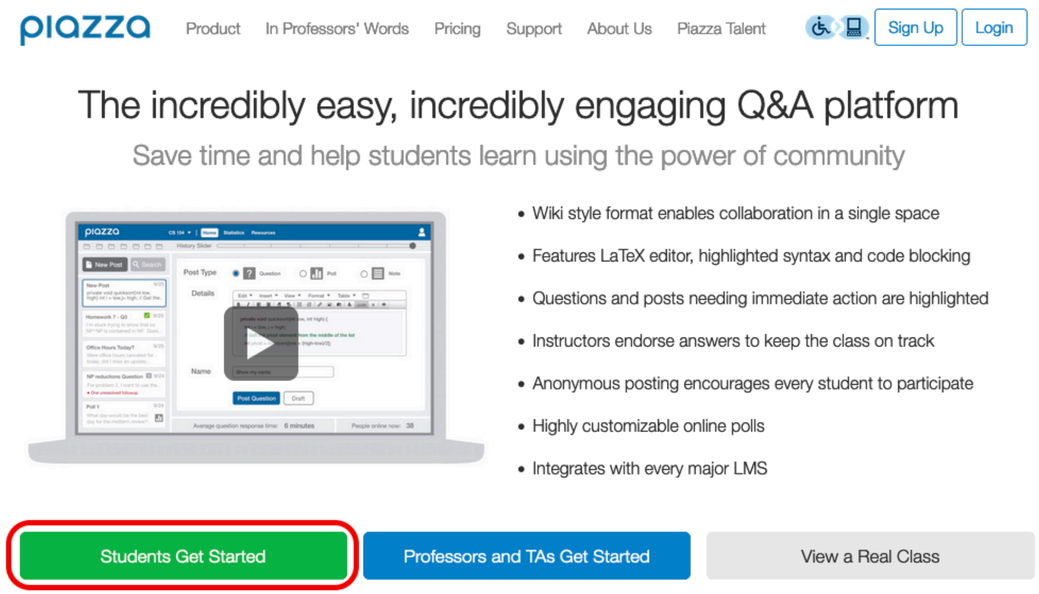
CS4740
CLOUD COMPUTING

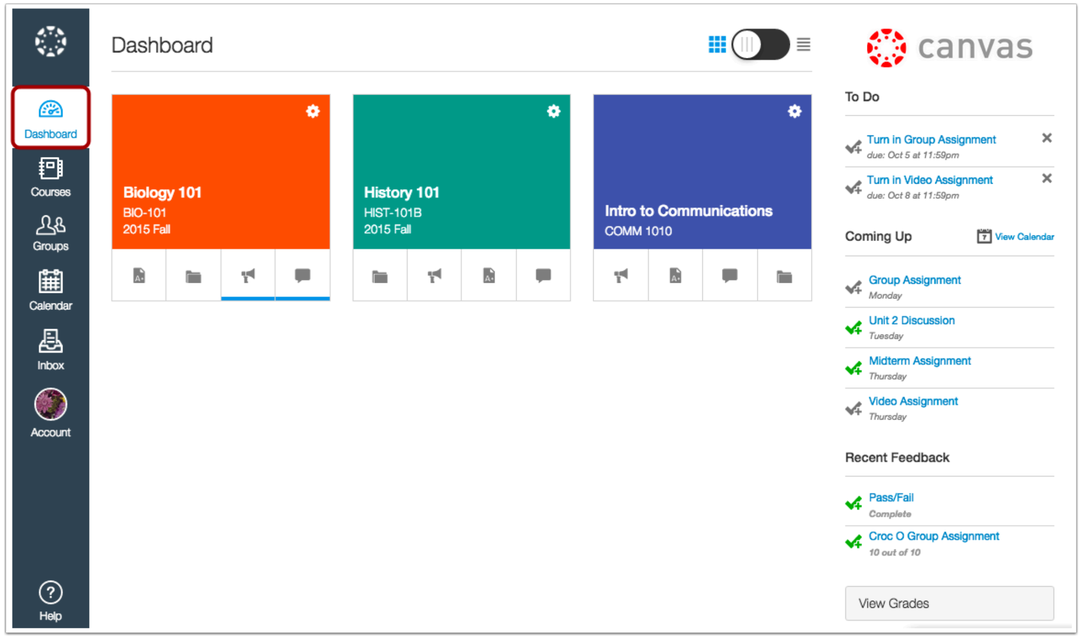
Intro to Cloud

Prof. Chang Lou, UVA CS, Spring 2024

STUDENT FEEDBACK: PIAZZA?

Do you prefer to switch discussion to Piazza for feat. like "Anonymous Posting"?





