CPSC 436C
Cloud Computing for Data Science

Introduction

Maryam R.Aliabadi

mraiyata@cs.ubc.ca

Spring 2024

Todays' Agenda

- 1st part
  - Getting to know each other
  - What this course is about
  - How the course is structured
  - How to successfully pass the course

- 2nd part
  - Fundamentals of Data centers and Cloud

Course Info

- Instructor
  - Maryam Raiyat Aliabadi
  - Email: mraiyata@cs.ubc.ca
  - Class time : Tue/Thu at 8AM - 9:30PM
  - Office hours: Thu at 10AM - 12PM

- Teaching Assistants
  - 1-Aryan Bhairaw
  - Email: baryan01@student.ubc.ca
  - 2-Arman Moztarzadeh
  - Email: arman88@student.ubc.ca
  - 3-Ryan Dick
  - Email: rdick01@student.ubc.ca
  - Office hours/lab tutorials : TBD

- Contents and notes will be available on Canvas

About me

- Education
  - PhD : Computer Science and Engineering, SBU/UBC
  - MSc. Computer (Software) Engineering, UBC

- Positions
  - Post-doctoral Teaching and Research Fellow, UBC
  - Teaching Faculty, New York Tech, Vancouver
  - Co-Founder and CEO, Kids Shield Services Inc (www.kidsshield.ca).

- More background
  - 4+ years' experience teaching in international universities
  - 10+ years' experience in IT industry
  - Interests: Cloud Computing, Cyber-Physical System Security, Software analysis and Testing,

Who does what?

Teacher does the TEACHING!

Students do the LEARNING!

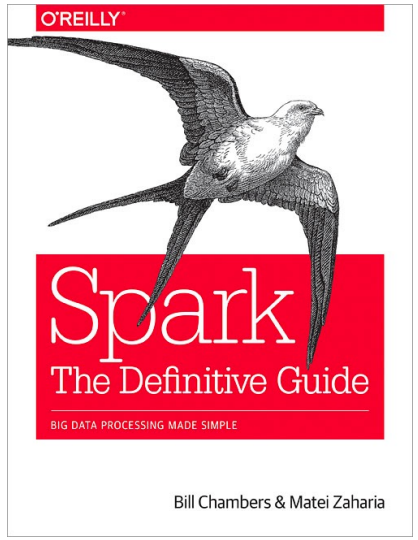
What is this course all about?

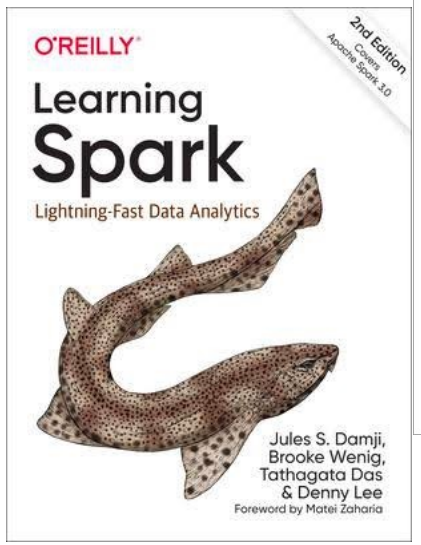
This course is an introduction to cloud computing designed for the students who wish to use the cloud for data science applications. It covers the topics of how cloud computing can be used to support data science workflows, including data storage, processing, analysis and visualization. It also includes the security considerations for the entire pipeline. Overall, the course provides students with the skills and knowledge necessary to effectively use cloud computing for Design, implementation, test, and deployment of data science applications

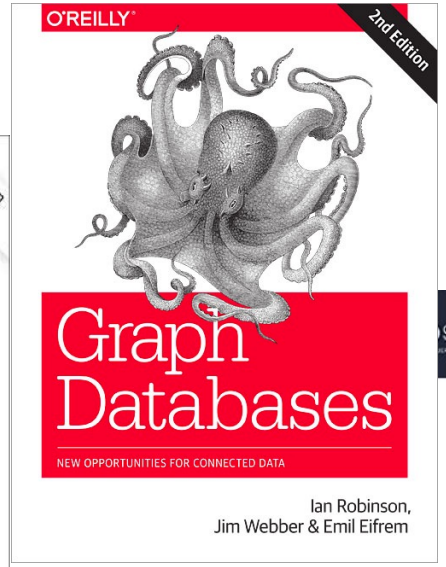
What will you learn in this course?

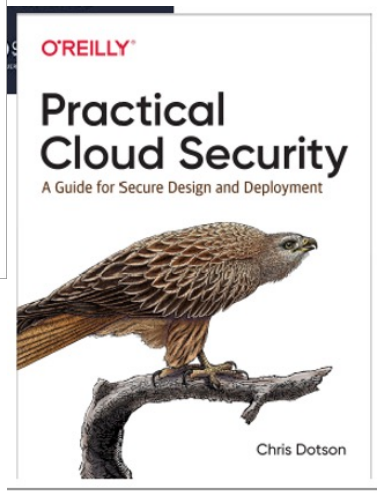
- **Use** different service delivery models such as virtualization, containers and serverless in cloud;
- **Identify** trade-offs in different data storage solutions;
- **Identify** the appropriate tools and architectures to implement a cloud-based design,
- **Create** distributed computing pipeline using cloud-based execution engines, including scheduling the jobs that comprise the pipeline;
- **Analyze** large datasets using distributed computing pipeline for different applications;
- **Analyze** the trade-offs between performance and cost using different designs to meet real-world constraints;
- **Use** basic cloud security services such as Identity and access management, network security, compliance and incident response;

Supplementary Optional Resources









Assessment

- **Group-based (40%)**
  - Assignments (40%)
    - Assignment 1-6 : 5 points each
    - Assignment 7 : 10 points each
- **Individual (60%)**
  - Midterm exam (25%)
  - Final Exam (30%)
  - In-Class Participation (5%)

PERCENTAGE	GRADE
90 – 100	A
85 – 90	A-
80 – 85	B+
70 – 80	B
65 – 70	B-
60 – 65	C+
55 – 60	C
0 – 55	F

Assignments

- Done in a group of maximum three students. Let TAs know if you want to work in a team but have no team-mates.

- There will be one submission from a group on Canvas.

- Assignments should be submitted before due date to avoid penalty of late submission. Submission are accepted by two weeks before final exam.

Labs/Office hours

- Labs/Office hours are designed to
  - teach you the alphabets of working with Cloud
  - explain the assignments
  - answer your questions

- Two sections per week

- Two cloud platforms: AWS and Azure

Midterm and Final Exam

- Short answer questions
- Long answer questions
- Multiple-choice questions
- Closed book

- Computer-Based Testing Facility (CBTF)
  - Synchronous

SYLLABUS

1	**Week 1: Introduction to Data Center and Cloud**
2	**Week 2 : Service Delivery Models - Function as a service & Containers**
3	**Week 3 : Service Delivery Models - Virtualization**
4	**Week 4 : Big Data**
5	**Week 5 : Data Store**
6	**Week 6 : Data Management Systems**
7	**Week 6 : Mid-term Exam**
8	**Week 7: Data Processing**
9	**Week 8 : Structured Data Processing and Distributed Machine Learning**
10	**Week 9 : Stream Processing**
11	**Week 10 : Graph Processing**
12	**Week 11 : Resource Management**
13	**Week 12 : Cloud Security and Advanced topics**
14	**Week 13 : Final Exam**

UBC Respectful Environment Policy

The University of British Columbia envisions a climate in which students, faculty and staff are provided with the best possible conditions for **learning**, **researching** and **working**, including an environment that is dedicated to excellence, **equity** and **mutual respect**. The University of British Columbia strives to realize this vision by establishing employment and educational practices that respect the **dignity** of individuals and make it possible for everyone to live, work and study in a **positive** and **supportive** environment, free from harmful behaviours such as bullying and harassment.

