COMP5349– Cloud Computing

Week 1: Introduction to Cloud Computing

Dr. Ying Zhou
School of Computer Science

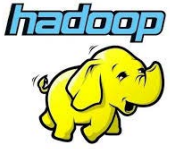
Outline

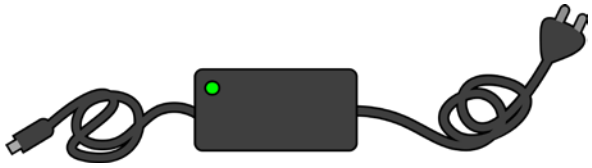
- **Cloud in layman term**
  - **A brief history of Cloud**
- **Cloud official definition**
- **The Role of Data Centers**

What is "Cloud"?

- Informally, we may view cloud computing as a way of ***renting*** IT resources
  - **Through Internet/Web**
  - Has an innovative way to specify, measure and charge the rented resources
  - Many other features...
- Not every kind of IT resources renting is called cloud
  - Lease from Dell to equip our labs
  - Rent some space from your ISP to set up a website
- Some forms of renting are closely related to Cloud and are considered as cloud's competitor

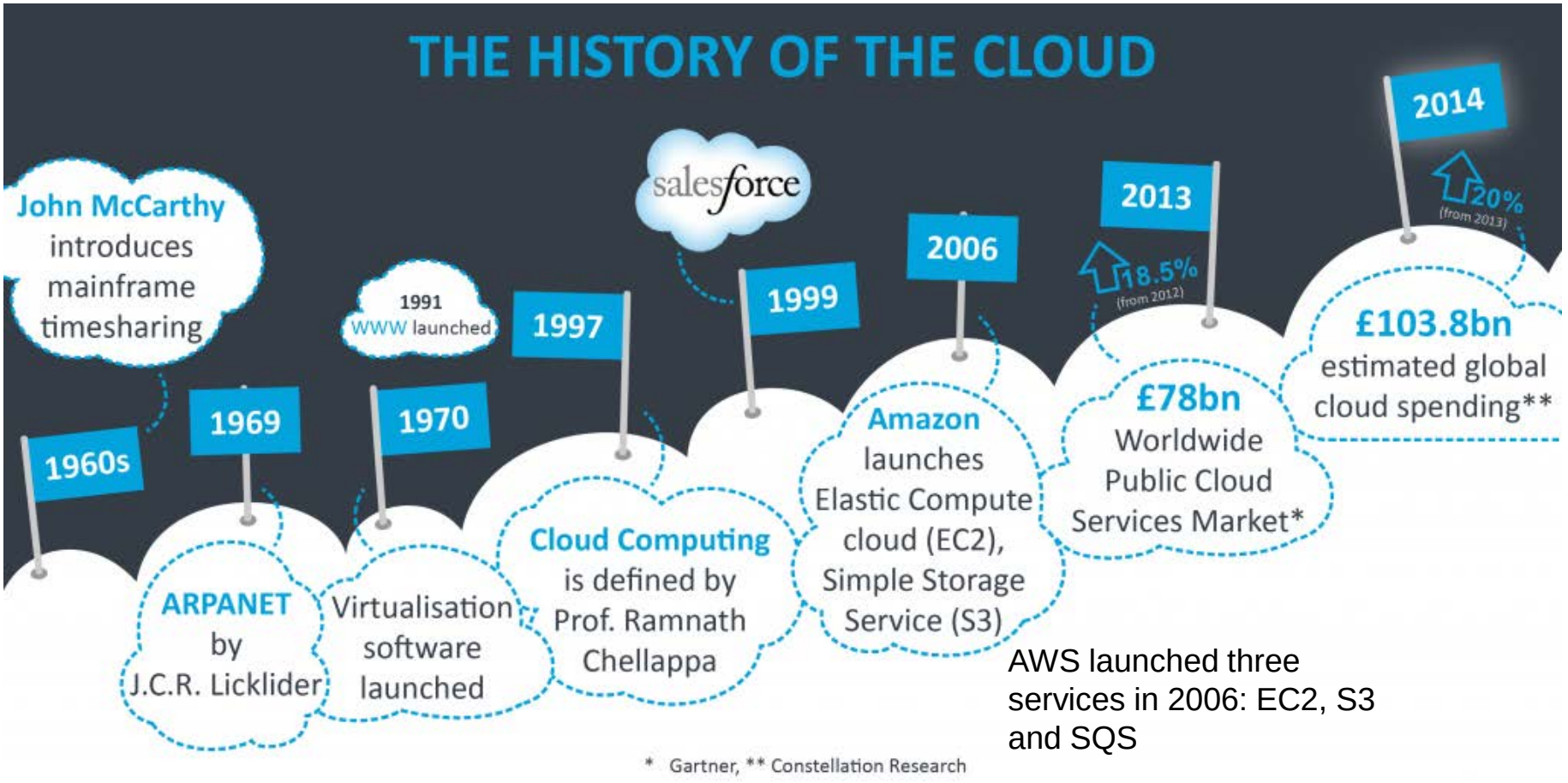
What are IT resources?





To run a renting business

- A way to package/measure/quantify your rental product/service
  - Leasing a Dell Desktop with certain specifications, each with different price tag
- A way to charge the customer
  - Hourly/daily/monthly/yearly rate
  - Subscription
- A way to deliver the produce/service
  - Truck, courier, pickup, or Internet
- A way to guarantee your product/service meet the client's requirements
- There are other forms of IT resources renting
- The particular form "Cloud" comes after supporting technologies are mature, and of course, good incentives for providers and market needs.



AWS launched three services in 2006: EC2, S3 and SQS

XML based SOAP web services was proposed in 1998

Amazon published a few SOAP bases services in 2002

Salesforce Now Live on Amazon Web Services Cloud Infrastructure in Australia (Oct. 2017)

https://timesofcloud.com/cloud-tutorial/history-and-vision-of-cloud-computing/

Other major players

- Microsoft
  - Microsoft Azure is officially released in 2010
  - It is now the second largest cloud provider with a market share less than half of the market share of AWS
- Google
  - Google cloud platform
  - Google App Engine was released in 2008 (an early PaaS service)
  - Google Cloud Platform was launched in 2011
- IBM
  - The latest one is called Bluemix and was released in 2014
  - IBM cloud has been through many other versions
  - IBM cloud is not that well recognized in general public

Outline

- Cloud in layman term
- **Cloud official definition**
  - **Various model**
  - **Specification, enabling technologies and pricing**
  - **Incentives from provider and consumer perspectives**
- The Role of Data Centers

Cloud Computing– a Broad Definition

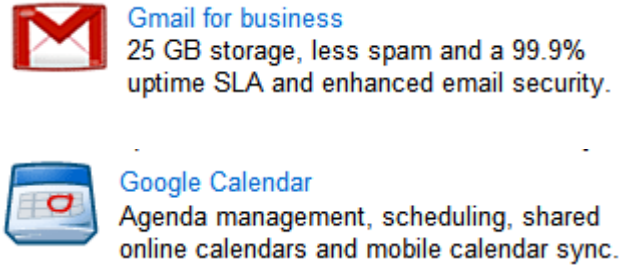
- A definition by US Governments' National Institute of Standard and Technology
  - "Cloud computing is a model for enabling ubiquitous, convenient, on-demand **network access** to a **shared pool** of **configurable computing resources** (e.g., *networks, servers, storage, applications*, *and services*)"
- In early days, we tend to differentiate three different models
  - Infrastructure as a Service (IaaS)
  - Platform as a Service (PaaS)
  - Software as a Service (SaaS)
- There are many new services and
  - Many providers are not restricted by a single service model
  - Many services cannot be categorized easily

