability of hitting a transmission path where data flow is impaired
2. **Bandwidth:** edge data centers process data locally, reducing the volume of traffic flowing to and from central servers. In turn, greater bandwidth across the user's broader network becomes available, which improves overall network performance
3. **Operating Cost:** because edge data centers reduce the volume of traffic flowing to and from central servers, they inherently reduce the cost of data transmission and routing, which is important for high-bandwidth applications. More specifically, edge data centers lessen the number of necessary high-cost circuits and interconnection hubs leading back to regional or cloud data centers, by moving compute and storage closer to end users
4. **Security:** edge data centers enhance security by: i) reducing the amount of sensitive data transmitted, ii) limiting the amount of data stored in any individual location, given their decentralized architecture, and iii) decreasing broader network vulnerabilities, because breaches can be ring-fenced to the portion of the network that they compromise

Edge Data Center: Introduction

Edge data centers are small data centers that are located close to the edge of a network. They provide the same devices found in traditional data centers, but are contained in a smaller footprint, closer to end users and devices.

Edge data centers can deliver cached content and cloud computing resources to these devices. The concept works off edge computing, which is a distributed IT architecture where client data is processed as close to the originating source as possible. Because the smaller data centers are positioned close to the end users, they are used to deliver fast services with minimal latency.

In an edge computing architecture, time-sensitive data may be processed at the point of origin by an intermediary server that is located in close geographical proximity to the client. The point is to provide the quickest content delivery to an end device that may need it, with as little latency as possible. Data that is less time-sensitive can be sent to a larger data center for historical analysis, big data analytics and long-term storage. Edge data centers work off of the same concept, except instead of just having one intermediary server in close geographical proximity to the client, it's a small data center -- that can be as small as a box. Even though it is not a new concept, edge data center is still a relatively new term.

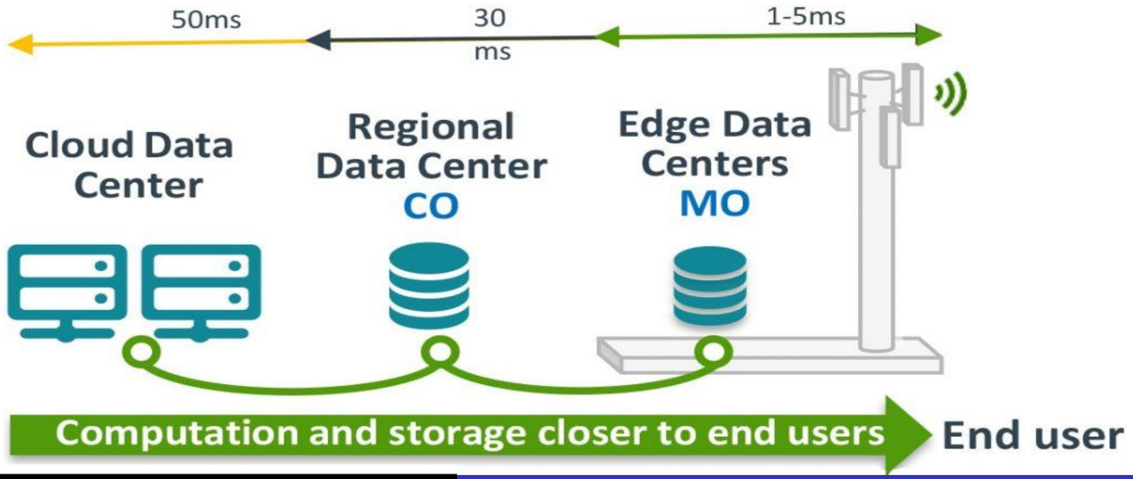


Edge Data Center: Introduction

The major benefit of an edge data center is the quick delivery of services with minimal latency, thanks to the use of edge caching. Latency may be a big issue for organizations that have to work with the internet of things (IoT), big data, cloud and streaming services.

Edge data centers can be used to provide high performance with low levels of latency to end users, making for a better user experience. Typically, edge data centers will connect to a larger, central data center or multiple other edge data centers.

Data is processed as close to the end user as possible, while less integral or time-centric data can be sent to a central data center for processing. This allows an organization to reduce latency.



Edge Data Center: Use Cases

1. **5G**: Where a decentralized cell network made of edge data centers can help provide low latency for 5G in use cases with high device density.