AGENDA

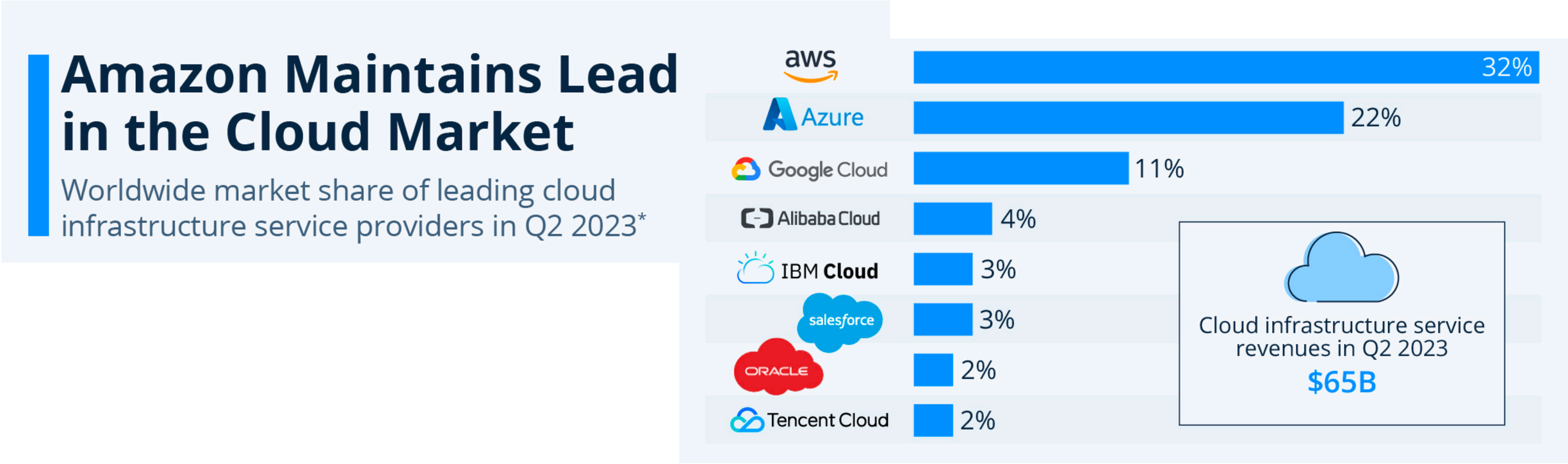
- What is a Cloud?
- Where does the idea of Cloud come from?
- What types of Cloud should I choose?

QUIZ

- Which company is the biggest player of today's Cloud market?
  - A. Amazon
  - B. Alibaba Cloud
  - C. Google
  - D. Microsoft

QUIZ

— Which company is the leading player of today's Cloud market?



(THERE WAS) THE HYPE!

- Forrester in 2010 – Cloud computing will go from $40.7 billion in 2010 to $241 billion in 2020.
- Today: Cloud Market is $**545B** (expected to reach $**1.2T** by 2027)

- 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 Cloud/Compute Engine/AppEngine
- Rightscale, Salesforce, EMC, Gigaspaces, 10gen, Datastax, Oracle, VMWare, Yahoo, Cloudera, and many many more!
- They become the backbone of modern computing.

TWO 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.
    - Recently: 0.09c to 2 c per GB month
  - Amazon EC2 (Elastic Compute Cloud): upload and run arbitrary OS images, pay per CPU hour used
    - Recently: 2 c per CPU hr to 40c per CPU hr (depending on strength), only CPUs not GPUs
  - Google cloud: similar pricing ranges as above
  - Google AppEngine/Compute Engine: develop applications within their appengine framework, upload data that will be imported into their format, and run

WHY WE NEED CLOUD?

- Customers Save Time and $$$!
  - Dave Power, Associate Information Consultant at Eli Lilly and Company: “With AWS, Powers said, a new server can be up and running in three minutes (it used to take Eli Lilly 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). … It's just shy of instantaneous.”
  - Jim Swartz, CIO, Sybase: “At Sybase, a private cloud of virtual servers inside its datacenter has saved nearly $US2 million annually since 2006, Swartz says, because the company can share computing power and storage resources across servers.”

What is a cloud?

WHAT IS A CLOUD?

— It’s a cluster!
— It’s a supercomputer!
— It’s a datastore!

WHAT IS A CLOUD?

- It’s a cluster!
- It’s a supercomputer!
- It’s a datastore!