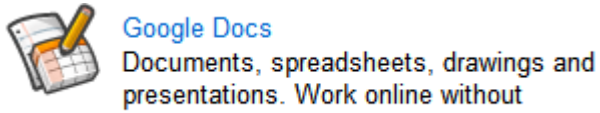
A Service Provider's offer: Azure

https://docs.microsoft.com/en-us/azure/architecture/aws-professional/services

SaaS Examples

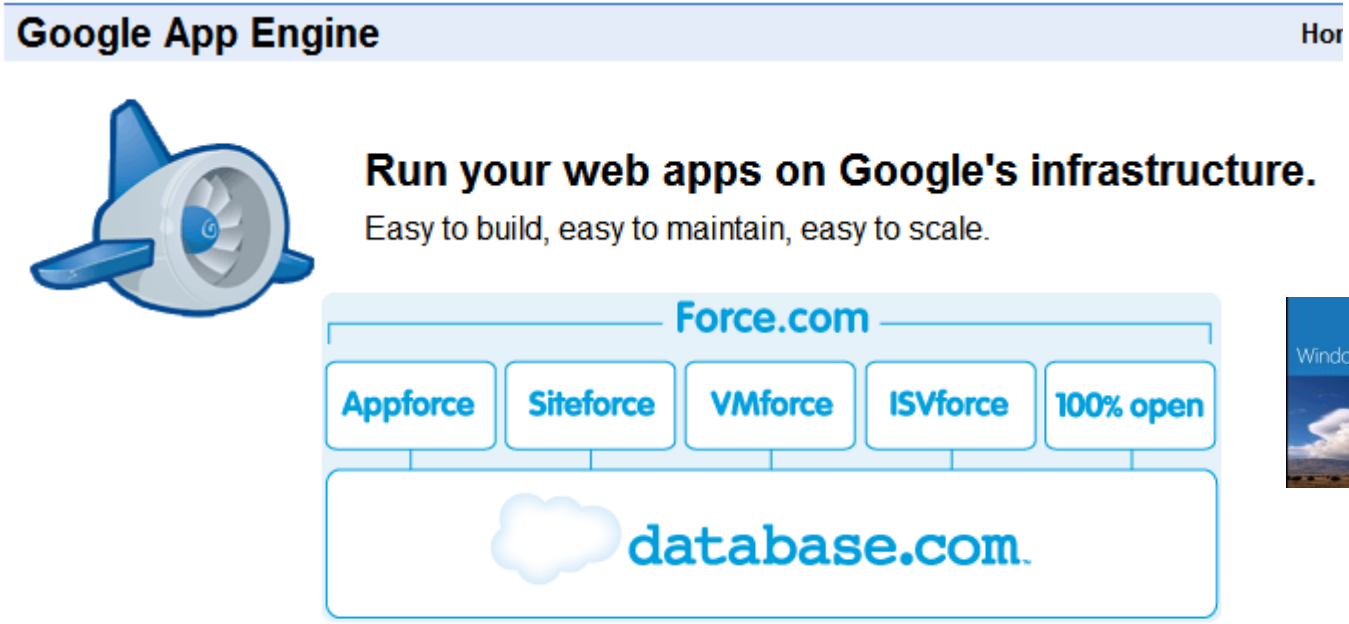
- **Software as a Service (SaaS):** The consumer uses an **application**, but does not control the operating system, hardware or network infrastructure on which it's running.
  - *Applications* are restricted to business applications or applications that may normally installed in a business network or personal computer
  - Examples
    - Business applications: CRM solutions from *salesforce*.com
    - Business/Personal applications: Gmail, Google Doc, etc.
- SaaS in many ways are different to the others and many times are not included when we say cloud

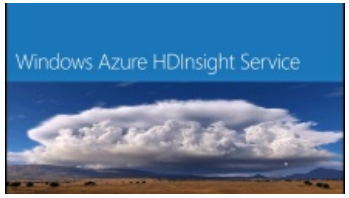




PaaS Examples

- **Platform as a Service (PaaS):** The consumer uses a **hosting environment** for their applications. The consumer controls the applications that run in the environment (and possibly has some control over the hosting environment), but does not control the operating system, hardware or network infrastructure on which they are running. The platform is typically an **application framework**.





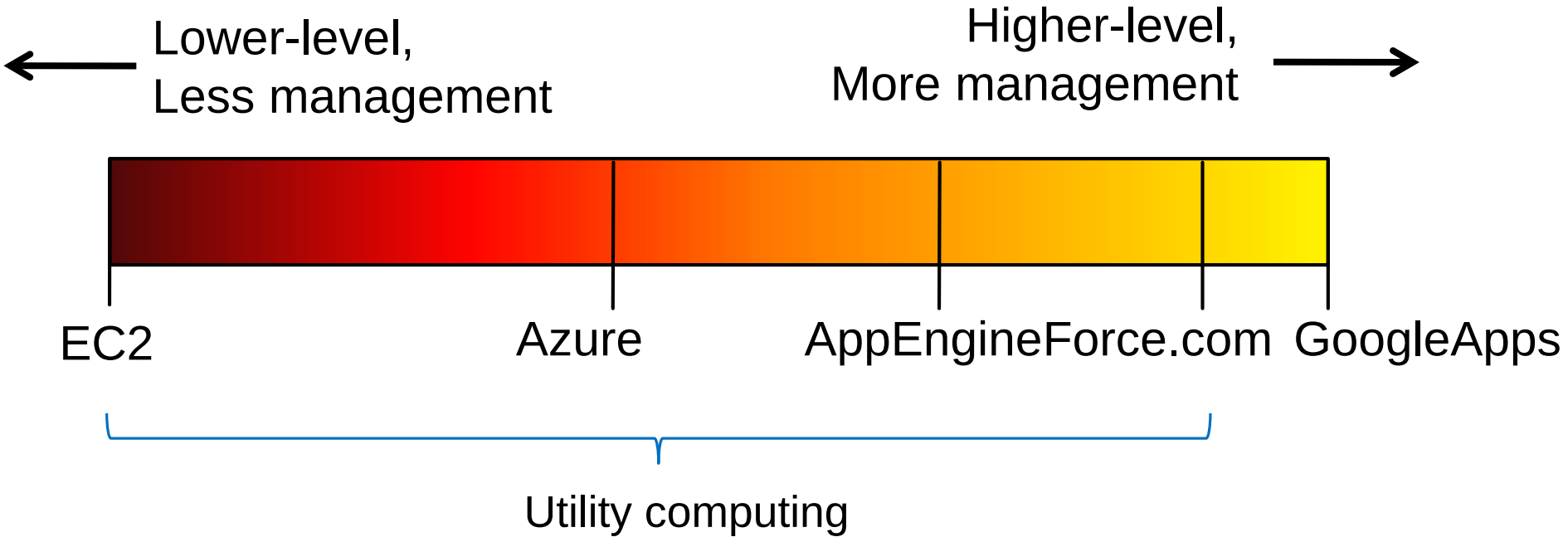
AWS Elastic MapReduce

IaaS Examples

- **Infrastructure as a Service (IaaS):** The consumer uses "**fundamental computing resources**" such as processing power, storage, networking components or middleware. The consumer can control the operating system, storage, deployed applications and possibly networking components such as firewalls and load balancers, but not the cloud infrastructure beneath them.

. . .

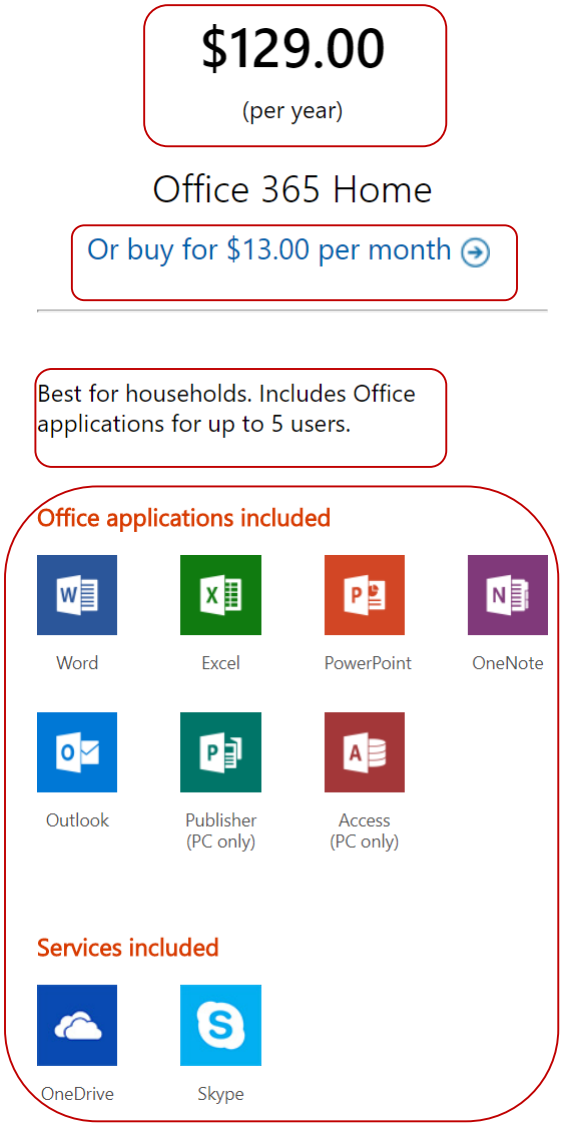
Spectrum of Cloud Services



From Berkeley Cloud presentation: http://berkeleyclouds.blogspot.com/

SaaS service specification and pricing

- SaaS
  - The service specification depends on the actual application, it could be the number of user account supported, the size of storage, etc
  - The pricing is usually subscription based, e.g. monthly or yearly price



https://products.office.com/en-au/compare-all-microsoft-office-products?tab=2 accessed 07/03/2018

