

Cloud Computing

#1 - Introduction

This Course

- Understand what cloud computing is and what the fundamental technologies are
- Get an overview of the current service providers and software projects
- Understand how data centers are designed and cloud services developed and deployed
- The end game: ability to formulate relevant research questions and assess ongoing research activities

This Course

- No book
- Mix of overview lectures and deep dive presentations by the participants
- Requirements:
 - Attendance
 - One presentation on a selected topic
 - Home assignments, small project, no exam.
- 7.5 credits
- Work in progress

Session Structure

- Course divided in sessions on different technologies that plays a major role in cloud
- Each session is divided into 1+N parts
 - Overview by Jorn or Johan (30 min)
 - #N Presentations by you (30 min each)
- Your presentations are done in collaboration with the session leader
 - Draft version to be submitted one week *before* the session take place
 - And you present *your* presentation

Session Assignments

- Distributed Systems 1 [jj] 3/3
 - Johan (TBD), Hassan (TBD)
- Distributed Systems 2 [jj] 10/3
 - Manfred (DHT), Victor (Paxos), Antonio (TBD)
- Datacenter Networking [je] 17/3
 - Harald (SDN), Torgny (vSwitch)
- Virtualization [je] 24/3
 - Linus (Containers & Hypervisors), Robban (HW virtualization)
- Storage [jj]: 31/3
 - William (GFS), Christopher (TBD)
- Datacenter OS & Applications [je]: 7/4
 - Ola (REST & SOA), Jonas (Resource Management)
- Programming Models [jj]: 14/4
 - Jens Andersson (Big Data prog.), Per (Big data problems: social graph, page rank, spam detection), Mehmet (TBD)
- Datacenter Security [je]: 21/4
 - Patrik Lantz, Joakim Persson

What is Cloud Computing?

National Institute of Standards and Technology (NIST)

- “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. This cloud model is composed of five essential characteristics, three service models, and four deployment models.”
- Essential Characteristics:
 - On-demand self-service
 - Broad network access
 - Resource pooling
 - Rapid elasticity
 - Measured service

Essential Characteristics (NIST)

- **On-demand self-service.** A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
- **Broad network access.** Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).
- **Resource pooling.** The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.
- **Rapid elasticity.** Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.
- **Measured service.** Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

Deployment Models (NIST)

- Private cloud
- Public cloud
- Hybrid cloud
- (Community cloud)

Deployment Models (NIST)

- **Private cloud.** The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.
- **Community cloud.** The cloud infrastructure is 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). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.
- **Public cloud.** The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.
- **Hybrid cloud.** The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

Deployment Models (NIST)

- **Private cloud.** The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.
- **Community cloud.** The cloud infrastructure is 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). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.
- **Public cloud.** The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.
- **Hybrid cloud.** The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

Deployment Models (NIST)

- **Private cloud.** The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.
- **Community cloud.** The cloud infrastructure is 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). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.
- **Public cloud.** The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.
- **Hybrid cloud.** The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

Service Models (NIST)

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

Service Models (NIST)

- **Software as a Service (SaaS)** The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.
- **Platform as a Service (PaaS)** The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider.³ The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.
- **Infrastructure as a Service (IaaS)** The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer 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).

What is Cloud Computing?

- The Guardian, Sept. 29, 2008
 - Richard Stallman, Free Software Foundation: “It’s worse than stupidity: it’s marketing hype. Somebody is saying this is inevitable - and whenever you hear that, it’s very likely to be a set of businesses campaigning to make it true.”
- Wall Street Journal, Sept. 26, 2008
 - Larry Ellison, CEO, Oracle: “The interesting thing about Cloud Computing is that we've redefined Cloud Computing to include everything that we already do.... I don't understand what we would do differently in the light of Cloud Computing other than change the wording of some of our ads.”