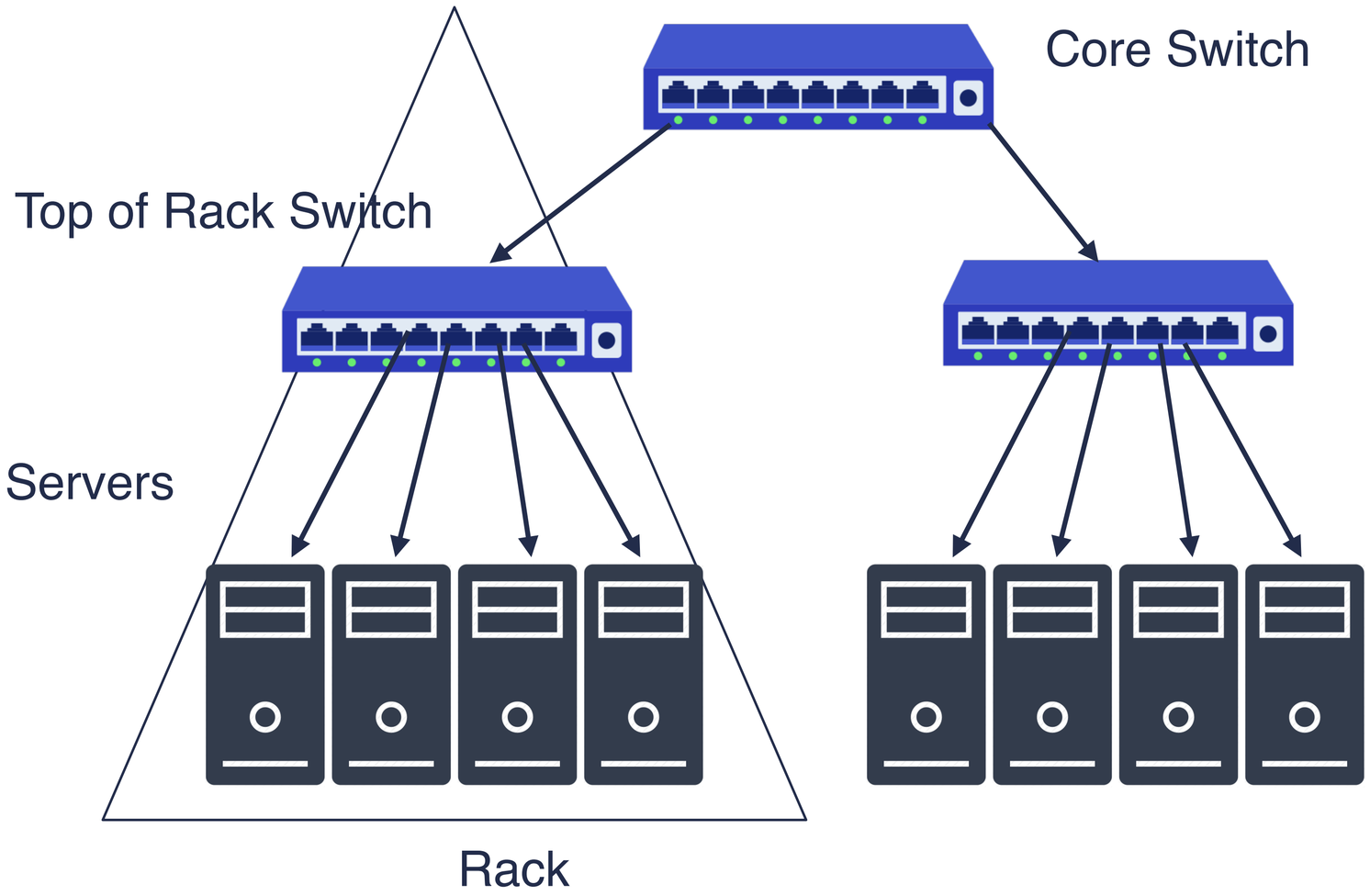
- Cloud = **Lots of storage + compute cycles nearby**

WHAT IS A CLOUD?

- A **single-site** cloud (aka “Datacenter”) consists of
  - Compute nodes (grouped into racks) (2)
  - Switches, connecting the racks
  - A network topology, e.g., hierarchical
  - Storage (backend) nodes connected to the network (3)
  - Front-end for submitting jobs and receiving client requests (1)
  - Software Services
- A geographically **distributed** cloud consists of
  - Multiple such sites
  - Each site perhaps with a different structure and services

*(1,2,3: Often called “three-tier architecture”)