IaaS Specification and Pricing

- IaaS
  - The specification is similar to the general spec when you purchase a computer. These include cpu speed, number of cores, memory, etc
  - At the beginning, most providers use fine grained pay-as-you-go hourly rate
  - Now many providers have even finer grained “Per Second Billing”[https://aws.amazon.com/ec2/pricing/ accessed 07/03/2018]

Model	GPUs	vCPU	Mem (GiB)	GPU Mem (GiB)	GPU P2P
p3.2xlarge	1	8	61	16	-
p3.8xlarge	4	32	244	64	NVLink
p3.16xlarge	8	64	488	128	NVLink

- Up to 8 NVIDIA Tesla V100 GPUs, each pairing 5,120 CUDA Cores and 640 Tensor Cores
- High frequency Intel Xeon E5-2686 v4 (Broadwell) processors

p3.2xlarge	$3.06 per Hour
p3.8xlarge	$12.24 per Hour
p3.16xlarge	$24.48 per Hour

NVIDIA Tesla V100 - GPU computing processor - Tesla V100 - 16 GB

Price:
$8,720.99

1 Add to Cart

https://aws.amazon.com/ec2/pricing/on-demand/
https://aws.amazon.com/ec2/instance-types/
http://www.zones.com/site/product/index.html?id=105374001
accessed 07/03/2018

Specification and Pricing

- PaaS
  - Somewhere in between, could be fine grained hourly rate or subscription based.
    - E.g. If you start a MapReduce cluster in Azure or AWS, you can specify how many node you want to have and the node type, you will be charged hourly (or secondly) based on those instances' price.

App Engine applications run as instances within the standard environment or the flexible environment.

Instances within the standard environment have access to a daily limit of resource usage that is provided at no charge defined by a set of quotas. Beyond that level, applications will incur charges as outlined below. To control your application costs, you can set a spending limit. To estimate costs for the standard environment, use the pricing calculator.

Sydney	
Instance class	Cost per hour per instance
B1	$0.068
B2	$0.135
B4	$0.270

https://cloud.google.com/appengine/pricing

Service packaging technologies

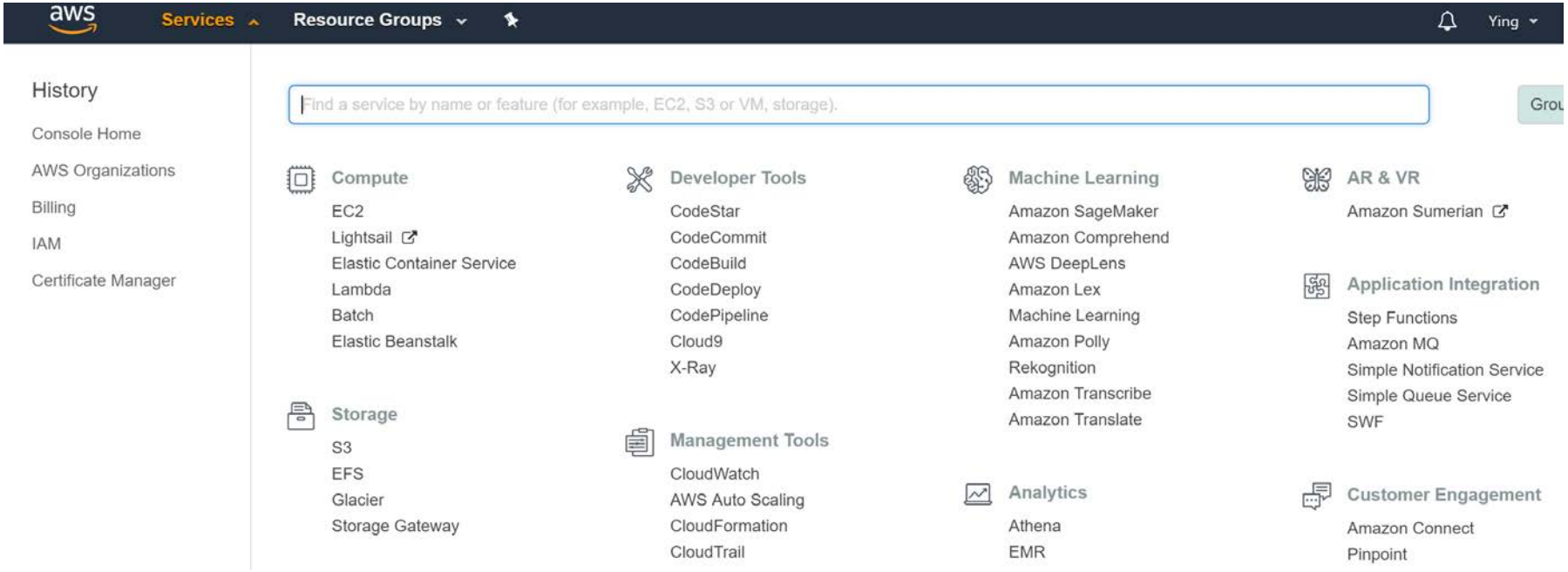
- IaaS
  - IaaS provides virtual machine together with storage and network as package
  - All providers use virtualization technology, the actual software used could be different
- PaaS
  - Virtualization technology can be used if the platform is presented as VM + some preinstalled software
  - Container technology may be used and the launching time could be greatly reduced
  - Provider may design their own software to let clients share an underlying platform
- SaaS
  - Again virtualization technology maybe used to provide each customer one or many VMs with preinstalled apps
  - Most providers write multi-tenancy based system to allow better resource utilization

Other services

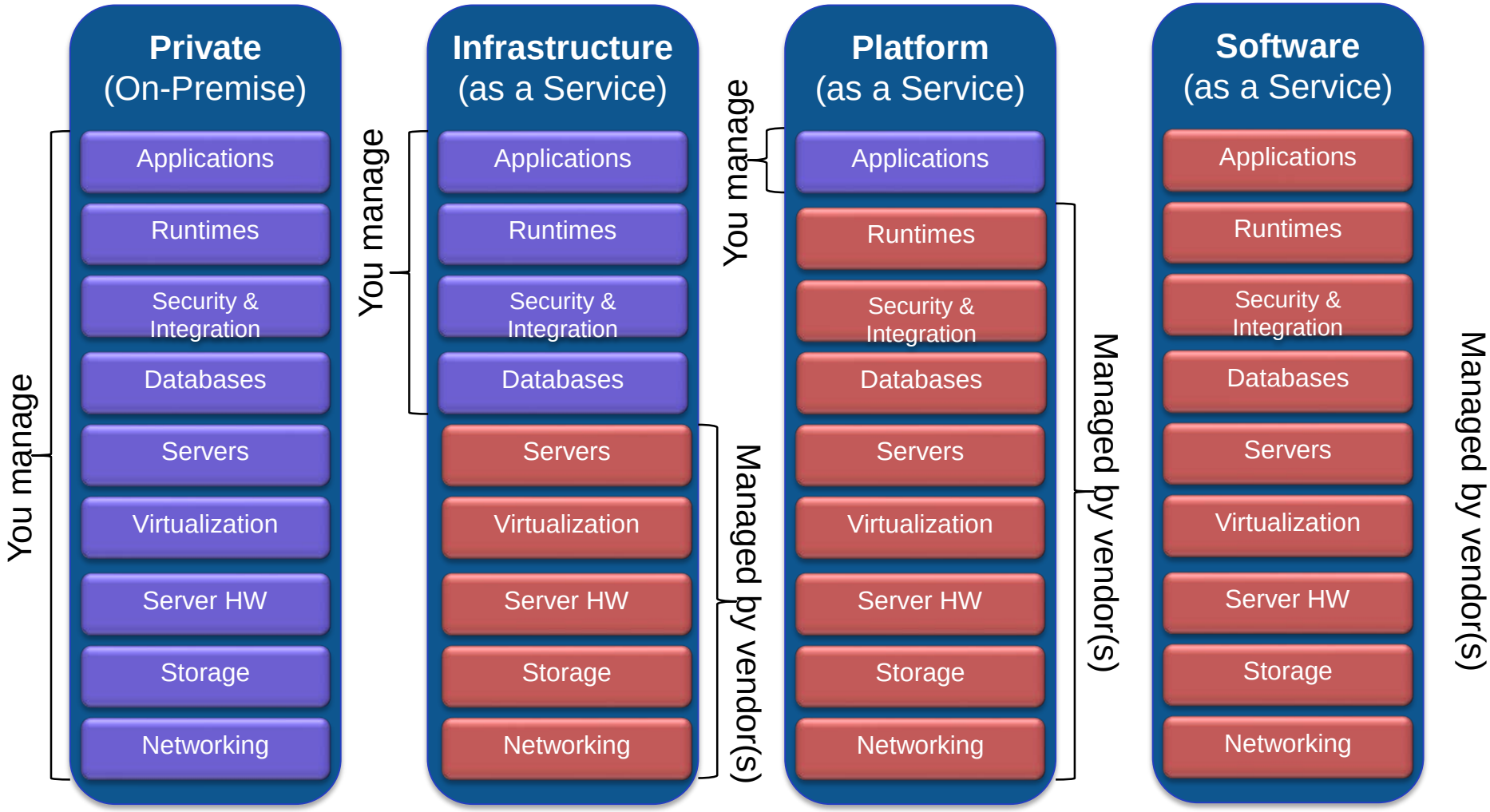
- Other services have their own way of describing, charging and enabling technologies
  - E.g. most storage services rely on company's own implementation of a planet scale storage system: Azure storage, DynamoDB, etc
  - Storage charging is more complicated as it has both the static and dynamic part
    - Actual storage size
    - Number of queries
    - Consistency and other quality requirement