Utility Computing

“ If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility... The computer utility could become the basis of a new and important industry. ”

—John McCarthy, speaking at the MIT Centennial in 1961^[2]

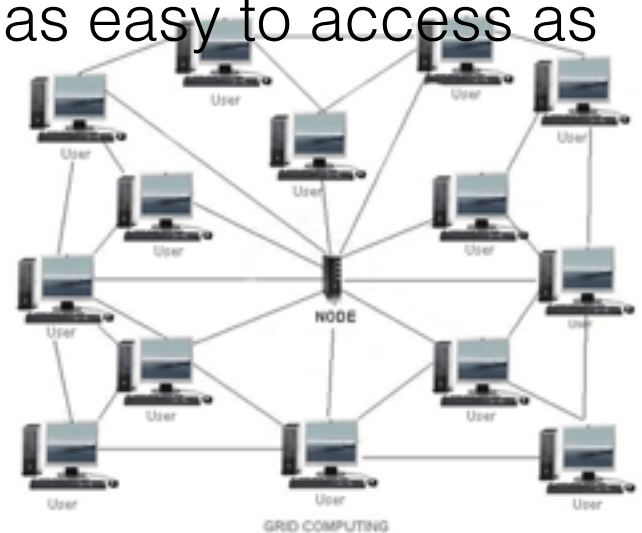
- The term utilities often refer to the set of services consumed by the public: electricity, natural gas, water, sewage, and telephone.
- Computing service through an on-demand, pay-per-use billing method is commonly referred to as a utility

Cluster Computing

- A collection of computers (often COTS hardware) interconnected by a high-speed network
 - Tightly connected (LAN)
 - Running one application
- Uses message passing for communication
- Work as an integrated collection of resources
- Homogeneous nodes, one owner
- Viewed as a single system w/ centralized management
- Supercomputing tasks
 - Scientific calculations
- Example: Nvidia Tesla

Grid Computing

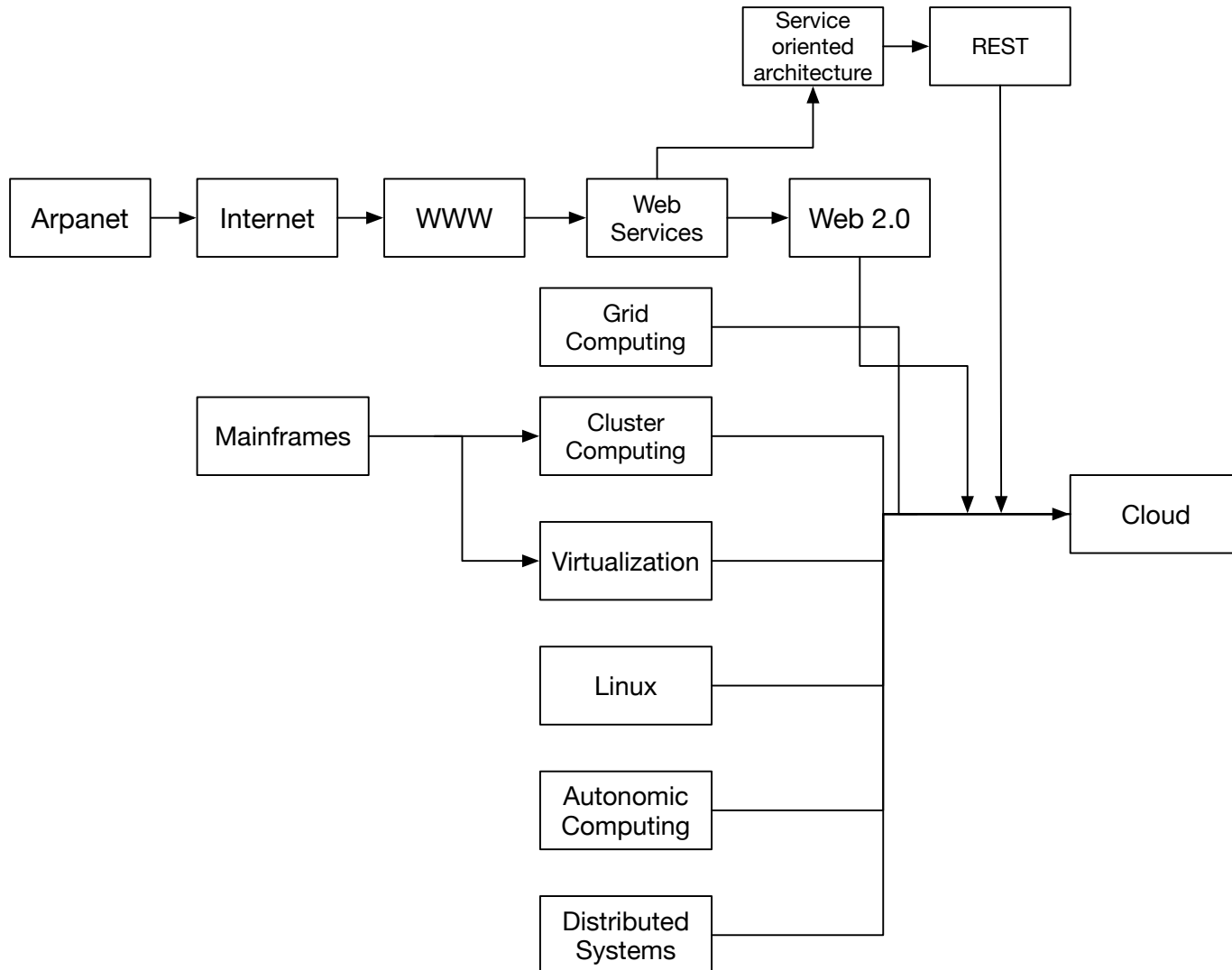
- Grid computing is a form of distributed computing whereby a "super and virtual computer" is composed of networked, loosely coupled computers, acting in concert to perform very large tasks
 - Geographically distributed.
 - Heterogeneous nodes, different owners
- The term grid computing originated in the early 1990s as a metaphor for making computer power as easy to access as an electric power grid
- Moderate number of nodes
 - Started by connecting clusters
- Large scale scientific problems