https://hr.ubc.ca/working-ubc/respectful-environment

Academic Integrity Policy

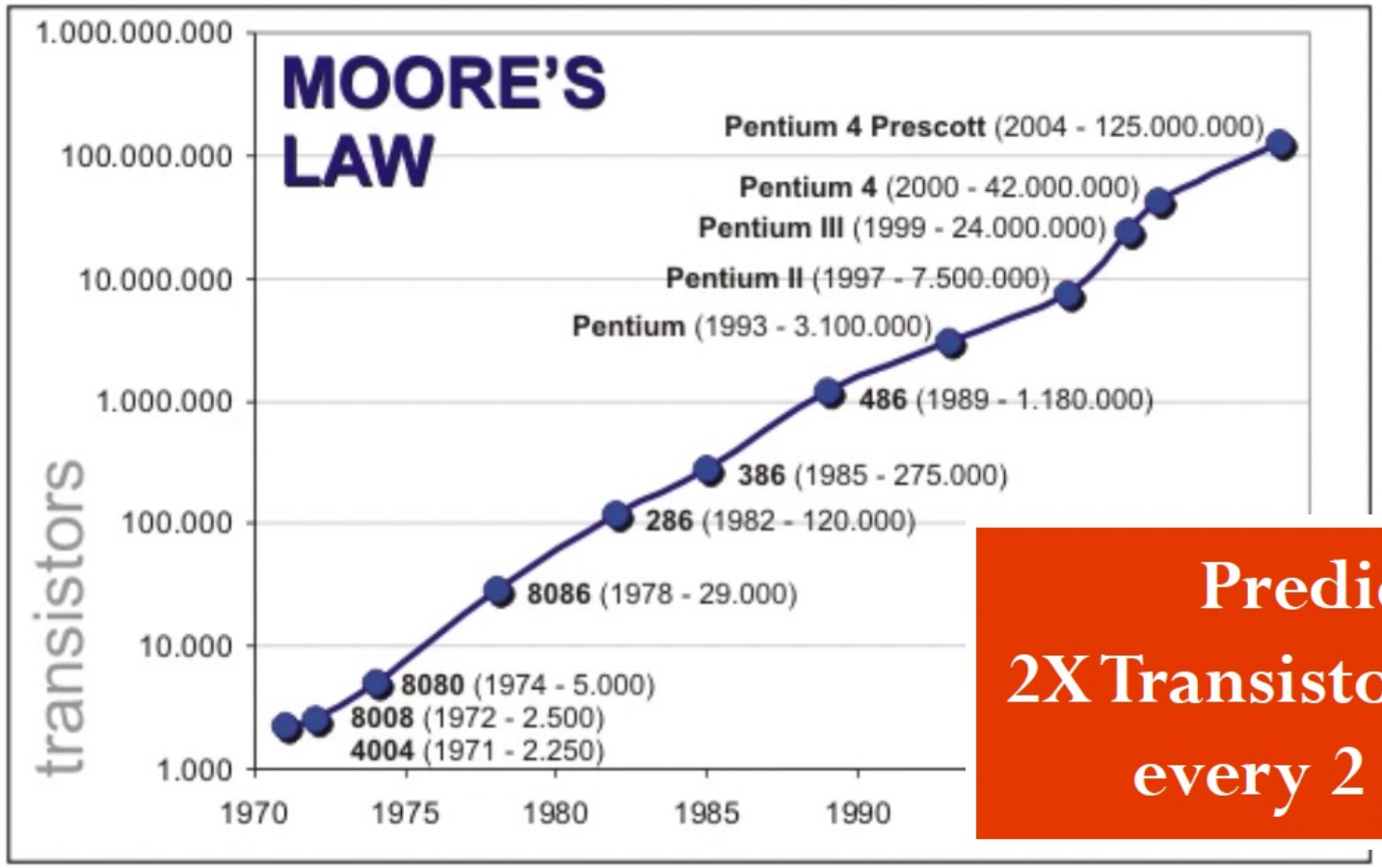
All of the work you submit must be done by you and your work must not be submitted by someone else. Plagiarism is academic fraud and is taken very seriously. The department uses software that compares programs for evidence of similar code/report. Please read the Rules and Regulations from the UBC Academic Integrity website: https://academicintegrity.ubc.ca/student-start/

Question?

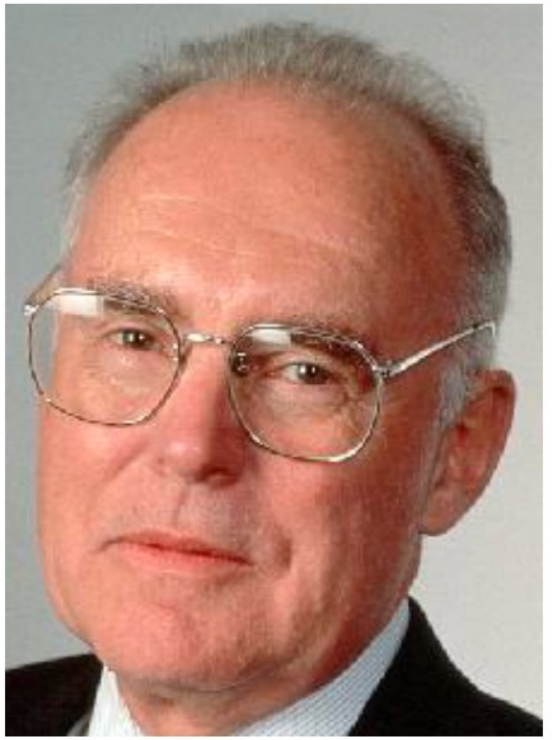
History

- ❑ Moore's Law (1965)
- ❑ Scaling Challenges (2020)
- ❑ Decline of Moore's Law from Power/Performance Perspective
- ❑ Transition from Uni-core to Multi-core:
- ❑ Data Centers

The computer revolution



Predicts:
2X Transistors / chip every 2 years

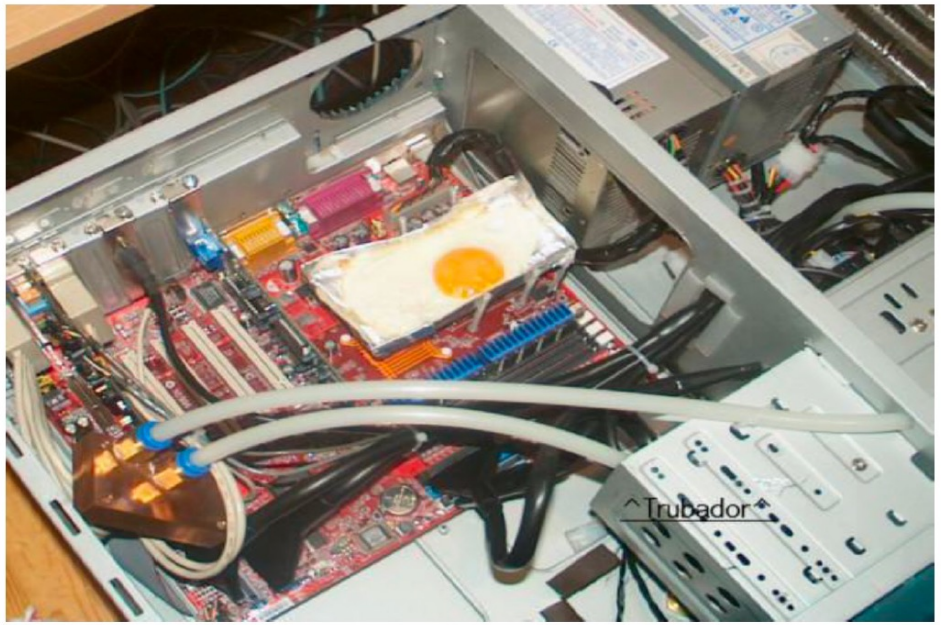


Gordon Moore
Intel Cofounder

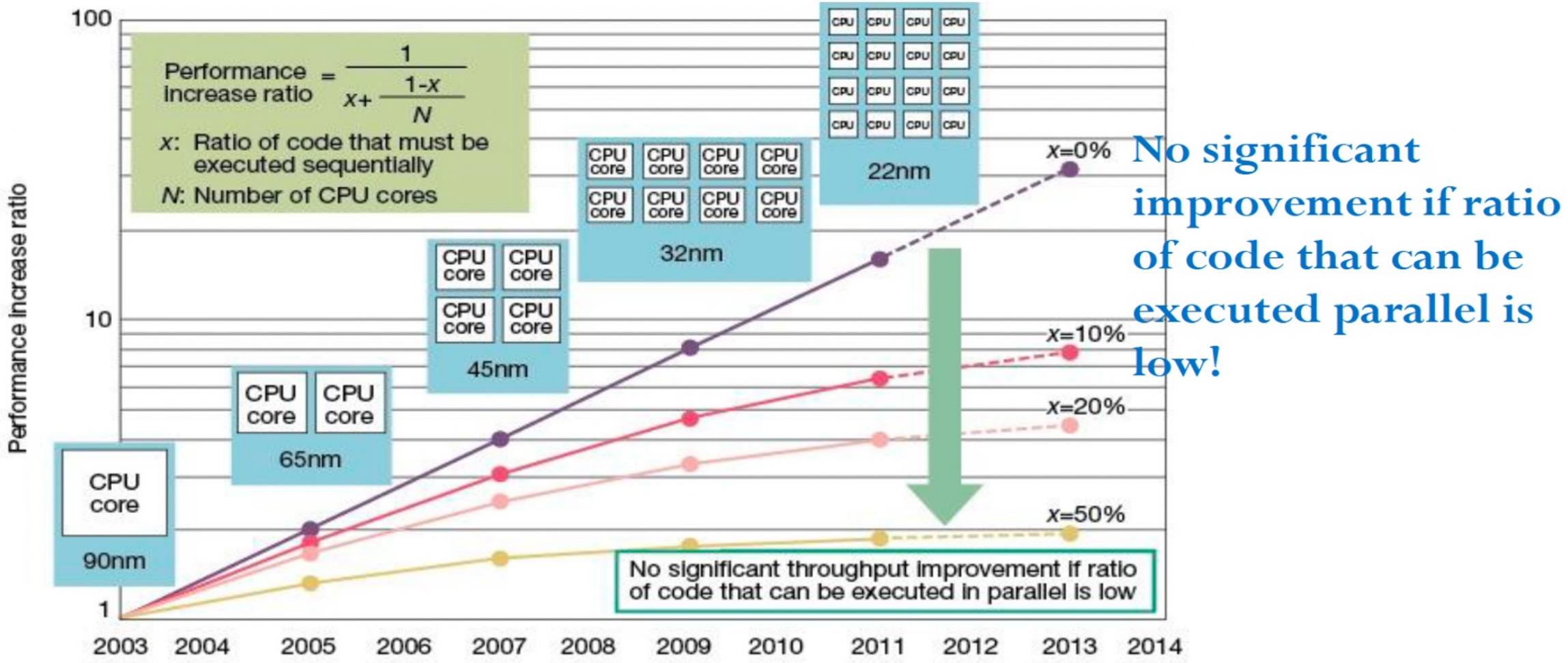
https://www.youtube.com/watch?v=CUnQNTwmHHo

Power Challenges

- Power is a challenge for integrated circuits for two reasons:
  - First, power must be brought in and distributed around the chip; modern microprocessors use hundreds of pins just for power and ground.
  - Second, power is dissipated as heat and must be removed. Server chips can burn more than 100 watts, and cooling the chip and the surrounding system is a major expense in Warehouse Scale Computers.
- How else can we improve performance?