Service Delivery

- All those XaaS models are delivered through Internet, with a web interface



Clouds Servicing Models



Using the Renting Terms

- All cloud serving models are based on the subject of renting
  - If I rent you the whole computer (the virtual version) with Operating System preinstalled, I am providing Infrastructure as a service
  - If I rend you some development environment and let you build your own application on top it, I am providing Platform as a service
  - If I rent you the whole application for you to use, I am providing Software as a service
- How do you determining what you want?
- Or, why, as a consumer, I should consider any such option?
- Or, why, as a provider, I should provide such services?

Incentives for Cloud Providers

- IT resource utilization is usually low
  - Server's cpu, memory, IO, networking are in idle state most of the time
  - If each person only uses one quarter of their server's computing power, why not provide a way to let four person sharing one computer?
- Economy of Scale
  - Most cloud providers are companies with large amount of IT facilities

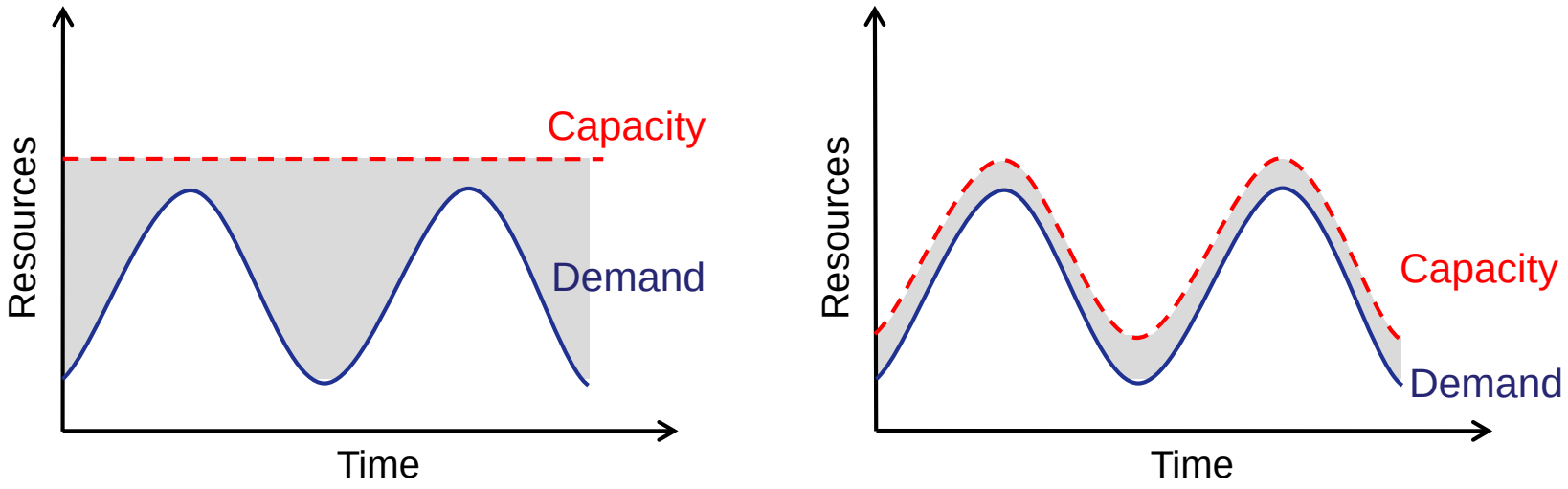
Resource	Cost in Medium DC	Cost in Very Large DC	Ratio
Network	$95 / Mbps / month	$13 / Mbps / month	7.1x
Storage	$2.20 / GB / month	$0.40 / GB / month	5.7x
Administration	≈140 servers/admin	>1000 servers/admin	7.1x

Incentives for Cloud Users

- Cloud user
  - Better **provisioning** through **elasticity** and **pay-as-you-go** and other fine-grained pricing models
  - Lift the burden of operational management
    - Installing certain software, e.g. a Hadoop cluster or GPU based tensorflow is time consuming and error prone
    - Upgrading existing software adds further headache
  - CapEx vs. OpEx tradeoff
- Example
  - **Netflix**: world's leading Internet subscription service for movies and TV shows
  - Netflix migrated from its own data centers to AWS in 2010
    - Capacity growth rate is accelerating, unpredictable
      - Year on year customer growth is 52%, year on year customers using streaming is up 145% (from ~4M to ~11M).
    - Product lunch spikes– iPhone, WII, PS3, XBox
    - Datacenter is large inflexible capital commitment

The Provisioning Problem

- It is very hard to predict usage and to provision sufficient capacities



Unused resources

From Berkeley Cloud presentation: http://berkeleyclouds.blogspot.com/

Elasticity and Pay-As-You-Go

- Elasticity
  - The cloud allows scaling up and scaling down of resource usage on an ‘as-needed’ basis. Elapsed time to increase or decrease usage is measured in seconds or minutes
- Pay-As-You-Go
  - Consumer is charged based on resourced they used (per instance or per cpu time), the charging unit has changed to hourly to secondly (actually minutely for most providers)

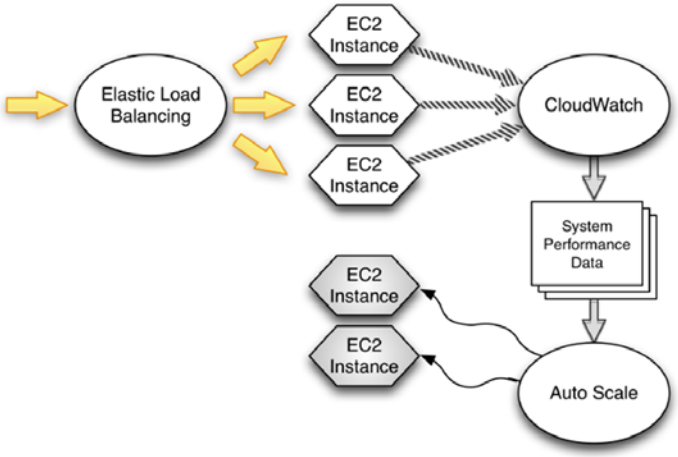
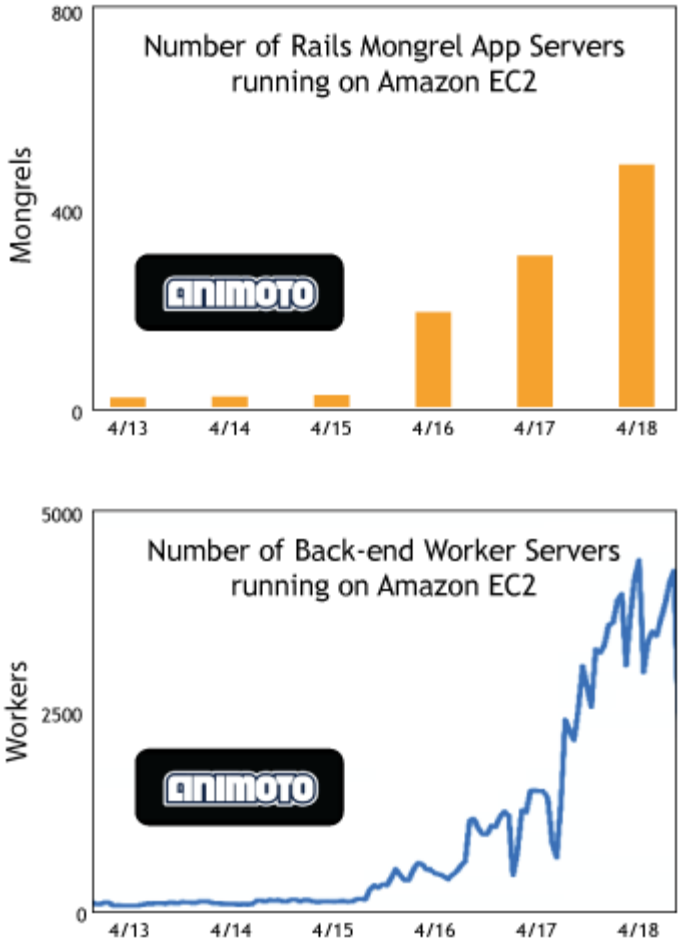