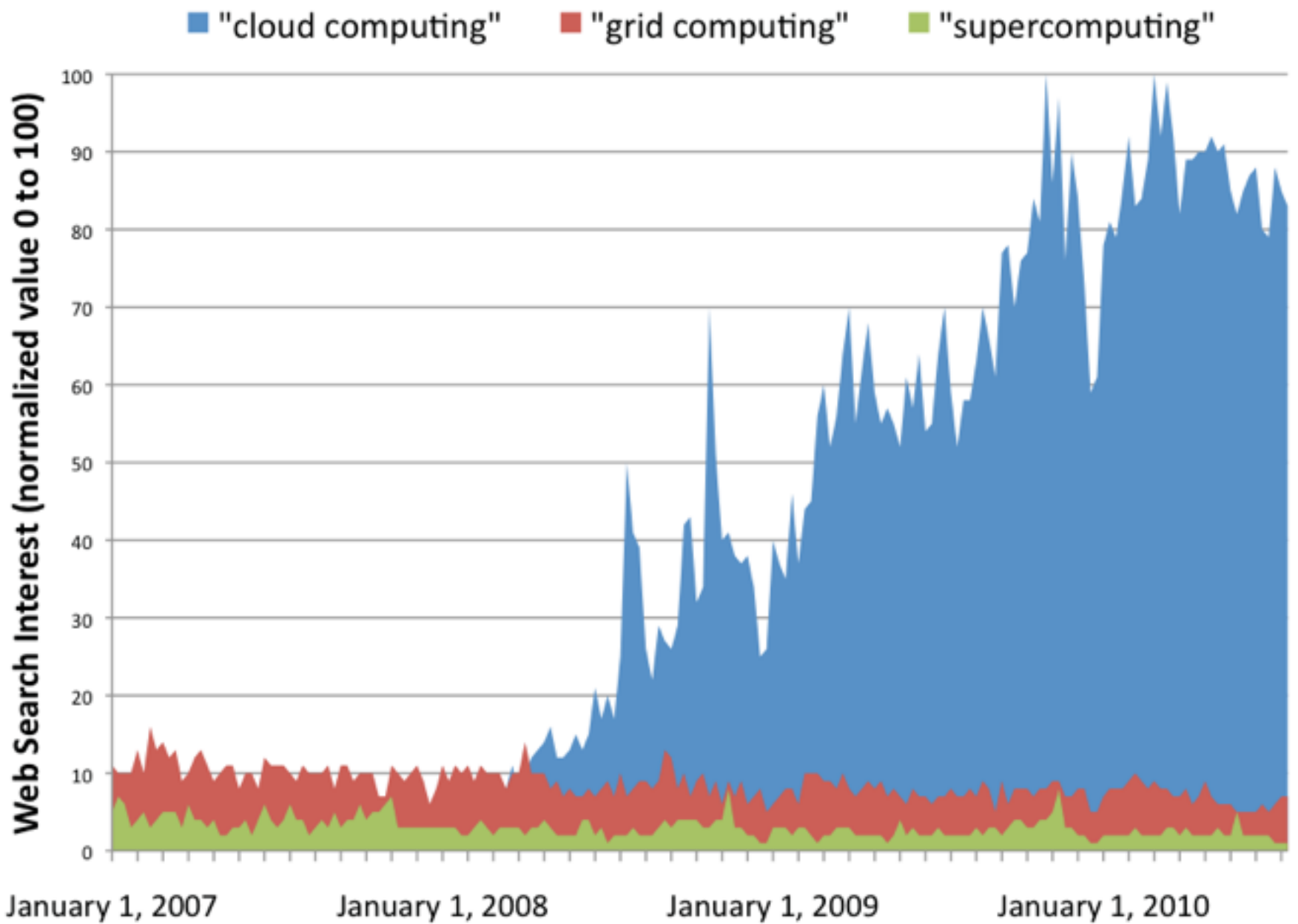


Peer-to-Peer

- A model of communication where every node in the network acts alike, as peers, without centralized management.
 - Geographically distributed.
 - Heterogeneous nodes, different owners
- The participants of such a network are both consumers & producers.