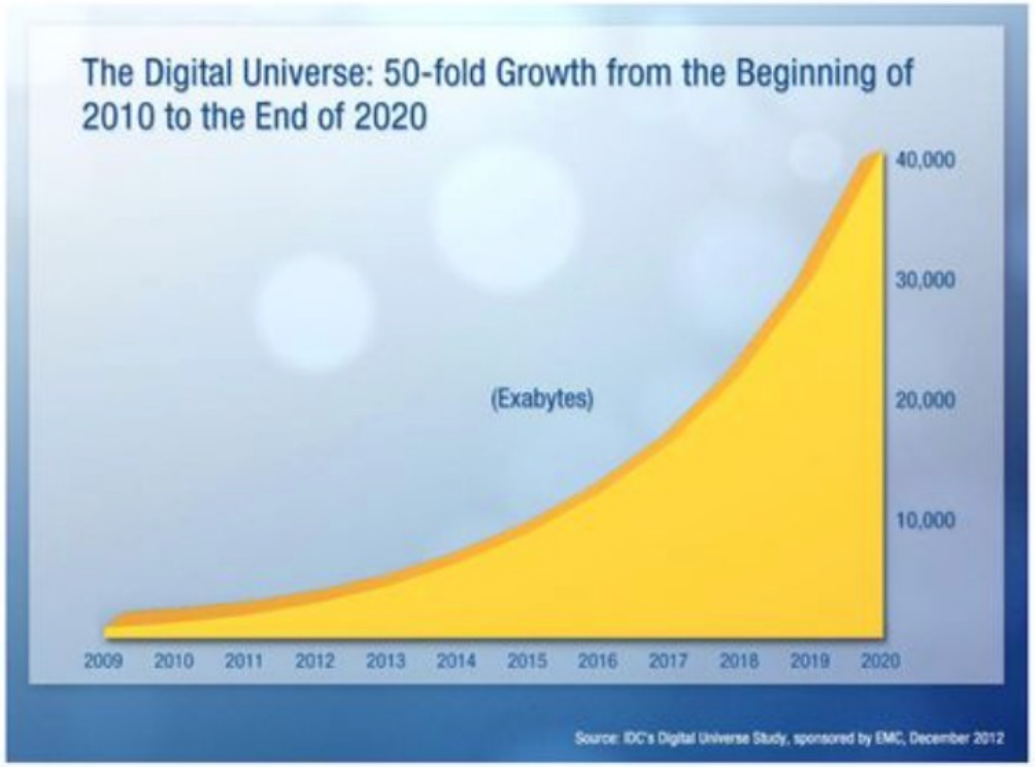


From Uniprocessors to Multiprocessors



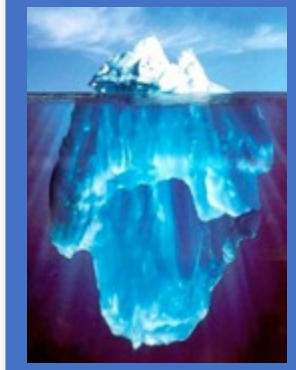
Why a Computer is not enough?

- Too much data
- Too little storage capacity
- Not enough I/O bandwidth
- Not enough computing capability



Sources Driving Big Data

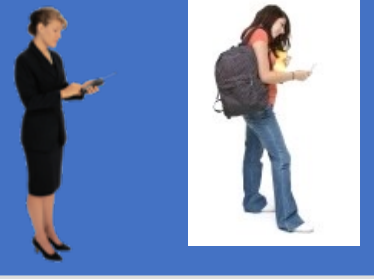
It's All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,..
Friend Request
Transaction
Network message
Fault
…

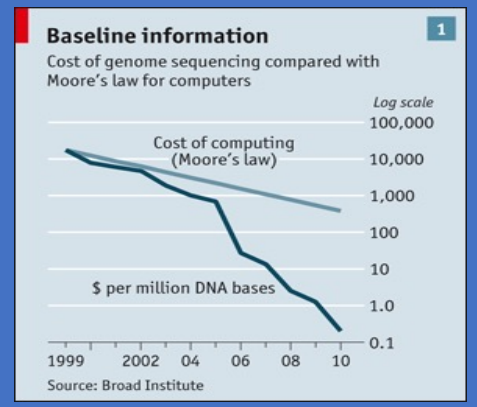
User Generated (Web & Mobile)

…
..