Figure 7.1. The relationship between elastic load balancing, CloudWatch, and auto scale

The famous Animoto Example



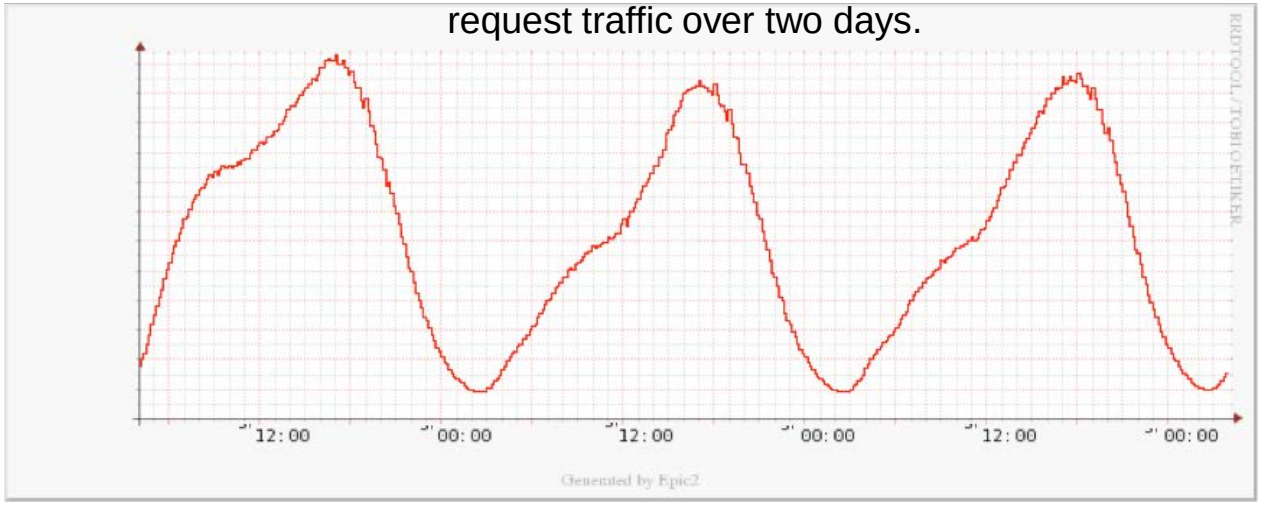
"They had 25,000 members on Monday, 50,000 on Tuesday, and 250,000 on Thursday. Their EC2 usage grew as well.

For the last month or so they had been using between 50 and 100 instances. On Tuesday their usage peaked at around 400, Wednesday it was 900, and then 3400 instances as of Friday morning."

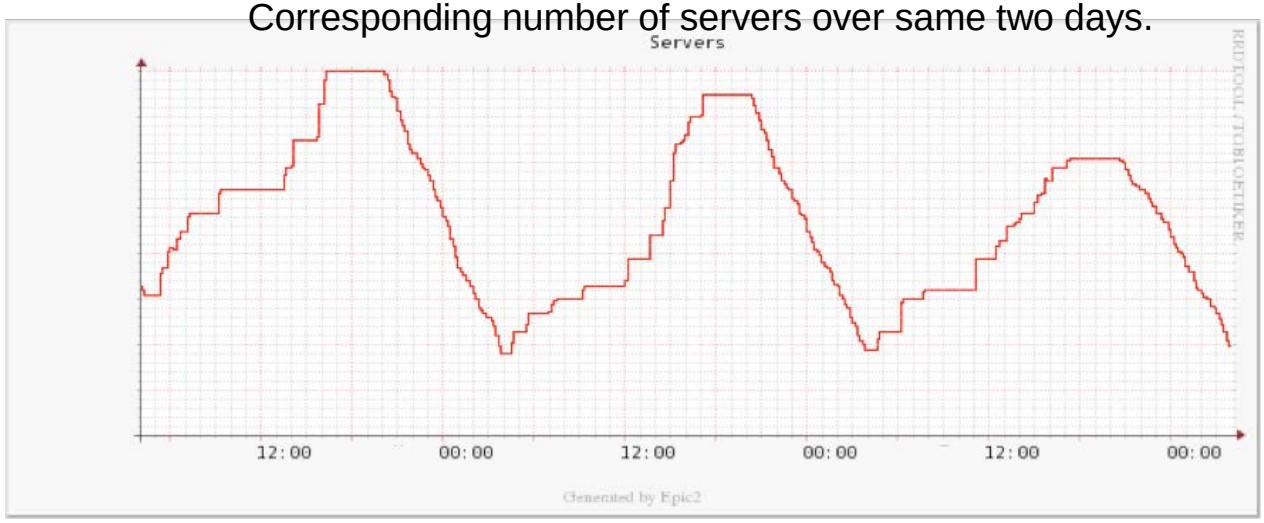
http://aws.typepad.com/aws/2008/04/animoto--scali.html

Netflix Auto-scaling Observations

request traffic over two days.

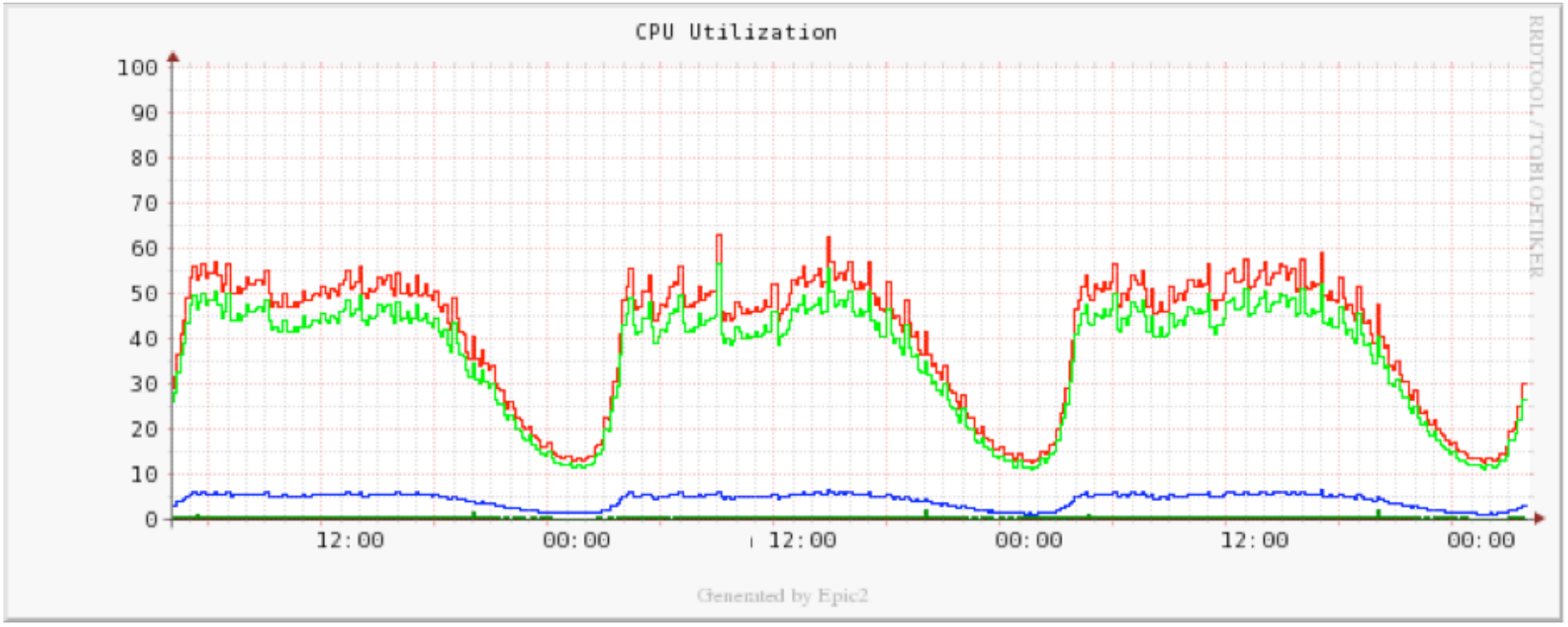


Corresponding number of servers over same two days.