- Ad hoc connectivity, nodes come & go
- Anonymity
- Scalability - Large number of participants
 - DHT - Distributed Hash Tables
- Example: Napster, Gnutella

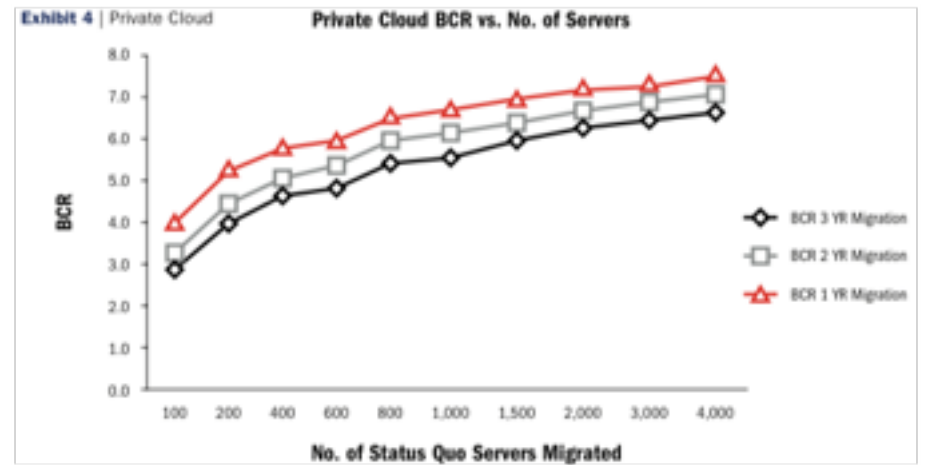
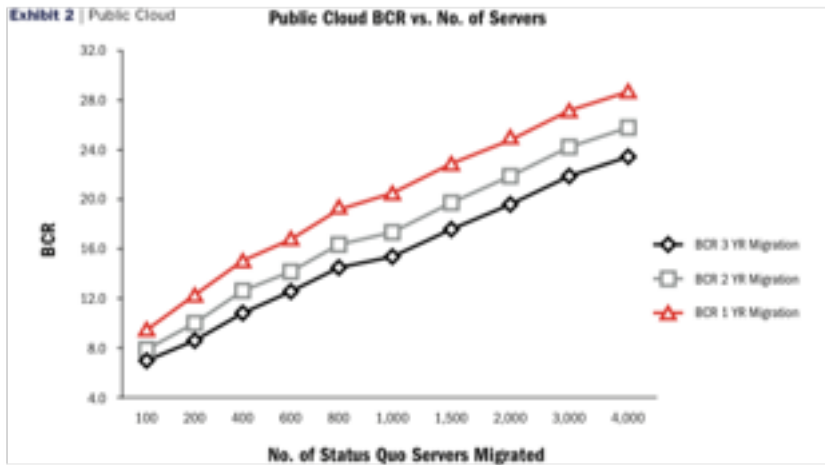
It All Came Together