Telecommunications companies. With cell-tower edge data centers, telecom companies can get better proximity to end users by connecting mobile phones and wireless sensors.

2. **IoT**: Edge data centers can be used for data generated by IoT devices. An edge data center would be used if data generated by devices needs more processing but is also too time-sensitive to be sent to a centralized server.

3. **Healthcare**: Some medical equipment, such as those used for robotic surgeries, would require extremely low latency and network consistency, of which, edge data centers can provide.

4. **Autonomous vehicles**: Edge data centers can be used to help collect, process and share data between vehicles and other networks, which also relies on low latency. A network of edge data centers can be used to collect data for auto manufacturers and emergency response services.

5. **Smart factories**: Edge data centers can be used for machine Predictive maintenance, as well as predictive quality management. It can also be used for efficiency regarding robotics used within inventory management.

Scalable Computing at network edge

- Evolutionary changes that have occurred in **distributed edge and cloud computing** over the past 30 years, **driven by applications with variable workloads, low-latency usecase and large data sets .**
- Evolutionary changes in machine architecture, operating system platform, network connectivity, and application workload.
- **Edge computing** uses multiple computers at network edge to solve large-scale problems locally and over the Internet. Thus, **distributed edge computing becomes data-intensive and network-centric.**
- **The emergence of distributed edge computing clouds** instead demands high-throughput computing (HTC) systems built with distributed computing technologies.
- **High-throughput computing (HTC)** appearing as computer clusters, service-oriented, computational grids, peer-to-peer networks, Internet clouds and edge, and the future Internet of Things.

The Hype of Cloud: Forecasting

- Gartner in 2009 – Cloud computing revenue will soar faster than expected and will **exceed $150 billion** by 2013. It will represent 19% of IT spending by 2015.

- IDC in 2009: “Spending on IT cloud services will triple in the next 5 years, reaching **$42 billion**.”

- Forrester in 2010 – Cloud computing will go from **$40.7 billion** in 2010 to **$241 billion** in 2020.

- Companies and even federal/state governments using cloud computing now: **fbo.gov**

Many Cloud Providers

- AWS: Amazon Web Services
  - EC2: Elastic Compute Cloud
  - S3: Simple Storage Service
  - EBS: Elastic Block Storage
- Microsoft Azure
- Google Compute Engine/AppEngine
- Rightscale, Salesforce, EMC, Gigaspaces, 10gen, Datastax, Oracle, VMWare, Yahoo, Cloudera
- And 100s more…

Categories of Clouds

- Can be either a (i) public cloud, or (ii) private cloud
- **Private clouds** are accessible only to company employees
- **Public clouds** provide service to any paying customer:
  - **Amazon S3 (Simple Storage Service):** store arbitrary datasets, pay per GB-month stored
  - **Amazon EC2 (Elastic Compute Cloud):** upload and run arbitrary OS images, pay per CPU hour used
  - **Google App Engine/Compute Engine:** develop applications within their App Engine framework, upload data that will be imported into their format, and run

Customers Save: Time and Money

- “With AWS, a new server can be up and running in **three minutes** compared to **seven and a half weeks** to deploy a server internally and a **64-node Linux cluster** can be online in five minutes (compared with three months internally.”

- “With Online Services, reduce the IT **operational costs** by roughly **30%** of spending”

- “A private cloud of virtual servers inside its datacenter has saved nearly **crores of rupees annually**, because the company can share computing power and storage resources across servers.”

- 100s of startups can harness large computing resources without buying their own machines.

What is a Cloud?