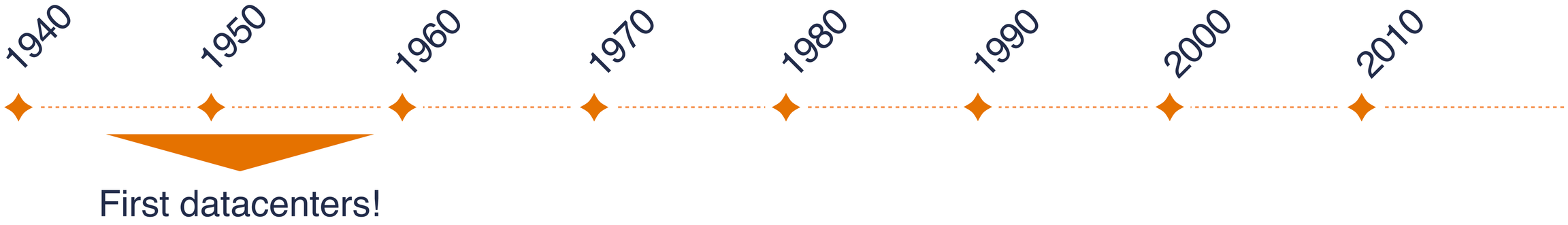
A SAMPLE CLOUD TOPOLOGY



A CLOUDY HISTORY OF TIME

A CLOUDY HISTORY OF TIME

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



A CLOUDY HISTORY OF TIME

- 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..

A CLOUDY HISTORY OF TIME

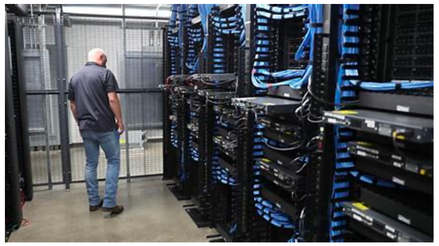
- Why time-sharing started to decline?
- Answer: Rise of PC

A CLOUDY HISTORY OF TIME

Open Science Grid

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

A CLOUDY HISTORY OF TIME



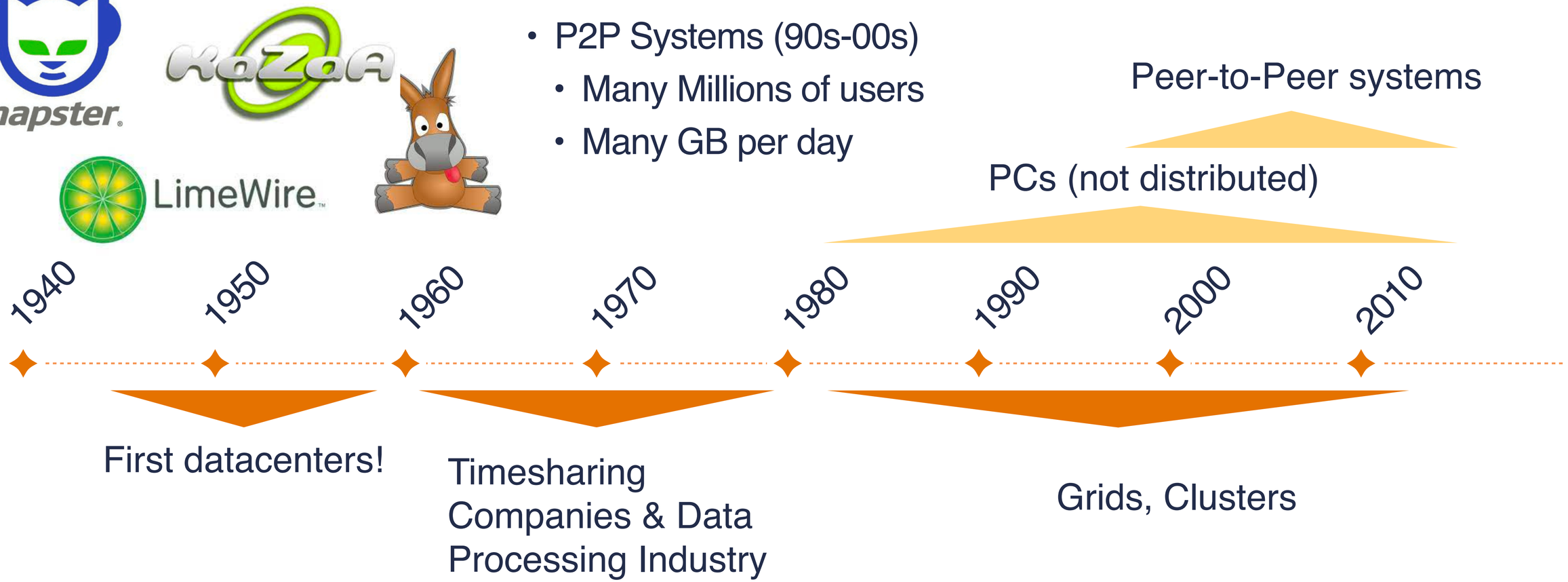
CRAY
THE SUPERCOMPUTER COMPANY

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

- Difference between Grids and Clusters?