Cost Reduction

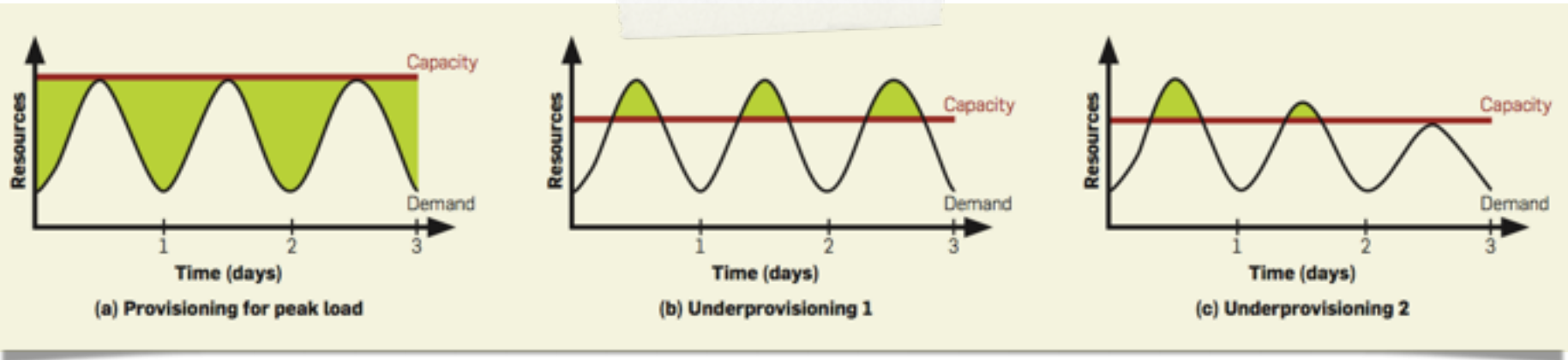
US administration moving to cloud saves 7-28 times