The Netflix Tech Blog: Auto Scaling in Amazon cloud (Jan. 18, 2012)
http://techblog.netflix.com/search/label/autoscaling

Netflix Auto-scaling Observations (cont)



Aggregated CPU utilization during this time period:
Note that under load the aggregate CPU is essentially flat

The Netflix Tech Blog: Auto Scaling in Amazon cloud (Jan. 18, 2012)
http://techblog.netflix.com/search/label/autoscaling

Outline

- Cloud in layman term
- Cloud official definition
- **The Role of Data Centers**

Cloud is very physical

- All rentable IT resources reside in data center
- What is a data center?
  - “A building or portion of a building whose primary purpose is to house a computer room and its support areas” [TIA 942]
    - But a computer lab is not a data center, your garage with several machines is not a data center, even our server room is not a data center
- Large companies like Amazon, Google, Microsoft operated their daily business in data centers before cloud era.

Data center Fundamentals

- Typical components of a data center
  - IT Equipment
    - Server, Switches, Storage, etc.
    - The way to organize and manage those equipment change all the time
  - Facility Equipment
    - Power and cooling
  - Ancillary Systems
    - Access control, CCTV, fire alarm, Data Center Infrastructure Management Systems
- Depends on complexity and robustness of data centers, they are categorized into four different tiers
  - Tier 1 being the most basic one
  - Tier 4 being the most robust one