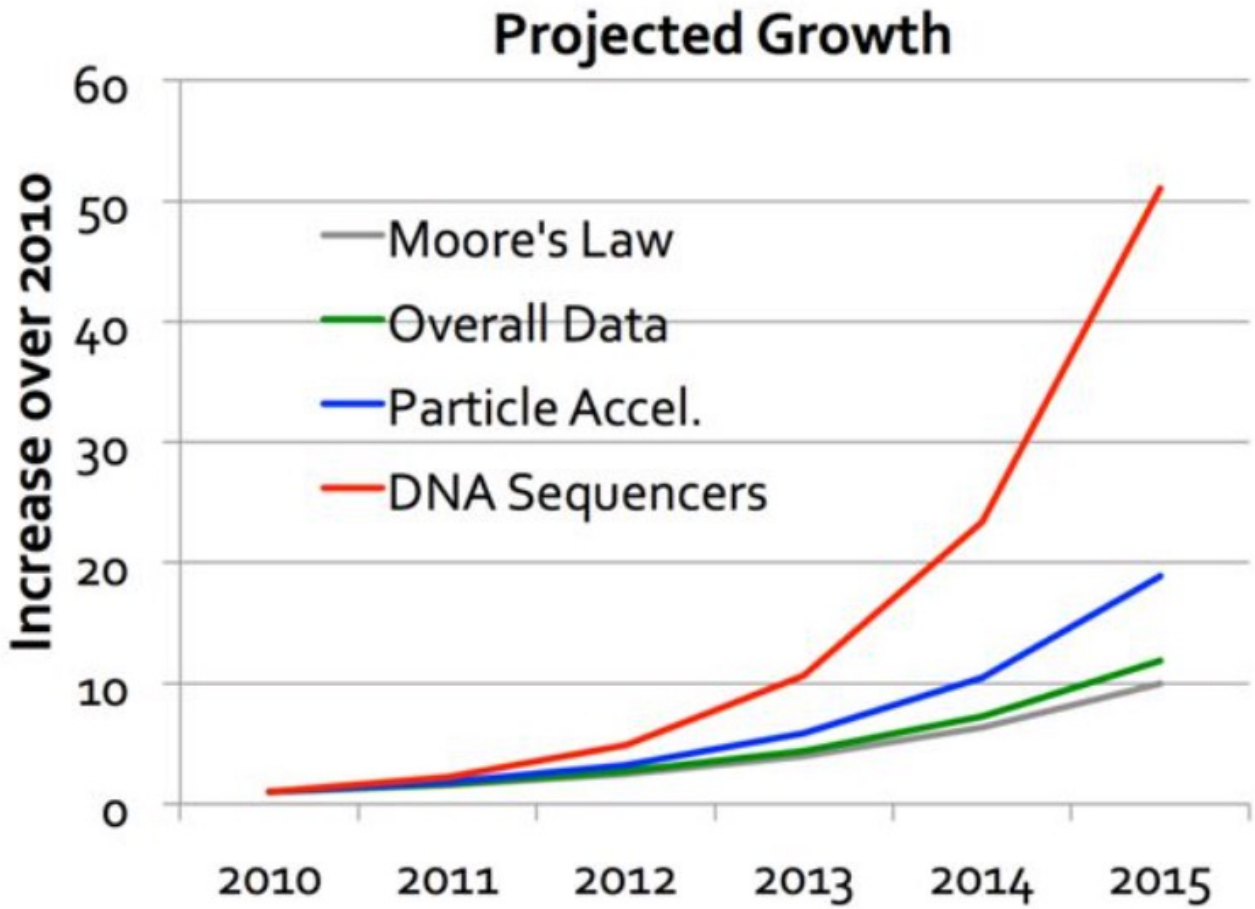


Internet of Things / M2M

Scientific Computing



Data Grows Faster than Moore's Law



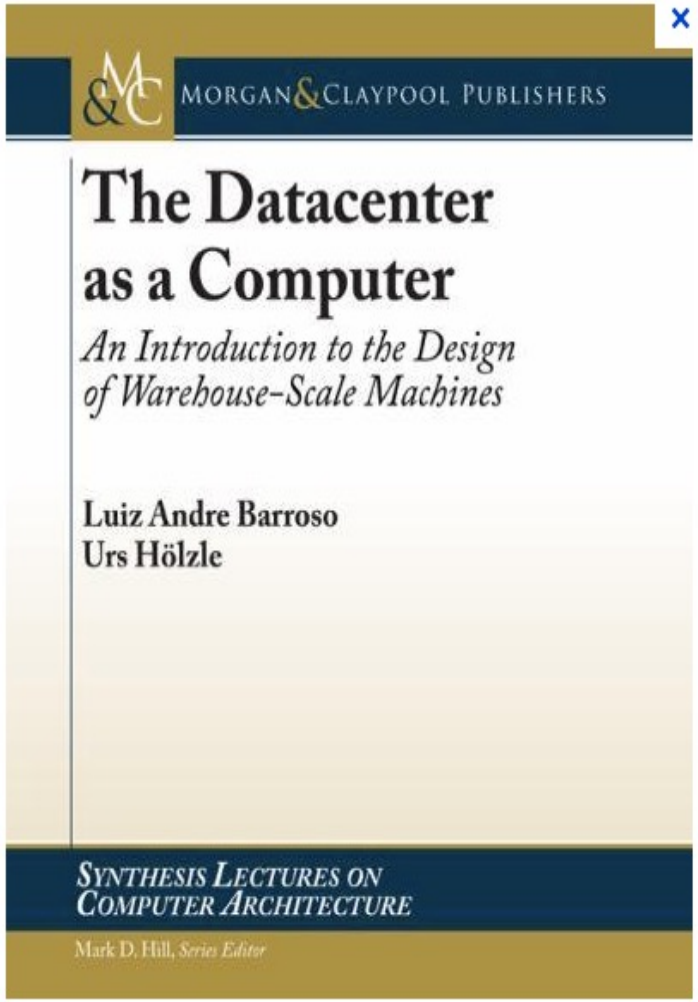
Text Data

Network Data

Multivariate Data

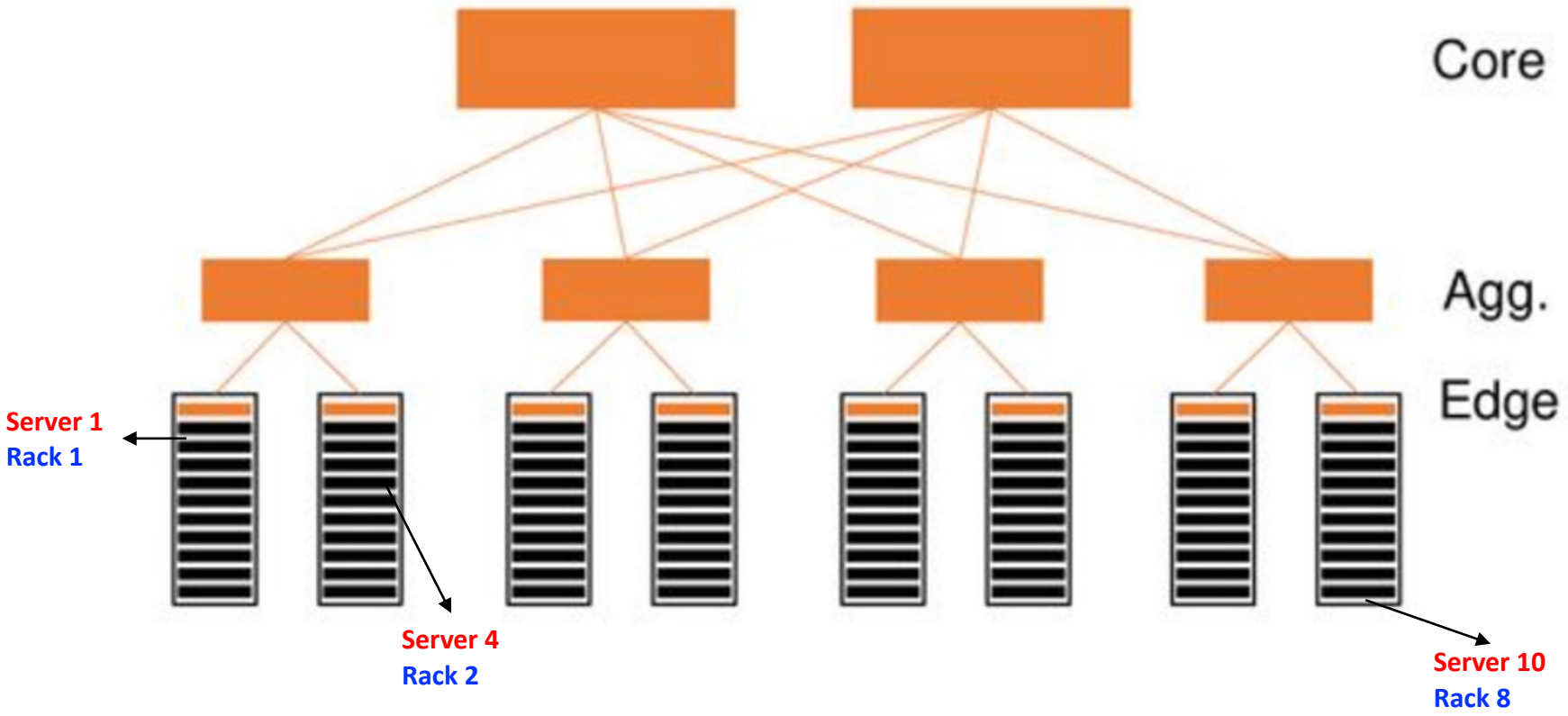
Solution: Data Centers

The Datacenter is the new Computer



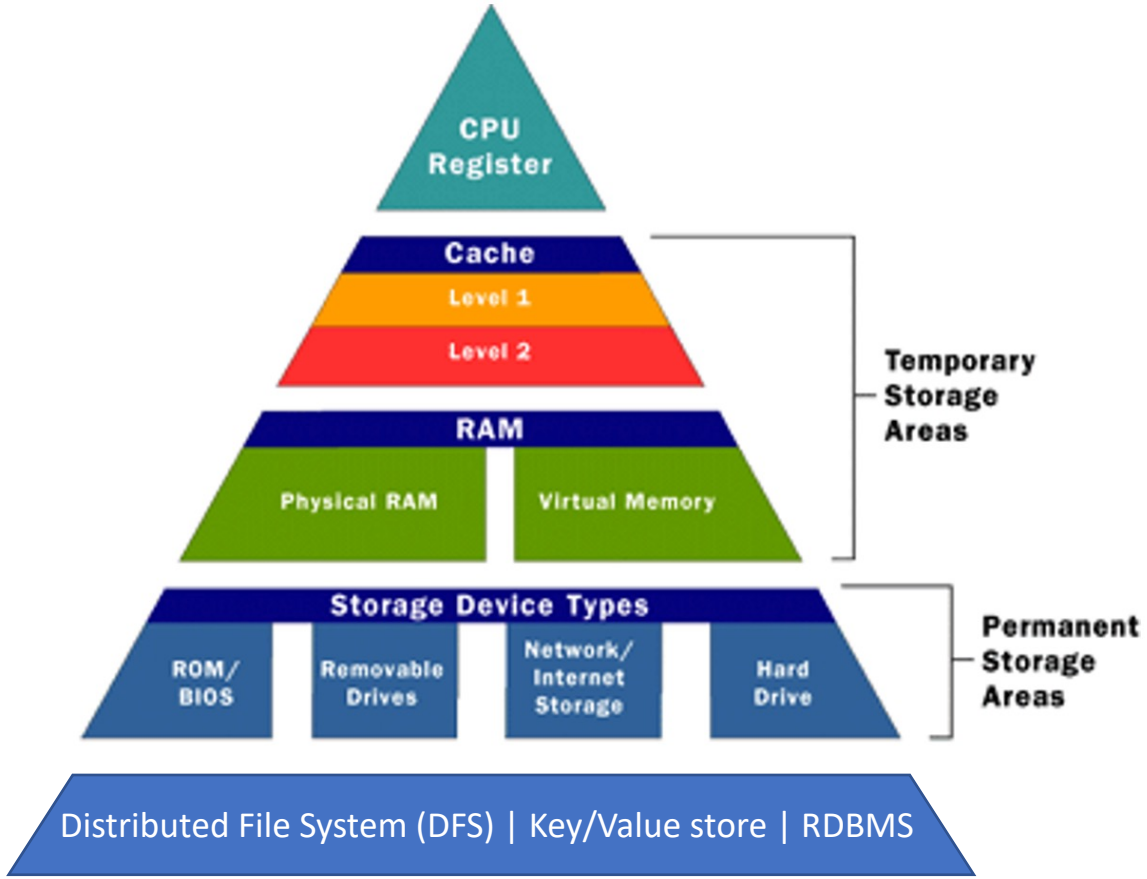
- *"Program"* == Web search, email, map/GIS, …
- *"Computer"* == 10,000's computers, storage, network
- Warehouse-sized facilities and workloads
- *Built from less reliable components than traditional datacenters*

Datacenter Networks



Storage Hierarchy

1. Registers and Cache
2. Main Memory (RAM)
3. Solid-State Drives (SSDs)
4. Hard Disk Drives (HDDs)
5. Network-Attached Storage (NAS) and Direct-Attached Storage (DAS)
6. Storage Area Network (SAN)
7. Data Center Storage



Datacenter Computing OS

- If the datacenter is the new computer
  - What is its **Operating System**?
  - Note that we are not talking about a host OS

Classical Operating Systems

- Data sharing
  - Inter-Process Communication, Remote Procedure Call, files, pipes, …
- Programming Abstractions
  - Libraries (libc), system calls, …
- Multiplexing of resources
  - Scheduling, virtual memory, file allocation/protection, …