BCR=Benefit-to-cost ratios

Calculated over a 13-year life cycle

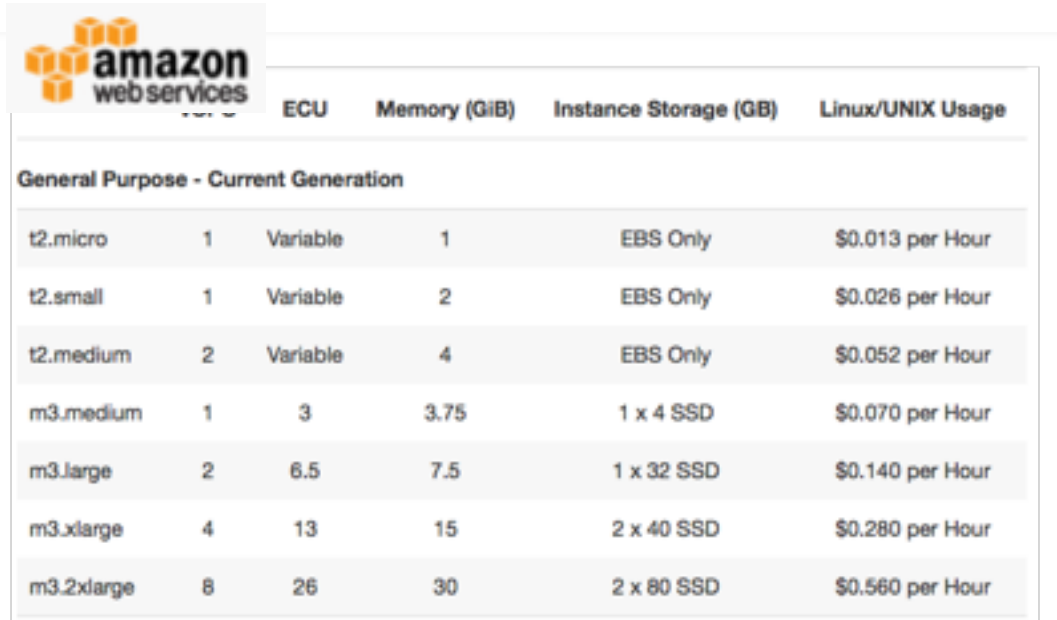
Difficult to dimension



- Workload varies much:
 - Death of Michael Jackson: 22% of tweets, 20% of Wikipedia traffic, Google thought they are under attack
 - Obama inauguration day: 5x increase in tweets
- Over-provisioning is expensive, under-provisioning may be worse

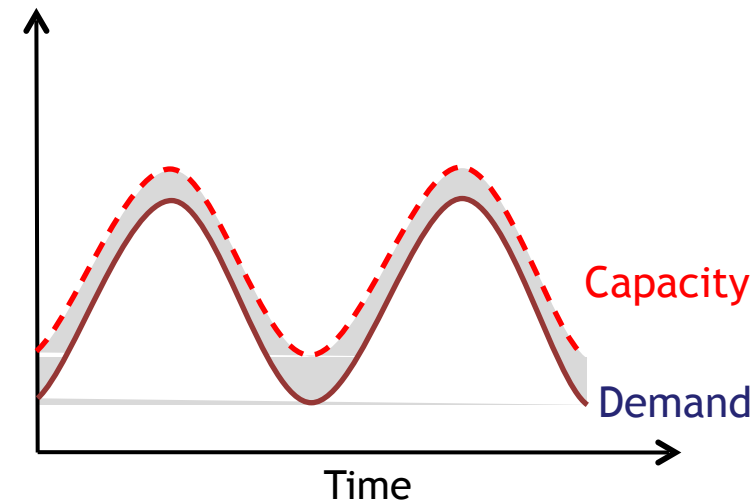
Rent a Datacenter

Pay by use - Rent a VM!



The image shows a screenshot of the Amazon Web Services (AWS) console, specifically the 'General Purpose - Current Generation' instance pricing table. The table lists various instance types with their respective specifications and hourly costs. The columns are: Instance Type, ECU, Memory (GiB), Instance Storage (GB), and Linux/UNIX Usage. The instance types listed are t2.micro, t2.small, t2.medium, m3.medium, m3.large, m3.xlarge, and m3.2xlarge.

	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage	
General Purpose - Current Generation					
t2.micro	1	Variable	1	EBS Only	\$0.013 per Hour
t2.small	1	Variable	2	EBS Only	\$0.026 per Hour
t2.medium	2	Variable	4	EBS Only	\$0.052 per Hour
m3.medium	1	3	3.75	1 x 4 SSD	\$0.070 per Hour
m3.large	2	6.5	7.5	1 x 32 SSD	\$0.140 per Hour
m3.xlarge	4	13	15	2 x 40 SSD	\$0.280 per Hour
m3.2xlarge	8	26	30	2 x 80 SSD	\$0.560 per Hour