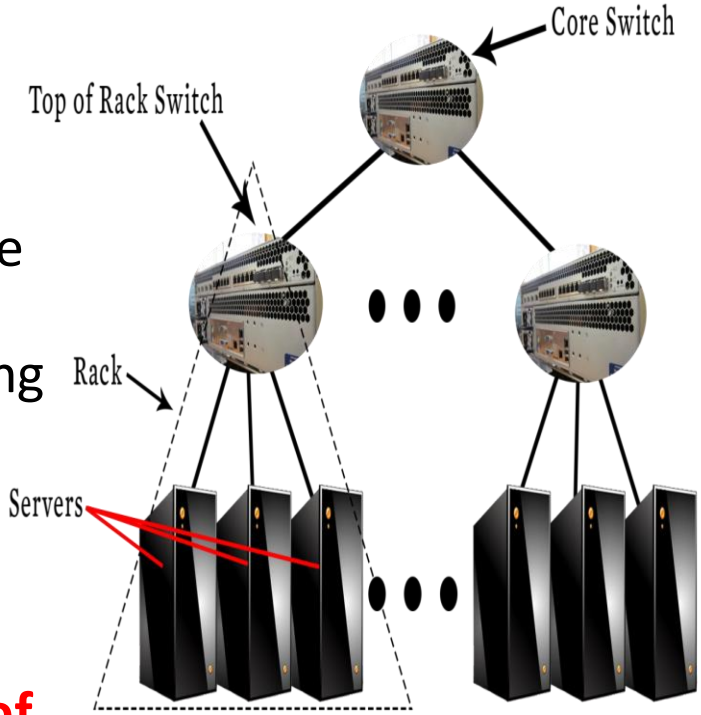
- Advances in virtualization make it possible to see the growth of Internet clouds **as a new computing paradigm**.
- i.e. dramatic differences between developing software for millions to use **as a service** versus distributing software to run on their PCs."

History:

- **In 1984, John Gage** Sun Microsystems gave the slogan, **"The network is the computer."**
- **In 2008, David Patterson** UC Berkeley said, **"The data center is the computer."**
- Recently, **Rajkumar Buyya** of Melbourne University simply said: **"The cloud is the computer."**
- Some people view **clouds as grids** or **clusters** with changes through virtualization, since clouds are anticipated to process huge data sets generated by the traditional Internet, social networks, and the future IoT.

What is a Cloud?

- **A single-site cloud (as known as "Datacenter") consists of**
  - Compute nodes (grouped into racks)
  - Switches, connecting the racks
  - A network topology, e.g., hierarchical
  - Storage (backend) nodes connected to the network
  - Front-end for submitting jobs and receiving client requests
  - (Often called "three-tier architecture")
  - Software Services

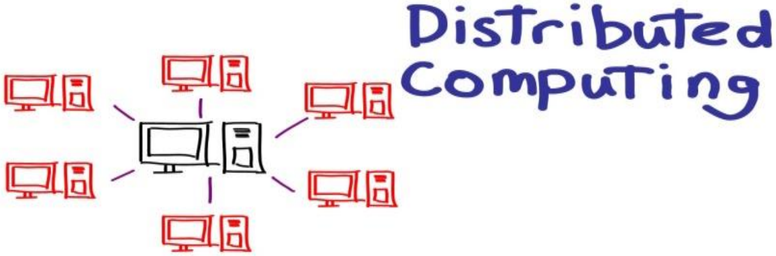


- **A geographically distributed cloud consists of**
  - Multiple such sites
  - Each site perhaps with a different structure and services

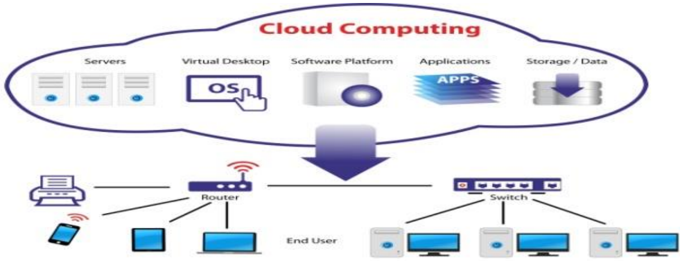
Computing Paradigm Distinctions

- Cloud computing overlaps with distributed computing.