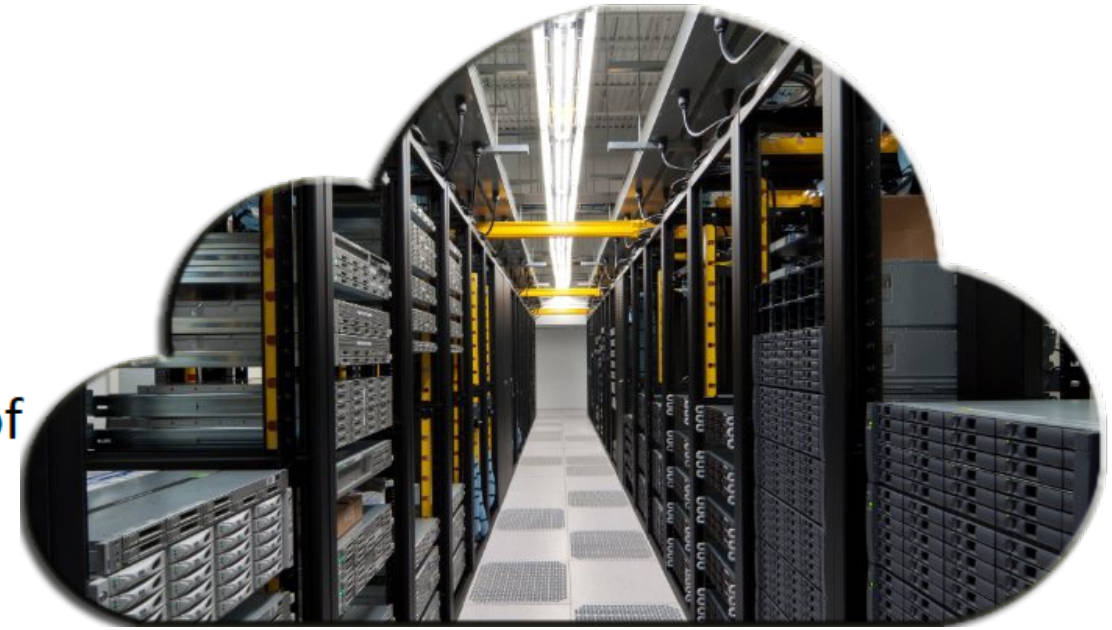
Datacenter Operating System

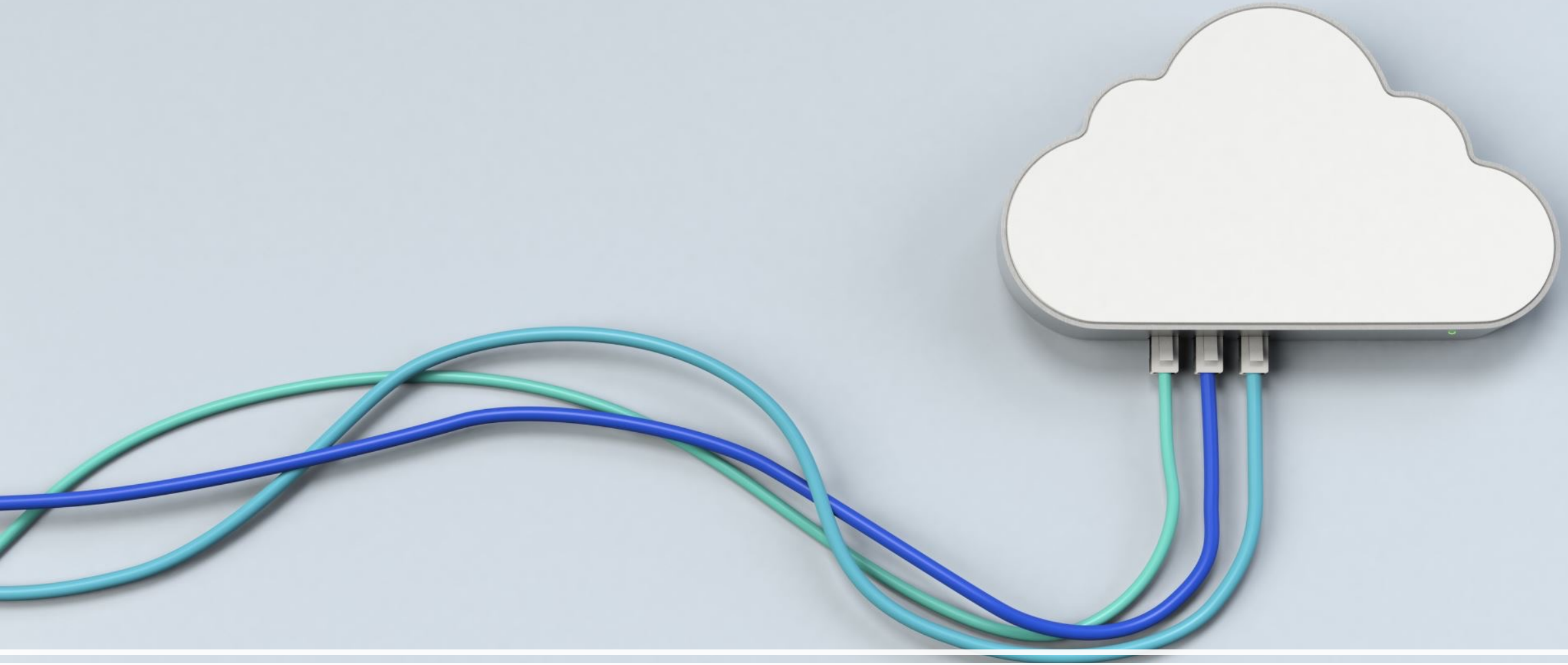
- Data sharing
  - Distributed File System, key/value stores
- Programming Abstractions
  - Google MapReduce, PIG, Hive, Spark
- Multiplexing of resources
  - Apache projects: Mesos, YARN (MRv2), ZooKeeper, DRF …

Data Centre Availability

- Datacenters have strict standards for reliability and availability
  - Tier 1: 99.671% Availability: 28 hours of downtime/year
  - Tier 2: 99.741% Availability: 22 hours of downtime/year
  - Tier 3: 99.982% Availability: 1.5 hours of downtime/year
  - Tier 4: 99.995% Availability: 26 minutes of downtime/year

R. Soundararajan, Datacenter 101, VMware, Inc.





Data Center Vs Cloud

Cloud Computing

- In 1960s the American computer scientist named John McCarthy stated that **computing will become a publicly available service in the future**. This is what has happened today and cloud computing has made it all possible.
- He claimed that computing might be sold in future in the same way as utilities are (electricity, water etc.).
- The first company to develop commercially successful cloud computing technology was Amazon.

Cloud Disclaimers

- "We've redefined Cloud Computing to include everything that we already do. I don't understand what we would do differently other than change the wording of some of our ads."
- "The computer industry is the only industry that is more fashion-driven than women's fashion," he said to a group of Oracle analysts.

Larry Ellison (Oracle CEO), 2008

What is Cloud Computing

- **Cloud computing** is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

NIST definition

National Institute of Standards and Technology

Cloud Attributes

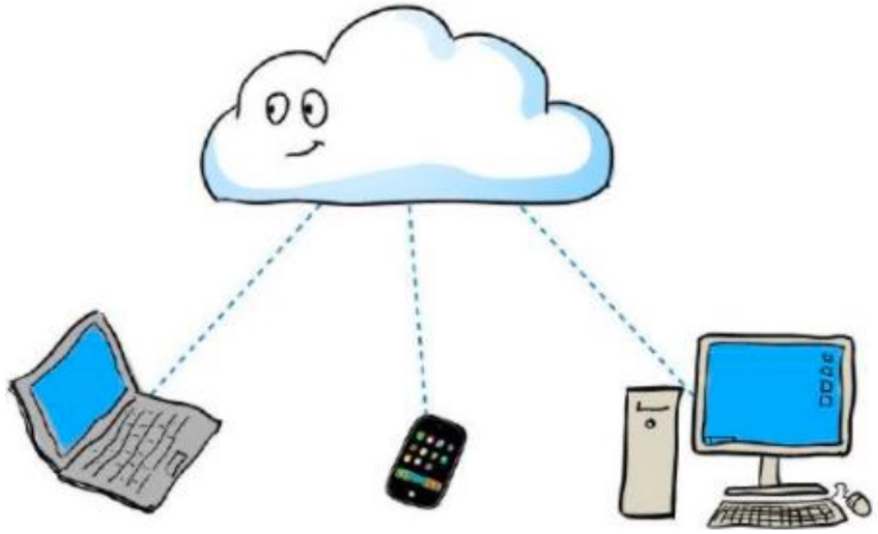
- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

On-demand Services

- Users are able to provision cloud computing resources without requiring human interaction, mostly done though a web-based self-service portal (management console).

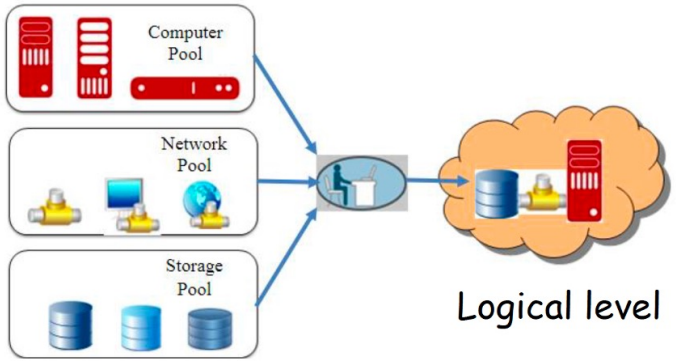
Broad Network Access

- Cloud computing resources are accessible over the network, supporting heterogeneous client platforms such as mobile devices, workstations, laptops, tablets, …



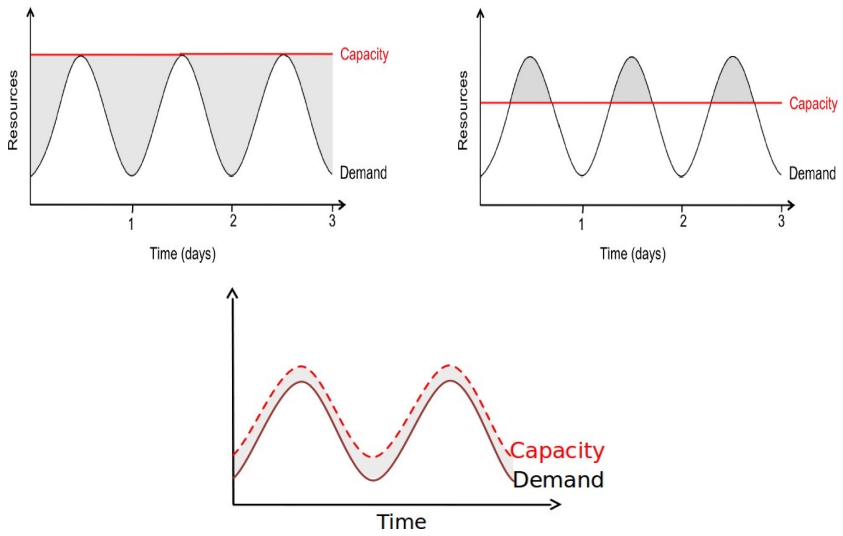
Resource Pooling

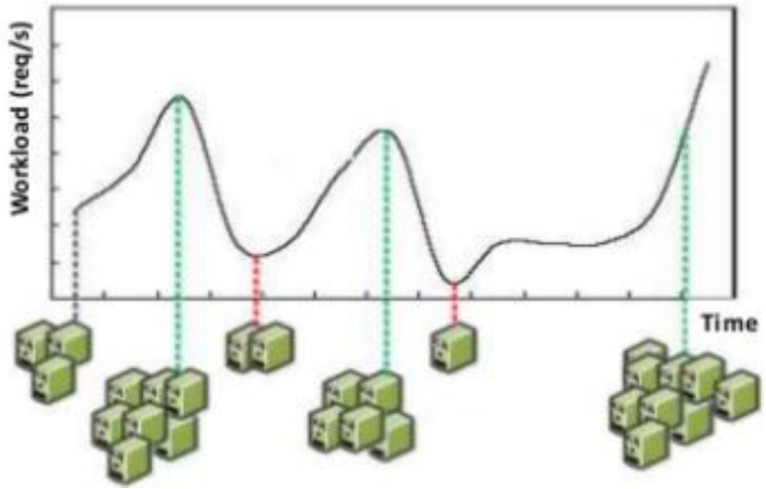
- Service multiple customers from the same physical resources, by securely separating the resources on logical level
- Computing resources:
  - Storage, processing, memory, network bandwidth and virtual machines
- Location independence
  - No control over the exact location of the resources