A CLOUDY HISTORY OF TIME

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



A CLOUDY HISTORY OF TIME

napster

KaZaA

LimeWire

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

Peer-to-Peer systems

PCs (not distributed)

1940 1950 1960 1970 1980 1990 2000 2010

First datacenters!

Timesharing Companies & Data Processing Industry

Grids, Clusters

Cloud

TRENDS: TECHNOLOGY

- Doubling Periods – storage: 12 mos, bandwidth: 9 mos, and (what law is this?) cpu compute capacity: 18 mos
- Then and Now
  - Bandwidth
    - 1985: mostly 56Kbps links nationwide
    - Today: Tbps links widespread
  - Disk capacity
    - Today's PCs have TBs, far more than a 1990 supercomputer

TRENDS: USERS

- Then and Now:
  - 1990: biologists were running small single-molecule simulations
  - Today: CERN’s Large Hadron Collider producing many PB/year

PROPHECIES

- In 1965, MIT's Fernando Corbató and the other designers 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 your favorite Intensive Compute & Communicate Application
  - Discussion: Have today’s clouds brought us closer to this reality?

FOUR FEATURES NEW IN TODAY'S CLOUDS

- (1) Massive scale.
- (2) On-demand access: Pay-as-you-go, no upfront commitment.
- (3) Data-intensive Nature: What was MBs has now become TBs, PBs and XBs.
- (4) New Cloud Programming Paradigms

(1) MASSIVE SCALE

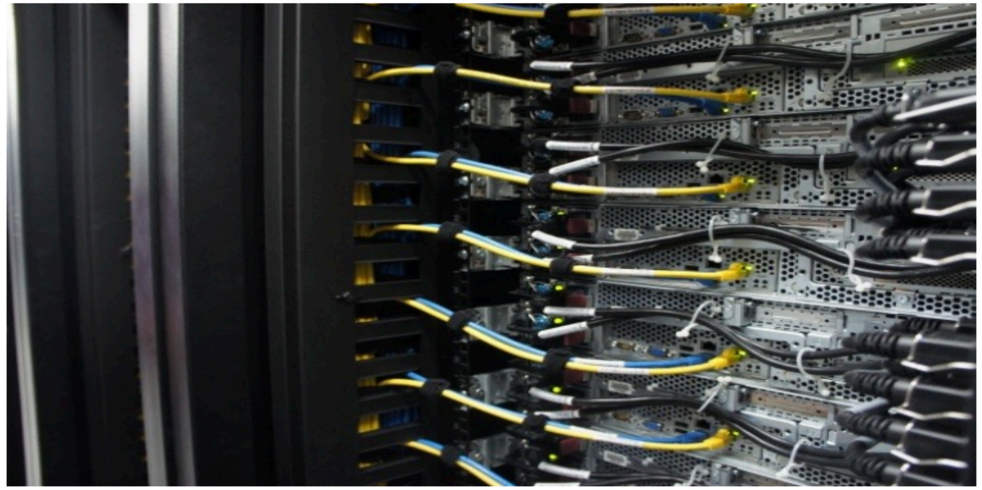
- Facebook [GigaOm, 2012]: 30K in 2009 -> 60K in 2010 -> 180K in 2012
- Microsoft [NYTimes, 2008]
  - 150K machines. Growth rate of 10K per month
  - 80K total running Bing
  - In 2013, Microsoft Cosmos had 110K machines (4 sites)
- Yahoo! [2009]:
  - – 100K, Split into clusters of 4000
- AWS EC2 [Randy Bias, 2009]
  - 40K machines – 8 cores/machine
- eBay [2012]: 50K machines
- HP [2012]: 380K in 180 DCs
- Google [2011, Data Center Knowledge] : 900K

Q: any guess on how many servers Microsoft Azure have today?

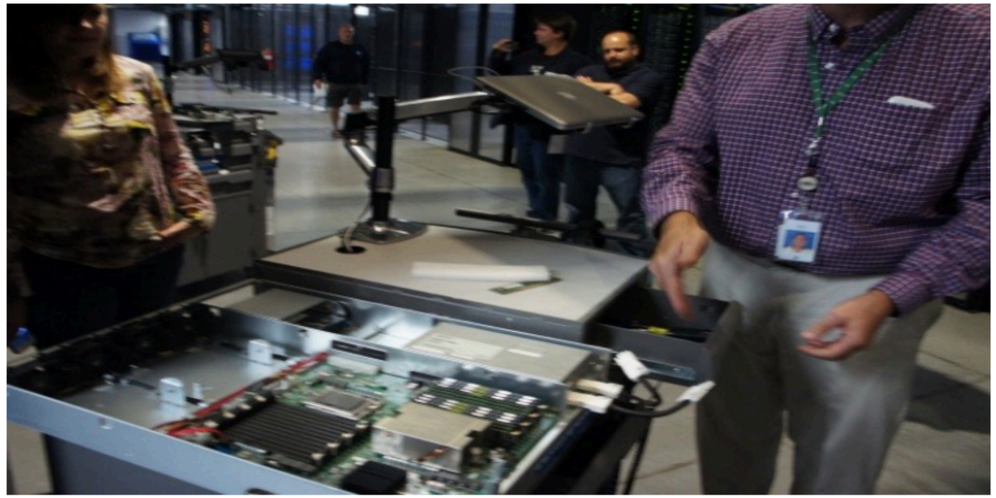
What does a datacenter look like from inside?