Data center main components

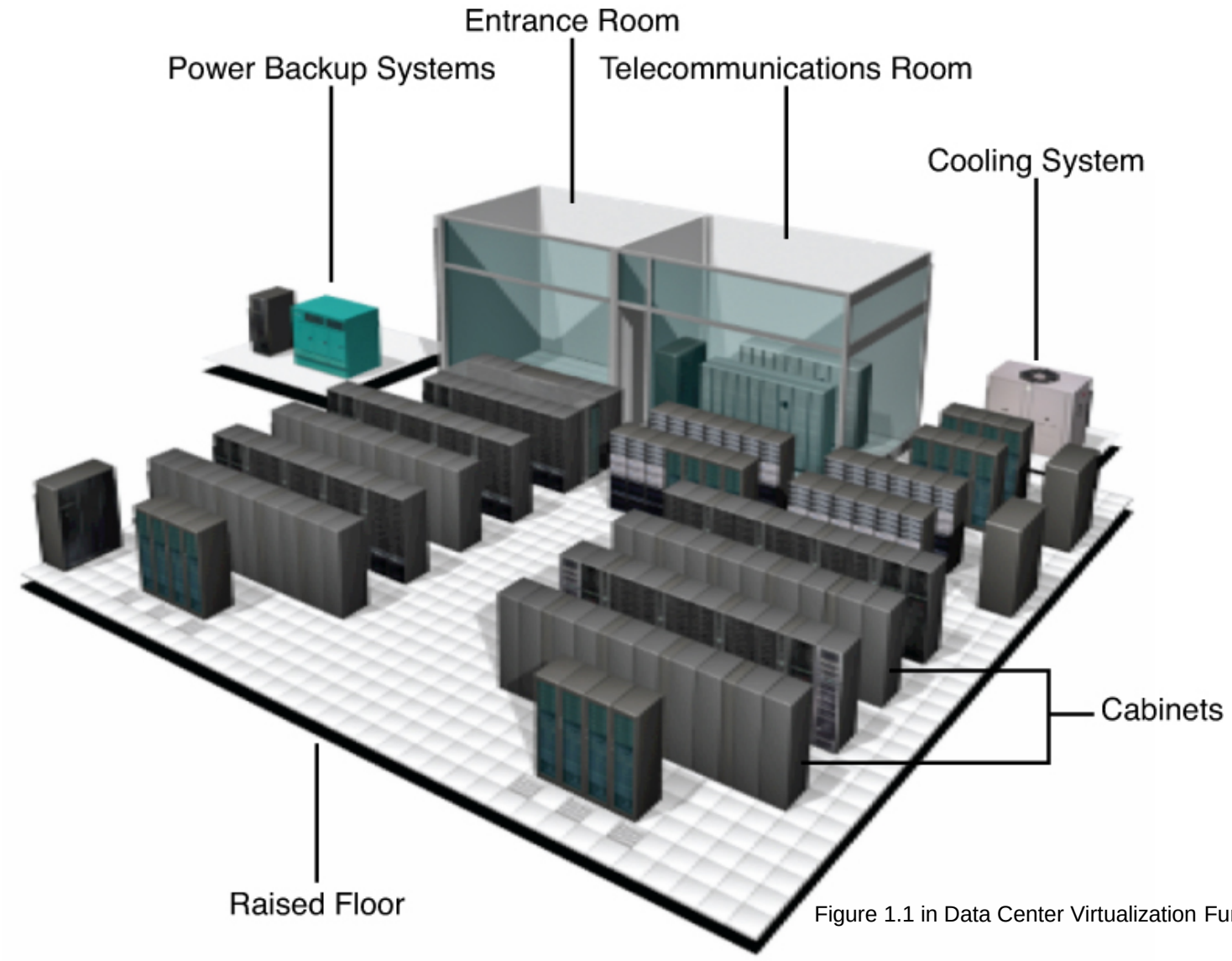


Figure 1.1 in Data Center Virtualization Fundamentals

The IT component

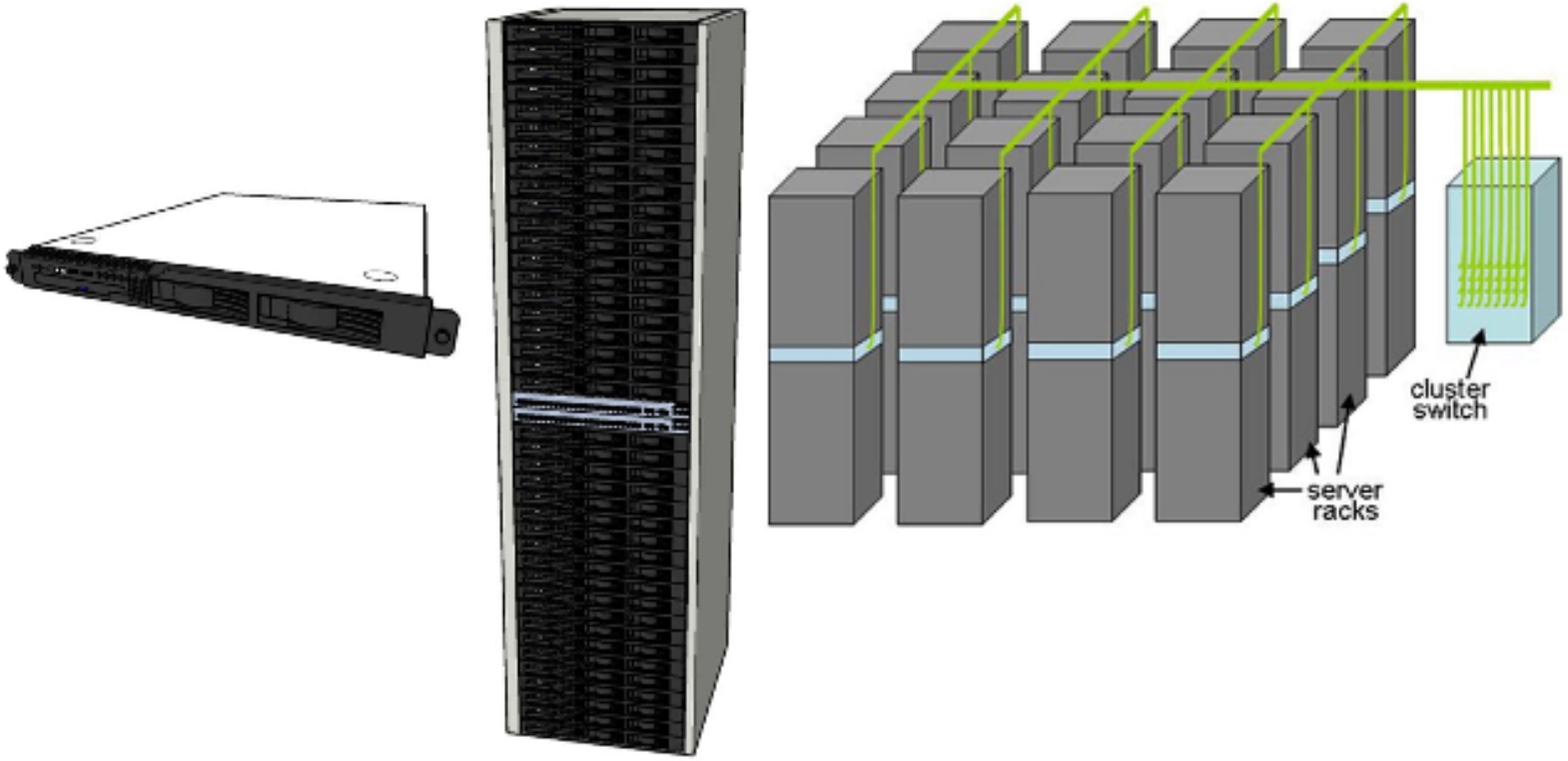


FIGURE 1.1: Typical elements in warehouse-scale systems: 1U server (left), 7′ rack with Ethernet switch (middle), and diagram of a small cluster with a cluster-level Ethernet switch/router (right).

http://www.morganclaypool.com/doi/abs/10.2200/S00516ED2V01Y201306CAC024

Typical Server Unit

Performance Consideration

- Traditionally: Computer Systems optimized for Performance (e.g. throughput)

- Nowadays, power is an increasingly important measure too
  - Costs of data center (DC) are dominated by **power** and **cooling** infrastructure
  - Carbon trading schemes will affect this even more
  - Hence DCs to optimize for *work done*/*Watt*

How is DC Energy Efficiency Measured?

- Commonly used metric for assessing data centers (DCs):
- **Power Usage Efficiency (PUE)**

$$\text{PUE} = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

- The average data center in the US has a PUE of 2.0 (Source: EPA - U.S. Environmental Protection Agency)
- According to James Hamilton, many have even PUE > 3.0
- High scale cloud services in the 1.2 to 1.5 range

Data Center Organization

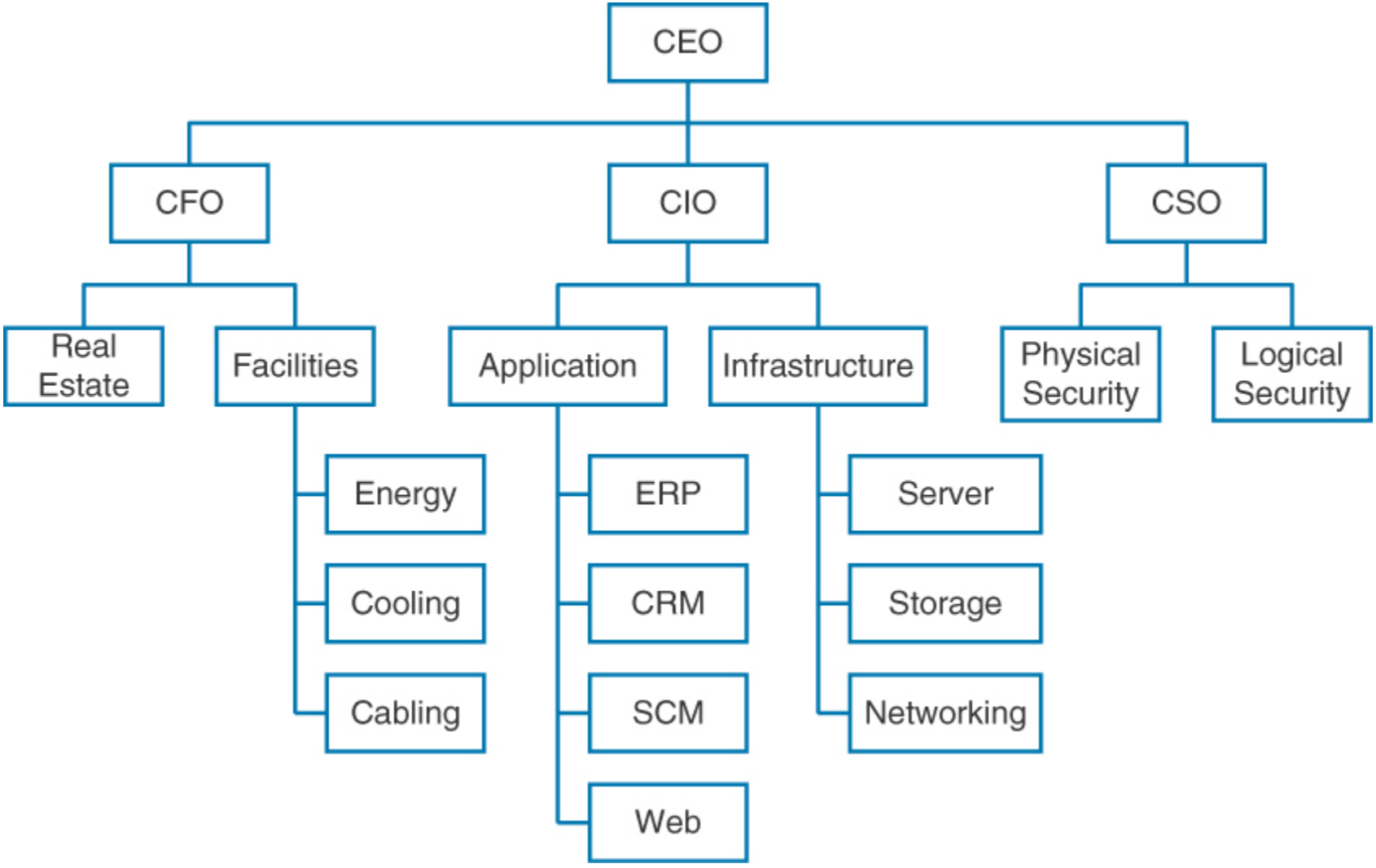


Figure 1.3 in Data Center Virtualization Fundamentals

Interdependent decision making

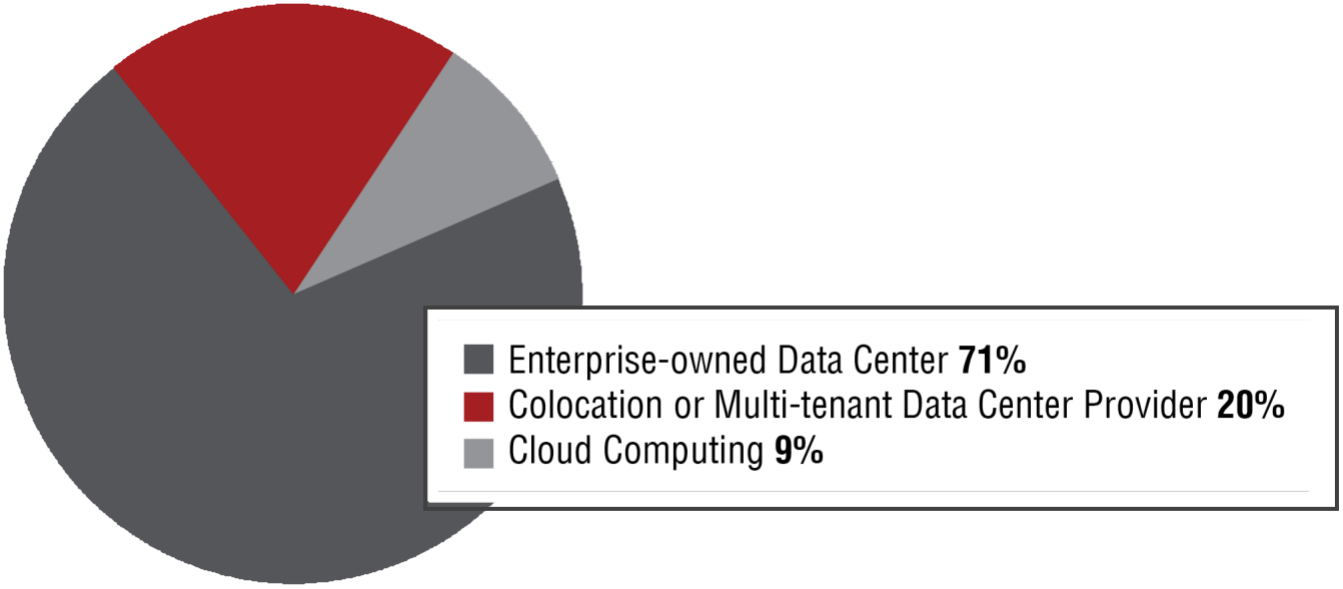
- The number of servers per cabinet depends on the power distribution design.
- A network design must be based on the knowledge of how many servers will the installed in each rack and how many interfaces they have.
- The network devices' physical position in the data center influences the structure's cabling project.
- Cabling can be laid out under the raised floor.
- The raised floor can have a direct influence over the cooling system, which is usually the highest contributor to the power system.