Rapid Elasticity

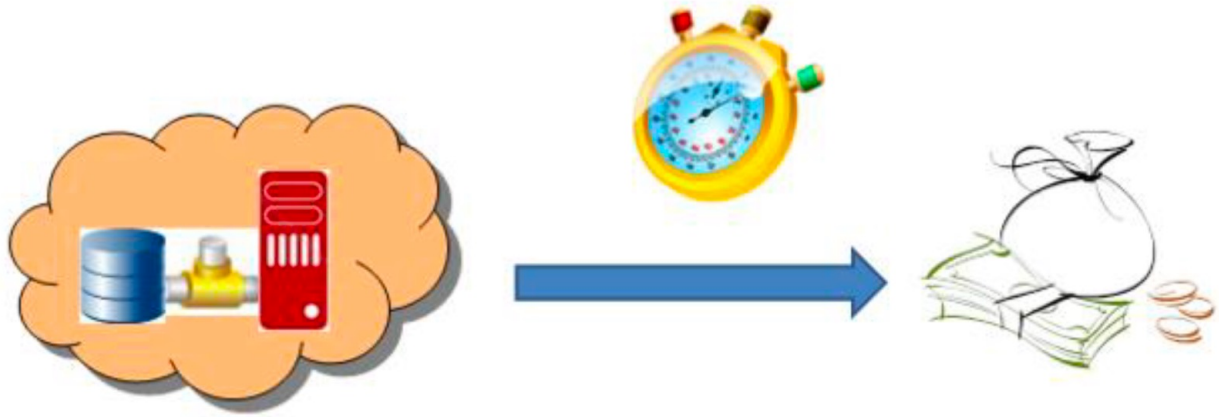
- The ability to scale resources both up and down as needed. To the consumer, the cloud appears to be infinite, and the consumer can purchase as much or as little computing power as they need.





Measured Service

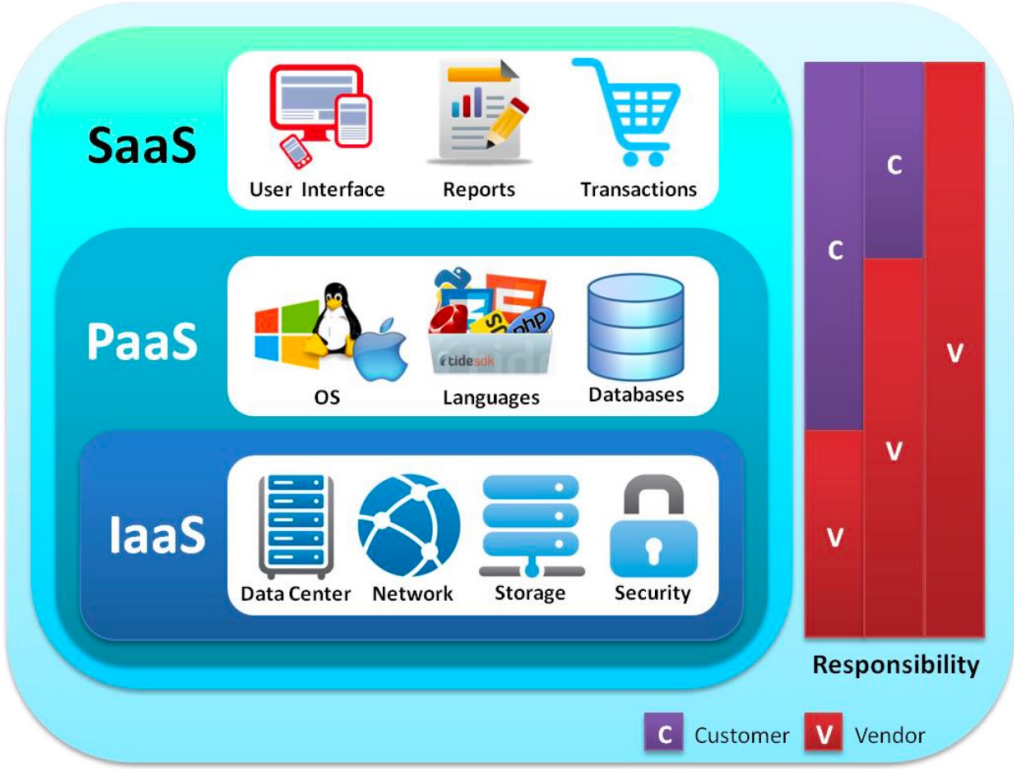
- Resource usage are monitored, measured, and reported (billed) transparently based on utilization.



Cloud Service Models

Cloud Service Models

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)



An Analogy

An Analogy

- Assume, you just moved to a city and you are looking for a place to live.
- What is your choice?
  - Built a new house?
  - Buy an empty house?
  - Live in a hotel?

Build a house

- Let's built a new house!
- You can fully control everything your like your new house to have.
- But that is a hard work.

Buy an empty house

- What if you buy an empty house?
- You can customize some part of your house.
- But never change the original architecture.

Live in a hotel

- How about live in a hotel?
- Live in a hotel will be a good idea if the only thing you care is enjoy your life.
- There is nothing you can do with the house except living in it.

Let's translate it to cloud computing

- Infrastructure as a Service (IaaS): similar to build a new house.
- Platform as a Service (PaaS): similar to buy an empty house.
- Software as a Service (SaaS): similar to live in a hotel.

Cloud Service Models

- IaaS:
  - hardware is provided by an external provider and managed for you
- PaaS:
  - in addition to hardware, your operating system layer is managed for you
- SaaS:
  - further to the above, an application layer is provided and managed for you - you won't see or have to worry about the first two layers

Software-as-a-Service (SaaS)

- Applications are supplied by the service provider.
- The user does not manage or control the underlying cloud infrastructure or individual application capabilities.
- Services offered include:
  - Enterprise services such as: workflow management, group-ware and collaborative, supply chain, communications, digital signature, customer relationship management (CRM), desktop software, financial management, geo-spatial, and search.
  - Web 2.0 applications such as: metadata management, social networking, blogs, wiki services, and portal services.
- Not suitable for real-time applications or for those where data is not allowed to be hosted externally.
- Examples: Gmail, Google search engine.

Platform-as-a-Service (PaaS)

- Allows a cloud user to deploy consumer-created or acquired applications using programming languages and tools supported by the service provider.
- The user:
  - Has control over the deployed applications and, possibly, application hosting environment configurations.
  - Does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage.
- Not particularly useful when:
  - The application must be portable.
  - Proprietary programming languages are used.
  - The hardware and software must be customized to improve the performance of the application.

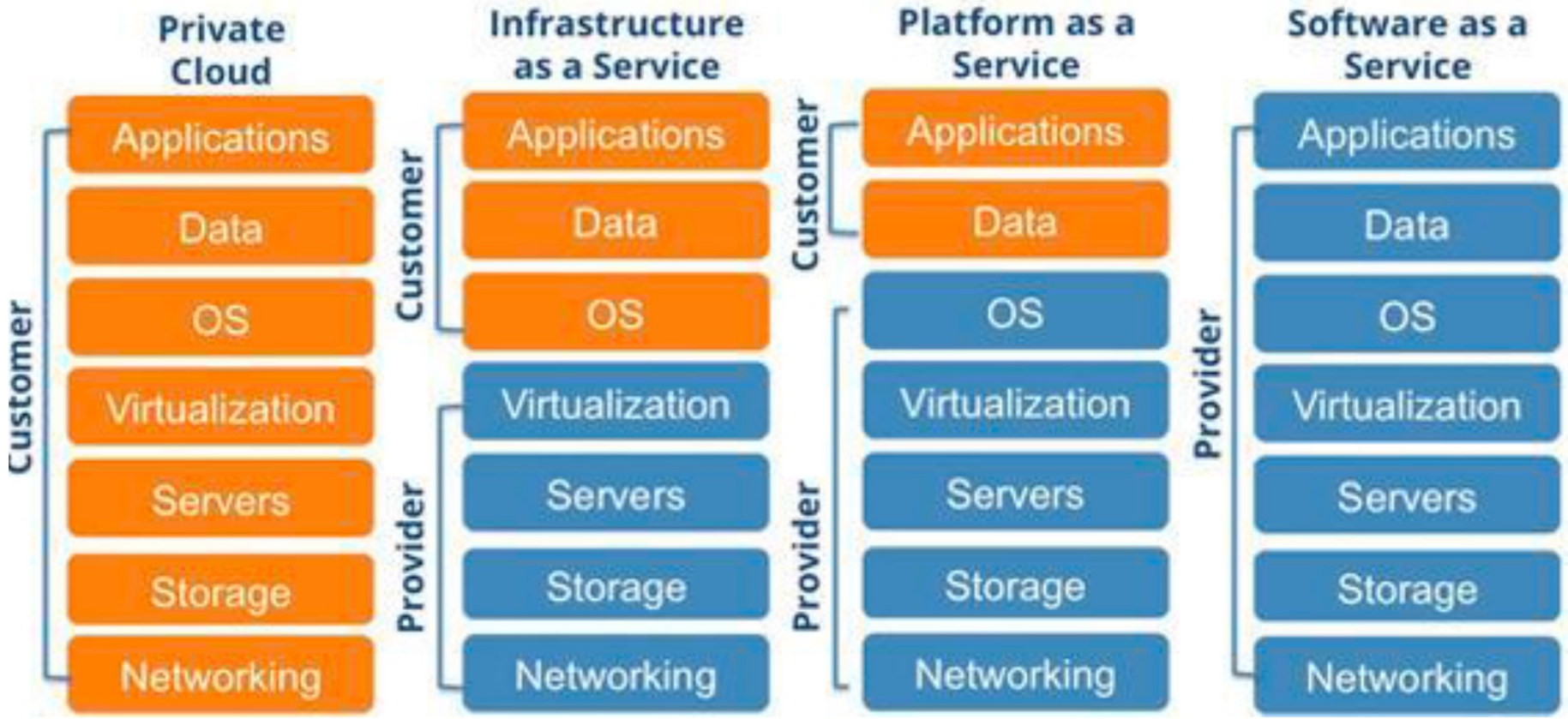
Infrastructure-as-a-Service (IaaS)