Servers

Front



Back



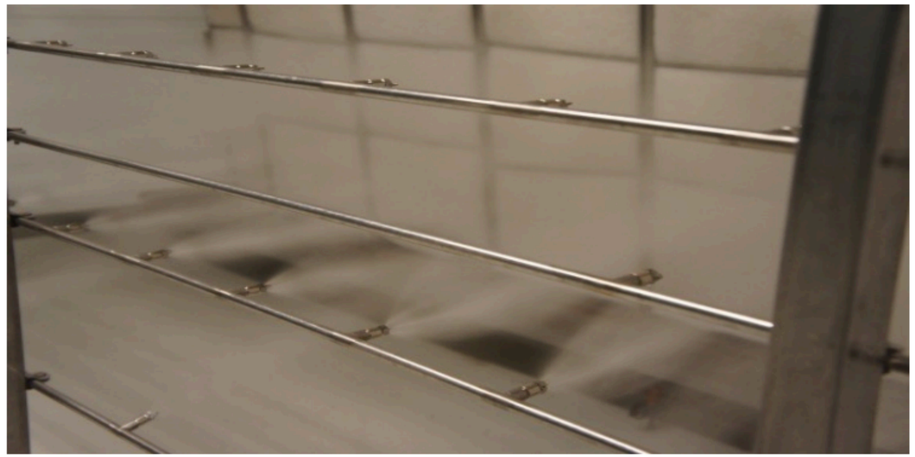
In

Some highly secure (e.g., financial info)

Cooling

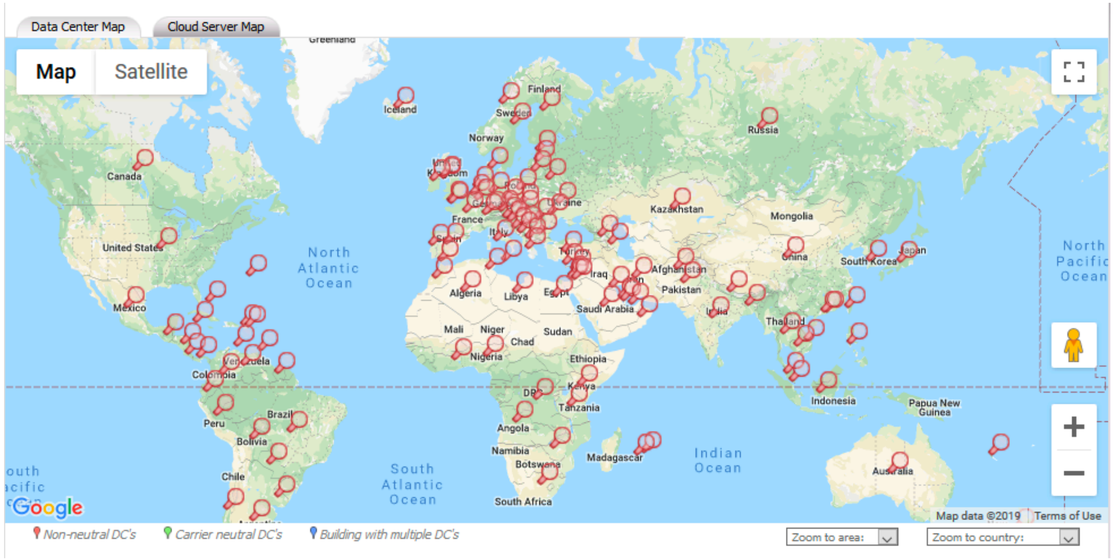
Air sucked in from top (also, Bugzappers)