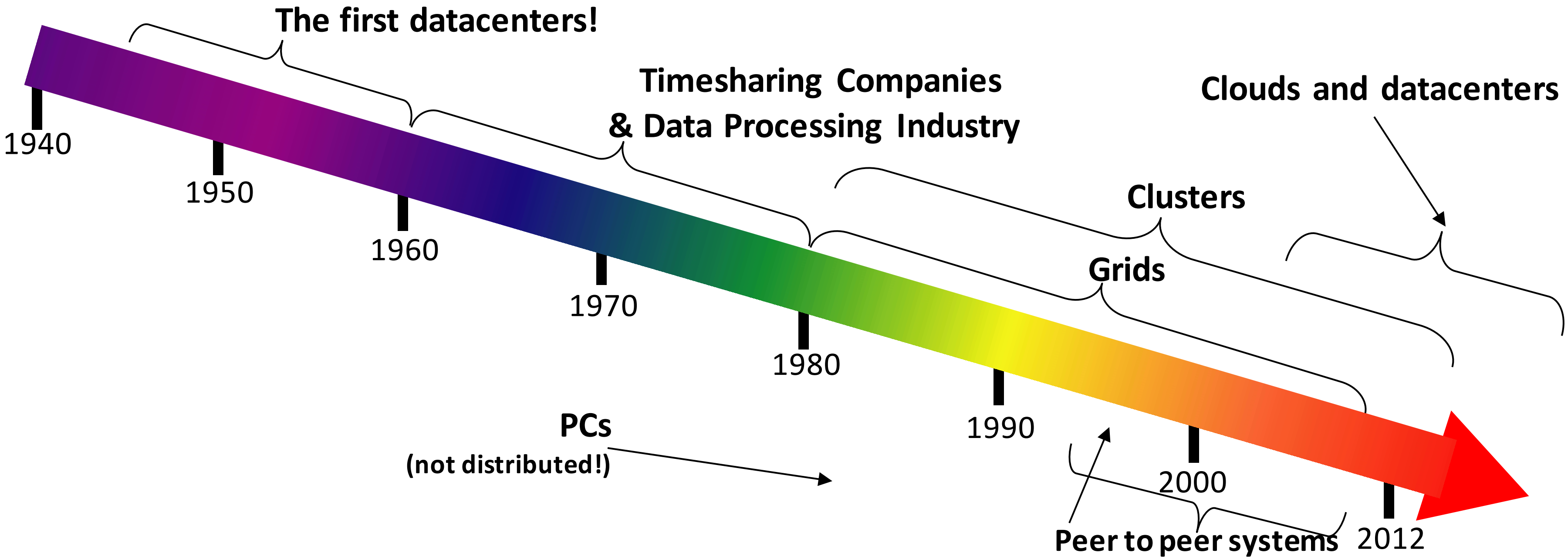
- **Distributed computing:** A ***distributed system*** consists of multiple autonomous computers, having its own memory, communicating through ***message passing***.



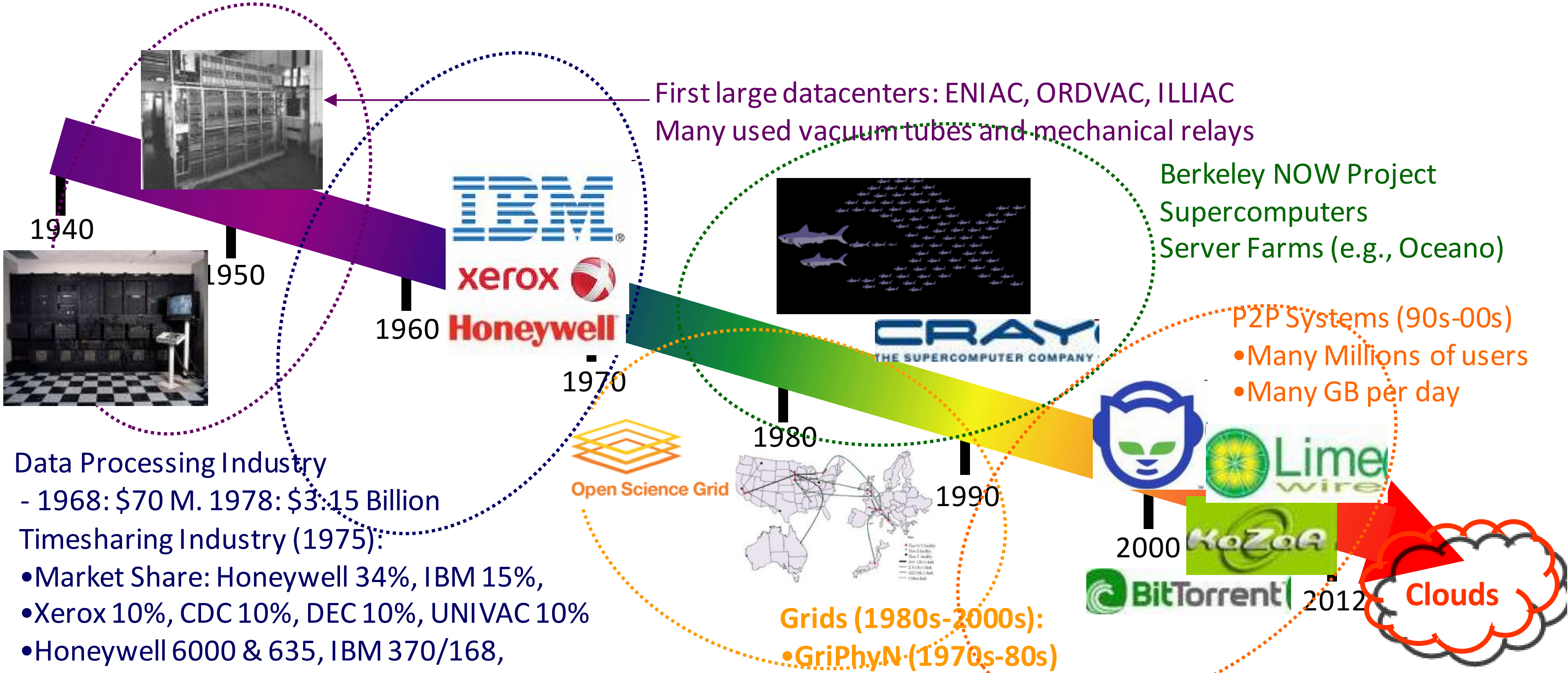
- **Cloud computing:** Clouds can be built with physical or virtualized resources over large data centers that are distributed systems. Cloud computing is also considered to be a form of ***utility computing or service computing.***