- The user is able to deploy and run arbitrary software, which can include operating systems and applications.

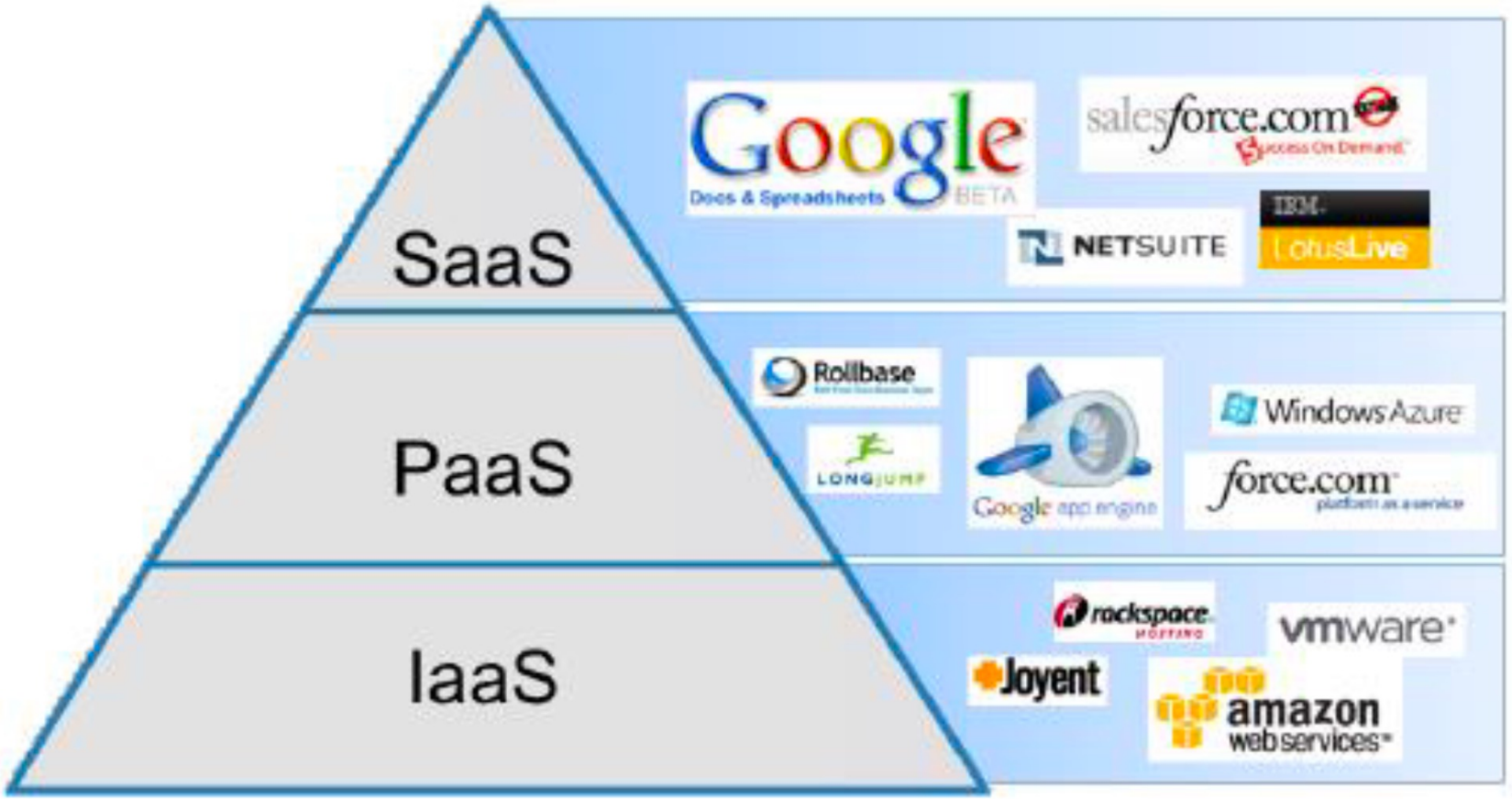
- The user does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of some networking components, e.g., host firewalls.

- Services offered by this delivery model include: server hosting, Web servers, storage, computing hardware, operating systems, virtual instances, load balancing, Internet access, and bandwidth provisioning.

Cloud Service Models



Cloud Service Model – Examples



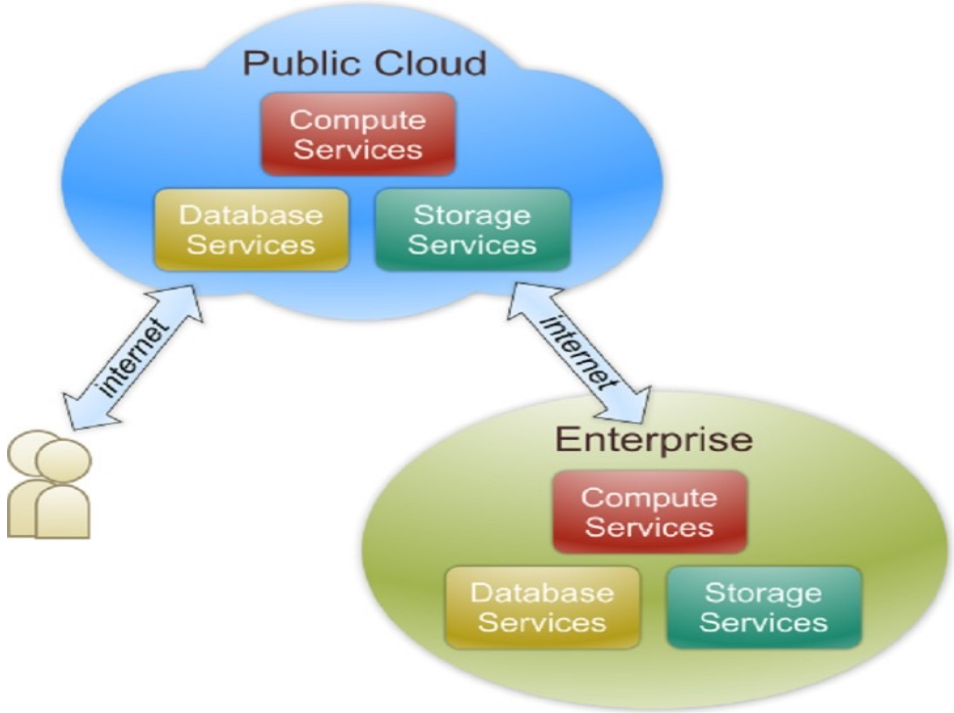
Cloud Deployment Models

Cloud Deployment models

- Public cloud
- Private cloud
- Hybrid cloud
- Community cloud

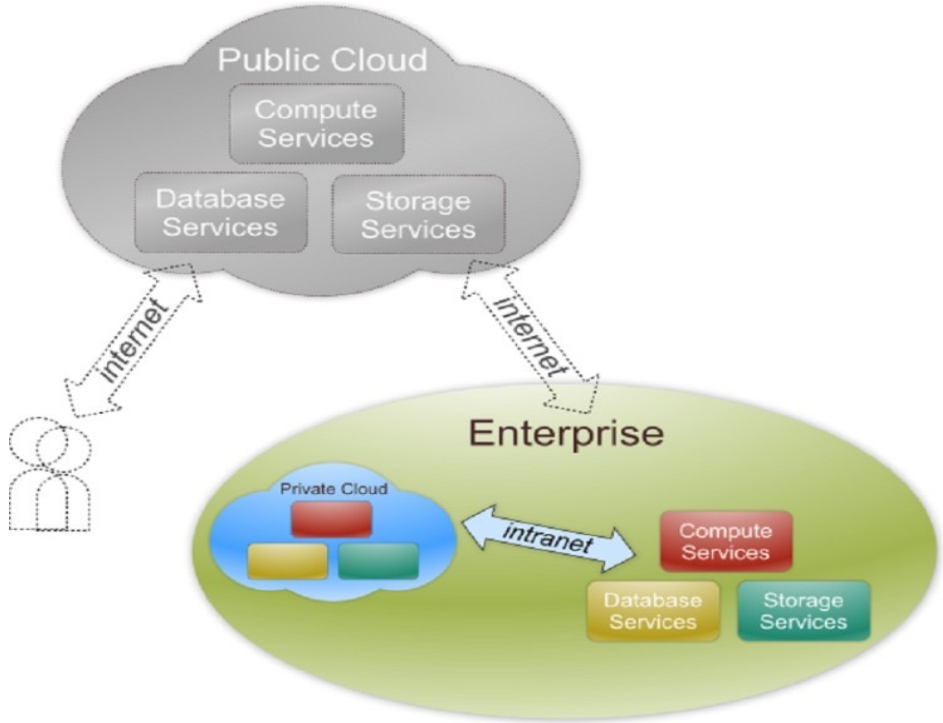
Public Cloud

- Infrastructure is made available to the general public.
- Owned by an organization selling cloud services.