Technology life cycle

- Design decision depends highly on the life cycle of relevant component to ensure future evolution
- **Building:** 10 to 15 years
- **Cabling:** 7 to 10 years
- **Network:** 3 to 5 years
- **Storage:** 1 to 2 years
- **Server:** 6 to 18 months

Data Center Authorities

- Uptime Institute
- The Telecommunications Industry Association (TIA)



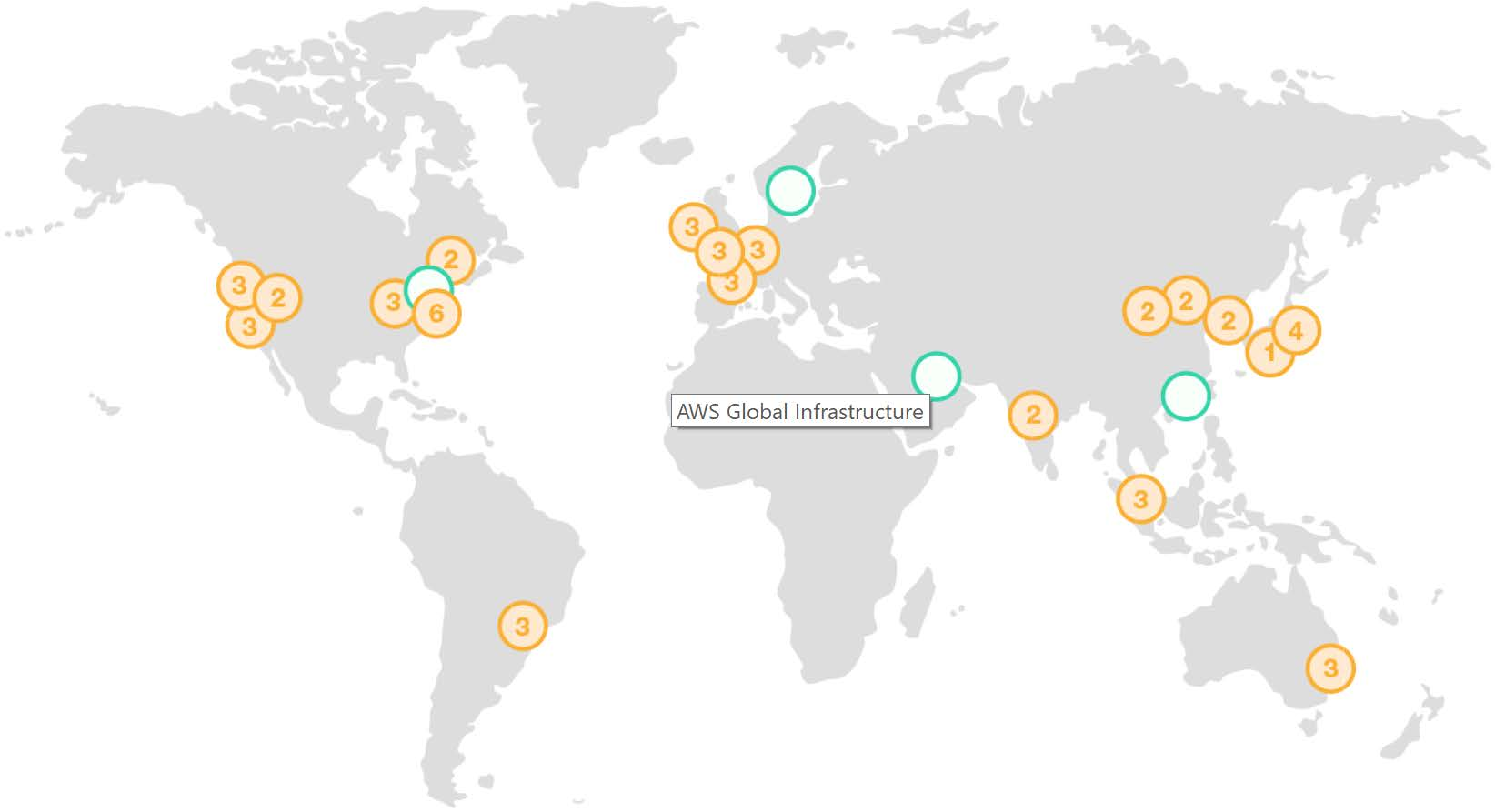
Uptime Institute Data Center Industry Survey results 2016

Data Center and Cloud

- Most cloud providers have their own data centers globally
- Some providers may use data centers owned by others through
  - Colocation provider, e.g. Equinox is a colo provider many cloud providers such as Azure, IBM and AWS rent space from
  - One type of cloud provider may user another provider's resources, e.g. Salesforce uses AWS in Australia
- "Colo is a type of data centre where equipment, space, and bandwidth are available for rental to retail customers."

  --[https://en.wikipedia.org/wiki/Colocation_centre]
  - Here equipment mostly refer to power, cooling and security ones
  - The actual IT equipment belongs to the company who rented the space, e.g. an entire floor

Amazon’s global infrstructure



Company or Organization's IT asset choices

- Enterprise owned data center
  - Private cloud
- Colo-data center
  - Private cloud
- Cloud
  - Private hosted cloud, e.g. Amazon VPC (Virtual Privte Cloud[https://aws.amazon.com/vpc/pricing/])
- "**Hybrid infrastructure models now the norm**"(Uptime Institute Data Center Industry Survey results 2016)
  - Our university is an example of using hybrid infrastructure model
  - We are big users of Microsoft and AWS
  - We also use colo providers

Summary

- **Cloud**
  - History and current status
  - Various cloud models
    - Description of the services
    - How to package the services
    - Pricing of the services
  - Data center as the physical part of cloud

Next Week: Homework

- **Homework for week 1**
  - **We prepared a self test homework to help you assess your Python programming skills**
  - **You may consider withdrawing from this unit if you have difficulty in completing it independently within a couple of hours.**

Next Week: Readings

- Gustavo A., A. Santana: **Data Center Virtualization Fundamentals: Understanding Techniques and Designs for Highly Efficient Data Centers with Cisco Nexus, UCS, MDS, and Beyond**
  - **Chapter 1, chapter 13**
- Matthew Portnoy: Virtualization Essentials, Second Edition
- Jim Smith (Author), Ravi Nair , **Virtual Machines: Versatile Platforms for Systems and Processes,** Morgan Kaufmann; 1 edition (June 17, 2005)
  - Chapter 1
  - Chapter 8 System Virtual Machines
- *Xen and the Art of Virtualization*. Paul Barham, Boris Dragovic, Keir Fraser, Steven Hand, Tim Harris, Alex Ho, Rolf Neugebauer, Ian Pratt, and Andrew Warfield. In Proceedings of the Nineteenth ACM Symposium on Operating systems principles (SOSP '03), 2003.

Reference

- Armbrustet al: **Above the Clouds: A Berkeley View of Cloud Computing**, TR EECS-2009-28, UC Berkeley, 2009.
- Gustavo A., A. Santana: **Data Center Virtualization Fundamentals: Understanding Techniques and Designs for Highly Efficient Data Centers with Cisco Nexus, UCS, MDS, and Beyond**
  - **Chapter 1**