"A Cloudy History of Time"



“A Cloudy History of Time”



First large datacenters: ENIAC, ORDVAC, ILLIAC
Many used vacuum tubes and mechanical relays

Berkeley NOW Project
Supercomputers
Server Farms (e.g., Oceano)

P2P Systems (90s-00s)
- Many Millions of users
- Many GB per day

Data Processing Industry
- 1968: $70 M. 1978: $3.15 Billion

Timesharing Industry (1975):
- Market Share: Honeywell 34%, IBM 15%,
- Xerox 10%, CDC 10%, DEC 10%, UNIVAC 10%
- Honeywell 6000 & 635, IBM 370/168, Xerox 940 & Sigma 9, DEC PDP-10, UNIVAC 1108

Grids (1980s-2000s):
- **GriPhyN (1970s-80s)**
- **Open Science Grid and Lambda Rail (2000s)**
- **Globus & other standards (1990s-2000s)**

Scalable Computing Trends: Technology

- **Doubling Periods** – storage: 12 months, bandwidth: 9 months, and CPU compute capacity: 18 months (what law is this?)
- **Moore's law** indicates that processor speed doubles every 18 months.
- **Gilder's law** indicates that network bandwidth has doubled each year in the past.

- Then and Now
  - Bandwidth
    - 1985: mostly 56Kbps links nationwide
    - 2015: Tbps links widespread
  - Disk capacity
    - Today's PCs have TBs, far more than a 1990 supercomputer

The Trend toward Utility Computing

- Aiming towards autonomic operations that can be self-organized to support dynamic discovery. Major computing paradigms are composable with ***QoS and SLAs (service-level agreements).***
- In 1965, MIT's Fernando Corbató of the Multics operating system envisioned a computer facility operating “like a power company or water company”.
- **Plug** your thin client into the computing Utility **and Play** Intensive Compute & Communicate Application
- **Utility computing** focuses on a business model in which customers receive computing resources from a paid service provider.
- All **grid/cloud platforms are regarded as utility service providers**.

Features of Today's Clouds

I. **Massive scale:** Very large data centers, contain tens of thousands sometimes hundreds of thousands of servers and you can run your computation across as many servers as you want and as many servers as your application will scale.

II. **On-demand access:** Pay-as-you-go, no upfront commitment.

- And anyone can access it

III. **Data-intensive Nature:** What was MBs has now become TBs, PBs and XBs.

- Daily logs, forensics, Web data, etc.