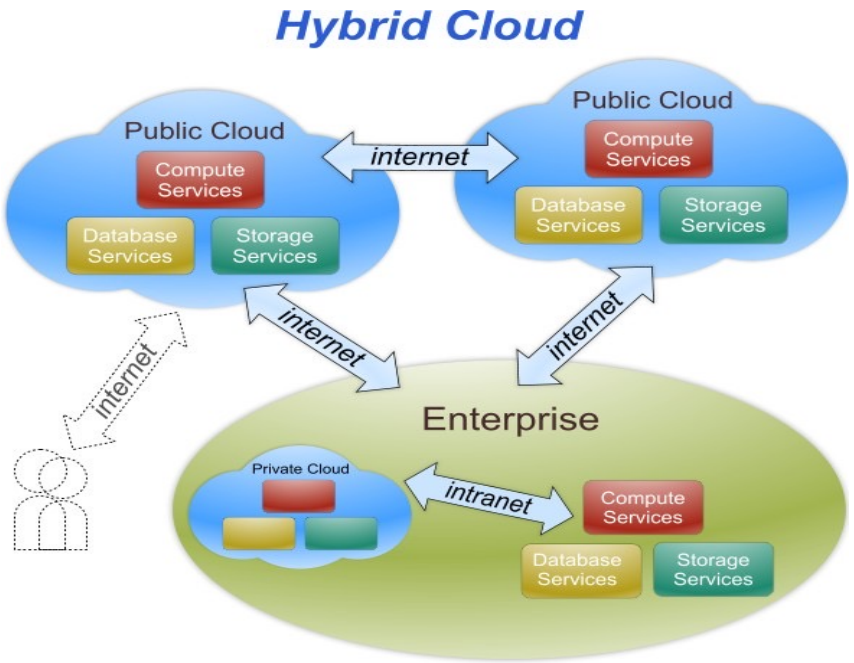
Private cloud

- Infrastructure is operated solely for an organization.
- Managed by the organization or by a third party.



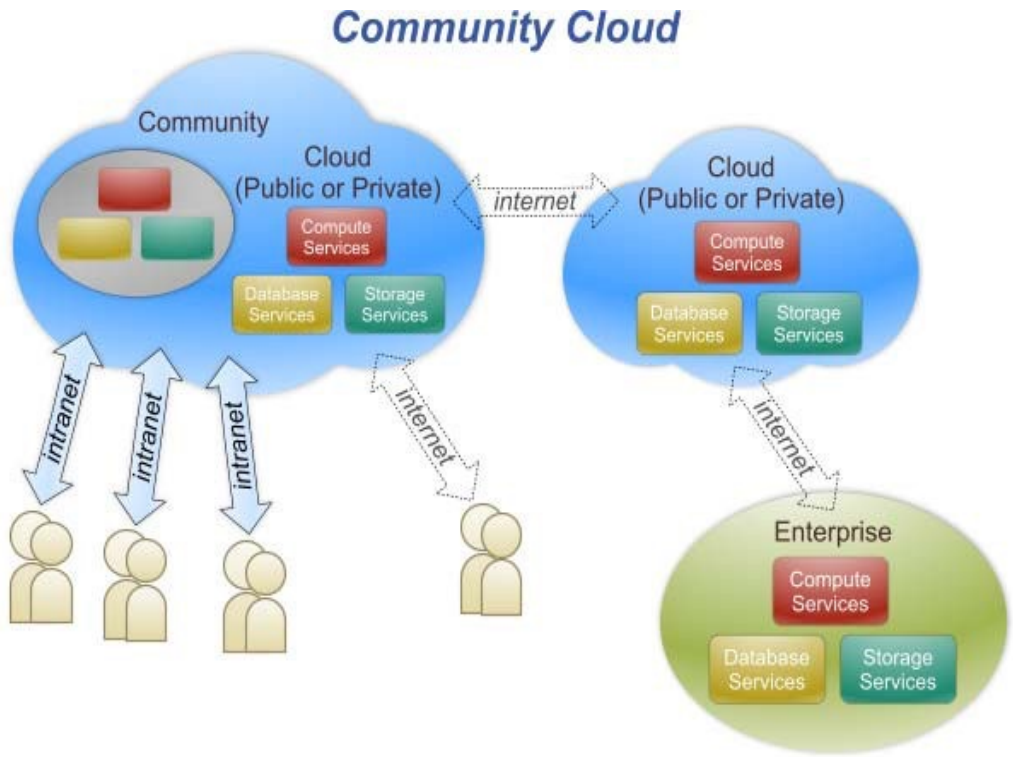
Hybrid Cloud

- Infrastructure is a composition of two or more clouds deployment models.
- Enables data and application portability.



Community Cloud

- Supports a specific community with common.
- Infrastructure is shared by several organizations.

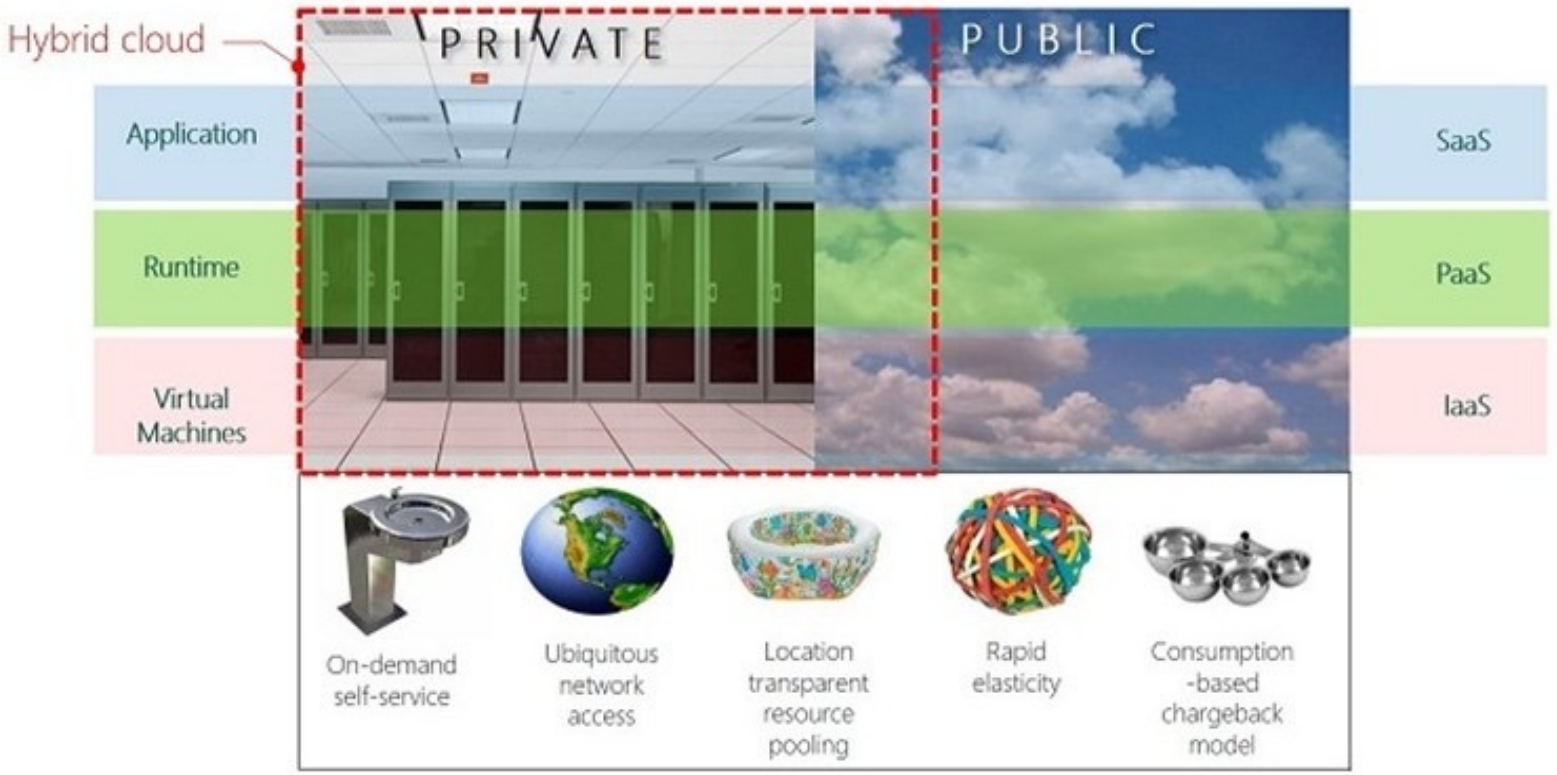


Cloud Main Services

- Computing (Virtual Machines, Containers, Serverless, …)
- Storage (file, block, object, …)
- Database (RDBMS, NoSQL, …)
- Big data analytics

Recap

- Why a computer is not enough?
- Data center as a computer
- Cloud Characteristics
- Cloud Service Models
- Cloud deployment Models



Next Topic:

Computing Services

Serverless and containers

Surveys

- Survey 1 – TA office hours
  - https://forms.gle/CXfrwXYrt5PGQn1y8
  - Questions/Discussions in Piazza

- Survey 2 – Team formation
  - https://forms.gle/NWfumogmLuzV3oNa7
  - Questions/Discussions in Piazza