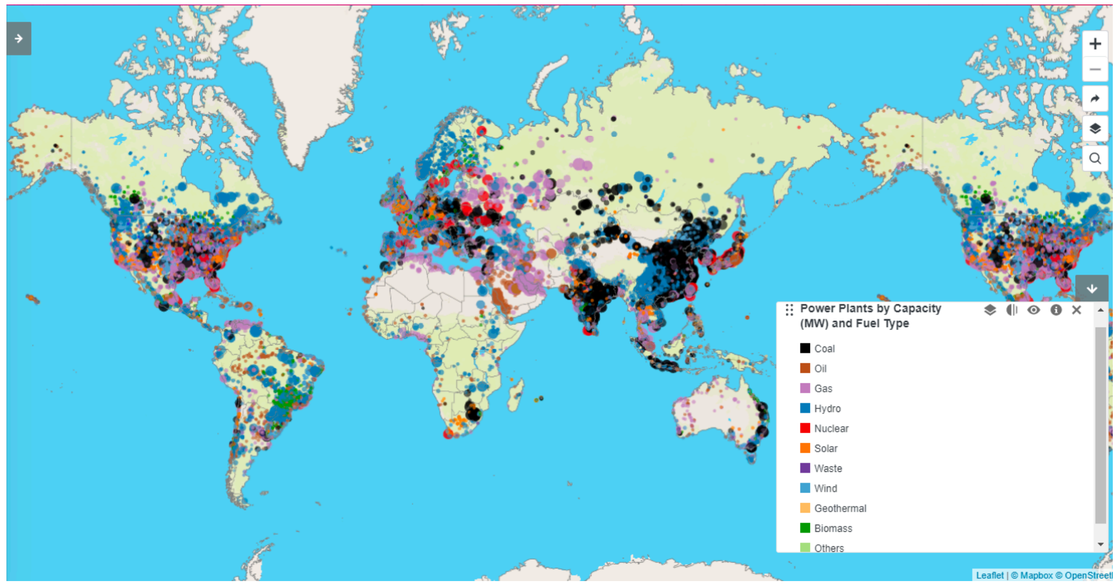
Water purified



Water sprayed into air

15 motors per server bank

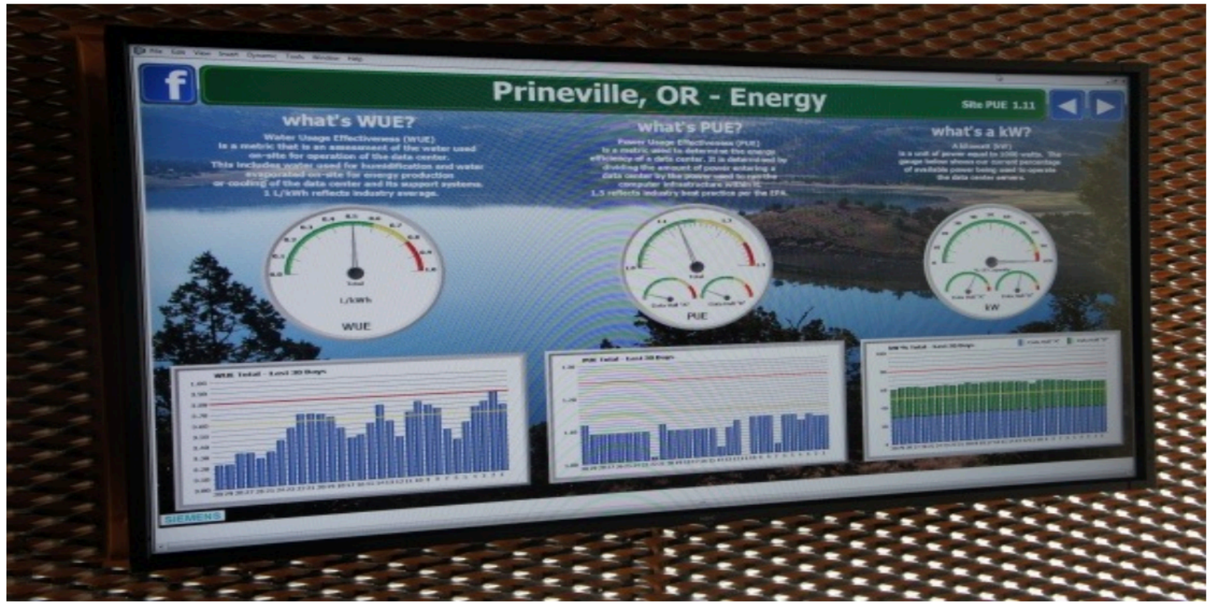




Similarities between data center locations and power plant locations?