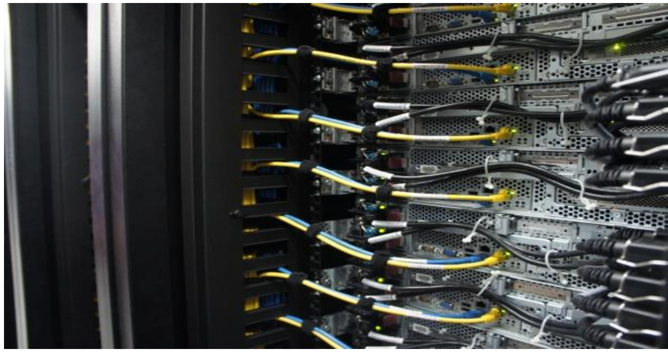
IV. **New Cloud Programming Paradigms:** MapReduce/Hadoop, NoSQL/Cassandra/MongoDB and many others.

- **Combination of one or more of these gives rise to novel and unsolved distributed computing problems in cloud computing.**

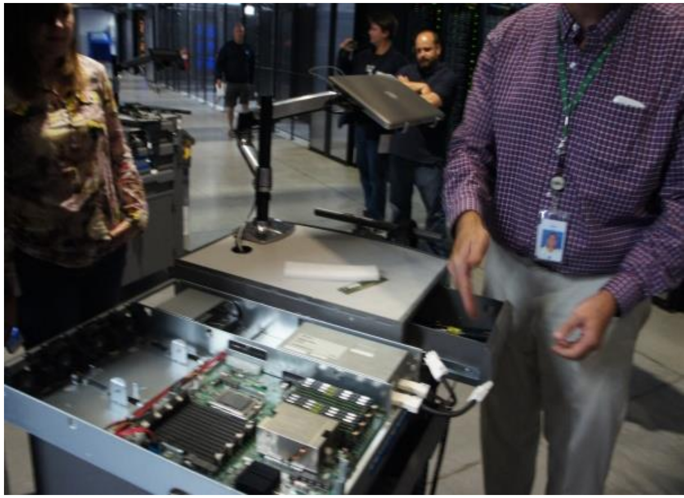
I. Massive Scale

- **Facebook [GigaOm, 2012]**
  - 30K in 2009 -> 60K in 2010 -> 180K in 2012
- **Microsoft [NYTimes, 2008]**
  - 150K machines
  - Growth rate of 10K per month
  - 80K total running Bing
  - In 2013, Microsoft Cosmos had 110K machines (4 sites)
- **Yahoo! [2009]:**
  - 100K
  - Split into clusters of 4000
- **AWS EC2 [Randy Bias, 2009]**
  - 40K machines
  - 8 cores/machine
- **eBay [2012]: 50K machines**
- **HP [2012]: 380K in 180 DCs**
- **Google: A lot**

What does a datacenter look like from inside?



Lots of Servers



Power and Energy

- WUE = Annual Water Usage / IT Equipment Energy (L/kWh)
  – low is good
- PUE = Total facility Power / IT Equipment Power
  – low is good (e.g., Google~1.11)

Off-site

On-site

Cooling

- Air sucked in
- Combined with purified water
- Moves cool air through system

II. On-demand access: *AAS Classification

- **On-demand:** renting vs. buying one. E.g.:
  - AWS Elastic Compute Cloud (EC2): a few cents to a few $ per CPU hour
  - AWS Simple Storage Service (S3): a few cents per GB-month
- **HaaS: Hardware as a Service**
  - Get access to barebones hardware machines, do whatever you want with them, Ex: Your own cluster
  - Not always a good idea because of security risks
- **IaaS: Infrastructure as a Service**
  - Get access to flexible computing and storage infrastructure. **Virtualization** is one way of achieving this. subsume HaaS.
  - Ex: Amazon Web Services (AWS: EC2 and S3), OpenStack, Eucalyptus, Rightscale, Microsoft Azure, Google Cloud.

II. On-demand access: *AAS Classification

- **PaaS: Platform as a Service**
  - Get access to flexible computing and storage infrastructure, coupled with a software platform (often tightly coupled)
  - Ex: Google's AppEngine (Python, Java, Go)

- **SaaS: Software as a Service**
  - Get access to software services, when you need them. subsume SOA (Service Oriented Architectures).
  - Ex: Google docs, MS Office on demand

III. Data-intensive Computing

- **Computation-Intensive Computing**
  - Example areas: MPI-based, High-performance computing, Grids
  - Typically run on supercomputers (e.g., NCSA Blue Waters)
- **Data-Intensive**
  - Typically store data at datacenters
  - Use compute nodes nearby
  - Compute nodes run computation services