Power

Off-site



On-site

- 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~1.1)

(2) ON-DEMAND ACCESS: *AAS CLASSIFICATION

- On-demand: calling a cab vs. renting a car 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

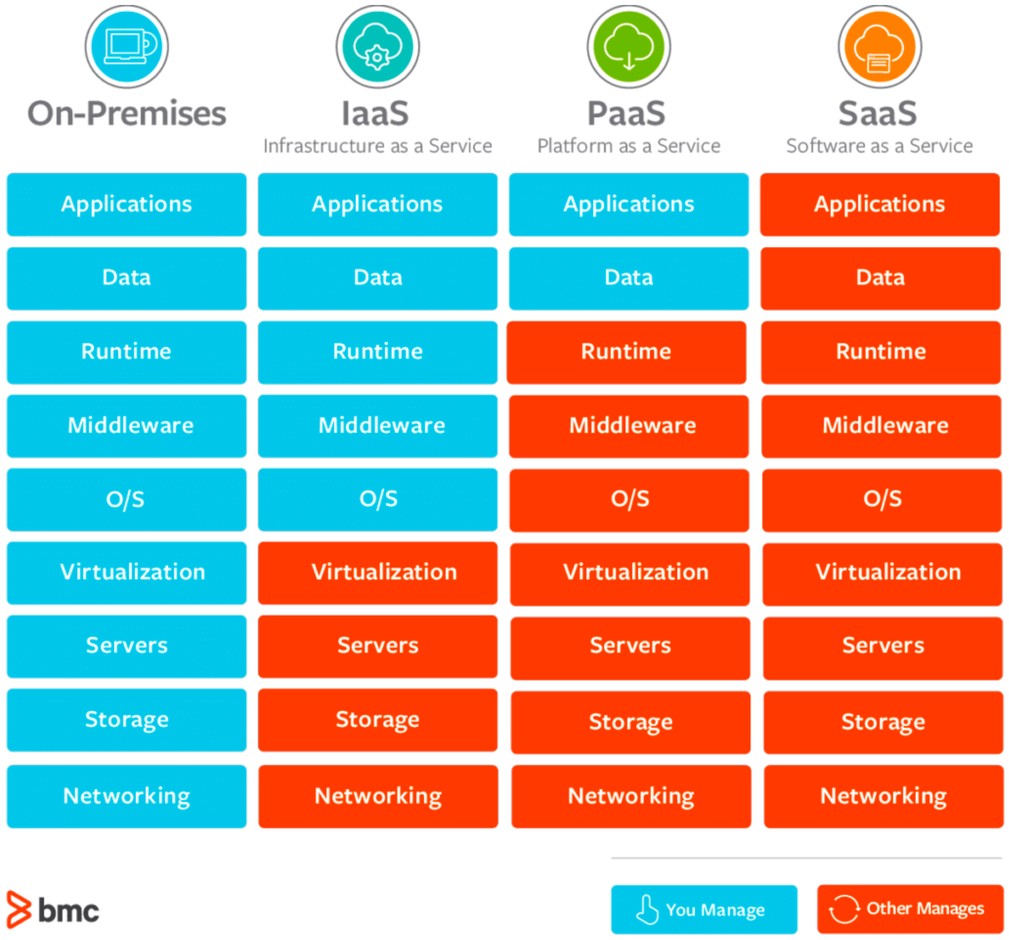
(2) ON-DEMAND ACCESS: *AAS CLASSIFICATION

- HaaS: Hardware as a Service
  - You 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
  - You get access to flexible computing and storage infrastructure. Virtualization or containerization is one way of achieving this (cgroups, Kubernetes, Dockers, VMs,…). Often said to subsume HaaS.
  - Ex: Amazon Web Services (AWS: EC2 and S3), OpenStack, Eucalyptus, Rightscale, Microsoft Azure, Google Cloud.

(2) ON-DEMAND ACCESS: *AAS CLASSIFICATION

- PaaS: Platform as a Service
  - You 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
  - You get access to software services, when you need them. Often said to subsume SOA (Service Oriented Architectures).
  - Ex: Google docs, MS Office 365 Online
- And new recently: FaaS = Function as a Service
  - Ex: AWS Lambda, Azure Functions, etc.

(2) ON-DEMAND ACCESS: *AAS CLASSIFICATION



(3) 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)

(4) NEW CLOUD PROGRAMMING PARADIGMS

- Easy to write and run applications in cloud programming paradigms:
  - Data Analytics (e.g. MapReduce): parallel programming
  - Serverless: auto-scaling, low-overhead
  - Microservice: easier for scaling and maintenance
  - ...

TAKEAWAYS

- Clouds build on many previous generations of distributed systems
- Especially the timesharing and data processing industry of the 1960-70s.
- Need to identify unique aspects of a problem to classify it as a new cloud computing problem
- Next class: distributed system foundation, with Starbucks!

ACKNOWLEDGEMENT

THIS COURSE IS DEVELOPED HEAVILY BASED ON COURSE MATERIALS SHARED BY PROF. INDRANIL GUPTA, PROF. ROBERT MORRIS AND PROF. ROXANA GEAMBASU. MANY APPRECIATIONS FOR GENEROUSLY SHARING THEIR MATERIALS AND TEACHING INSIGHTS.