- In data-intensive computing, the **focus shifts from computation to the data**:
- CPU utilization no longer the most important resource metric, instead I/O is (disk and/or network)

IV. New Cloud Programming Paradigms

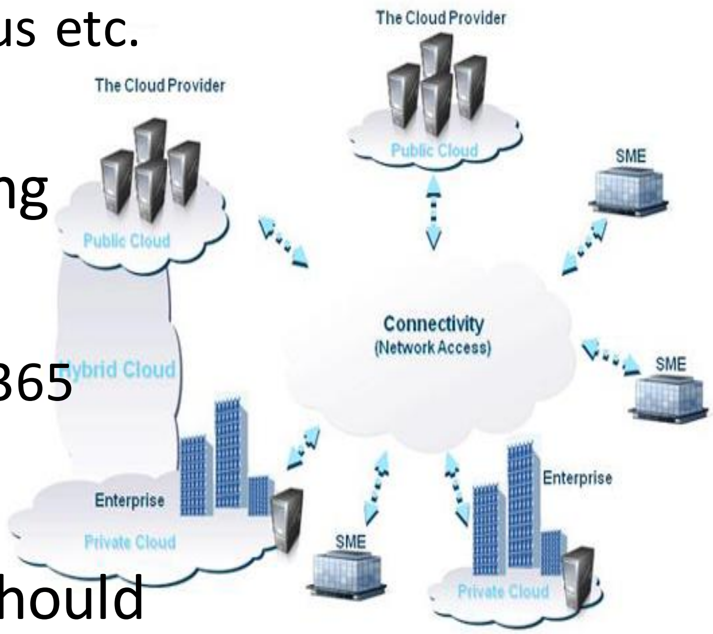
- Easy to write and run highly parallel programs in new cloud programming paradigms:
  - **Google:** MapReduce and Sawzall
  - **Amazon:** Elastic MapReduce service (pay-as-you-go)
  - Google (MapReduce)
    - Indexing: a chain of 24 MapReduce jobs
    - ~200K jobs processing 50PB/month (in 2006)
  - **Yahoo!** (Hadoop + Pig)
    - WebMap: a chain of several MapReduce jobs
    - 300 TB of data, 10K cores, many tens of hours (~2008)
  - **Facebook** (Hadoop + Hive)
    - ~300TB total, adding 2TB/day (in 2008)
    - 3K jobs processing 55TB/day
  - NoSQL: MySQL is an industry standard, but Cassandra is 2400 times faster

Two Categories of Clouds

- Can be either a (i) public cloud, or (ii) private cloud

- **Private clouds** are accessible only to company employees
- Example of popular vendors for creating private clouds are VMware, Microsoft Azure, Eucalyptus etc.

- **Public clouds** provide service to any paying customer
- Examples of large public cloud services include Amazon EC2, Google AppEngine, Gmail, Office365 and Dropbox etc.



- You're starting a new service/company: should you use a public cloud or purchase your own private cloud?

Single site Cloud: to Outsource or Own?

- Medium-sized organization: wishes to run a service for *M* months
  - Service requires 128 servers (1024 cores) and 524 TB
- **Outsource** (e.g., via AWS): *monthly* cost
  - S3 costs: $0.12 per GB month. EC2 costs: $0.10 per CPU hour (costs from 2009) Storage = $ 0.12 X 524 X 1000 ~ $62 K
  - Total = Storage + CPUs = $62 K + $0.10 X 1024 X 24 X 30 ~ $136 K
- **Own**: monthly cost
  - Storage ~ $349 K / *M* Total ~ $ 1555 K / *M* + 7.5 K (includes 1 sysadmin / 100 nodes)
    - using 0.45:0.4:0.15 split for hardware:power: network and 3 year lifetime of hardware

○ Breakeven analysis: **more preferable to own if:**

- $349 K / *M* < $62 K (storage)

- $ 1555 K / *M* + 7.5 K < $136 K (overall)

Breakeven points

○ *M* > 5.55 months (storage)

○ *M* > 12 months (overall)

-Startups use clouds a lot
-Cloud providers benefit monetarily most from storage

Conclusion

- Limitations of current cloud data center.
- Understanding the concept of edge data center.
- Clouds build on many previous generations of distributed systems
- Characteristics of cloud computing problem

  - **Scale, On-demand access, data-intensive, new programming**

THANK YOU!