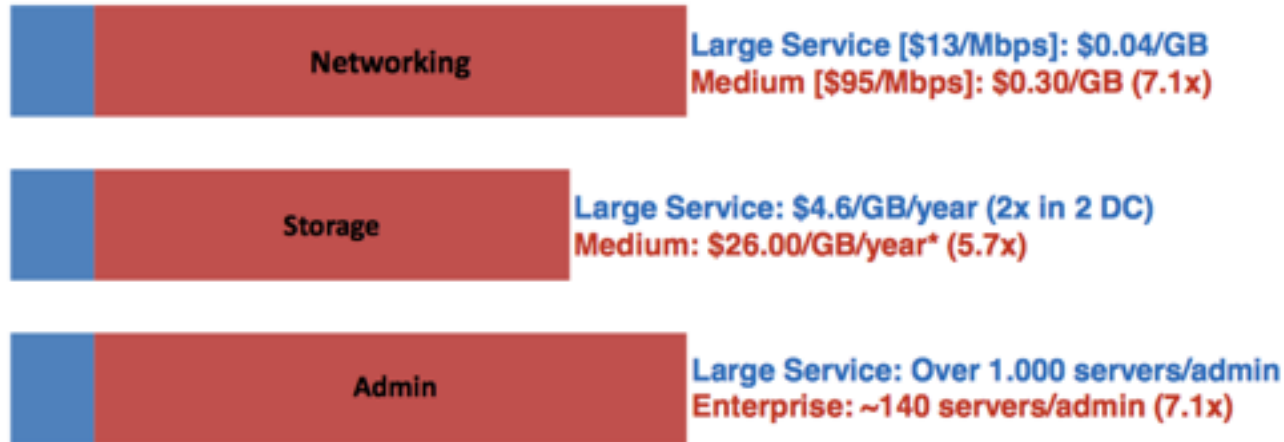


Computing resources in the cloud

1000 machines for 1 hour \Leftrightarrow 1 machine for 1000 hours

Bigger is Better

- Substantial economies of scale possible
- Compare a very large service with a small/mid-sized: (~1000 servers):



- High cost of entry
 - Physical plant expensive: 15MW roughly \$200M
- Summary: significant economies of scale but at very high cost of entry
 - Small number of large players likely outcome



Source: Gartner (May 2014)

AWS Growth Accelerates

Amazon S3 Usage

132% external usage growth YoY in data transfer



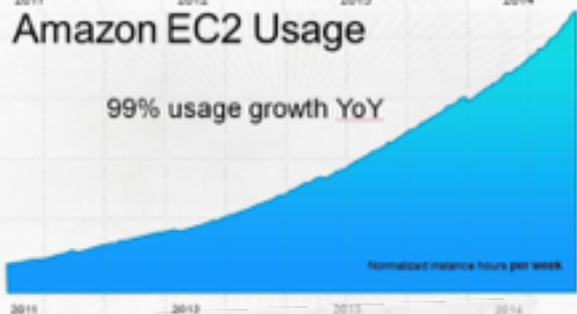
AWS Overall Business

Over 1,000,000 active customers



Amazon EC2 Usage

99% usage growth YoY

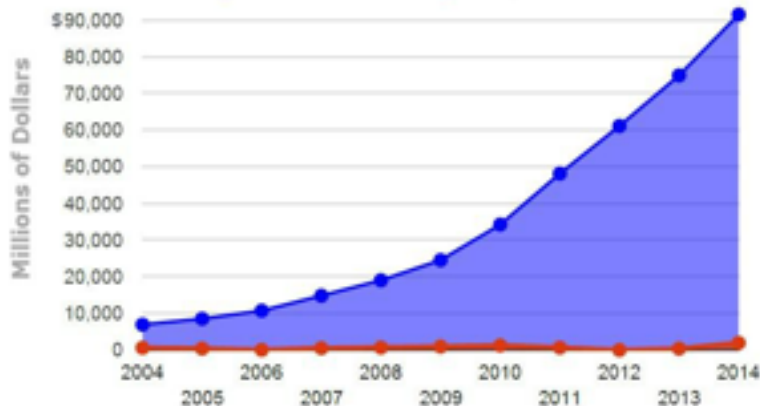


"5X the cloud capacity in use than the aggregate total of the other 14 providers"

Gartner.

Amazon.com's revenue & profit 2004-2014

● Revenue ● Net profit



It Has Only Just Begun

Although the use of cloud services is growing faster than the overall enterprise IT market, it is still a small part of overall IT spending, according Gartner, Inc. A recent Gartner survey on the future of IT services found that only 38 percent of all organizations surveyed indicate cloud services use today. However, 80 percent of organizations said that they intend to use cloud services in some form within 12 months, including 55 percent of the organizations not doing so today.

Source: Gartner <http://www.gartner.com/newsroom/id/2581315>

How much and when will cloud transform these markets? In this research, we project that the public cloud market will reach \$191 billion by 2020, from 2013's total of \$58 billion.



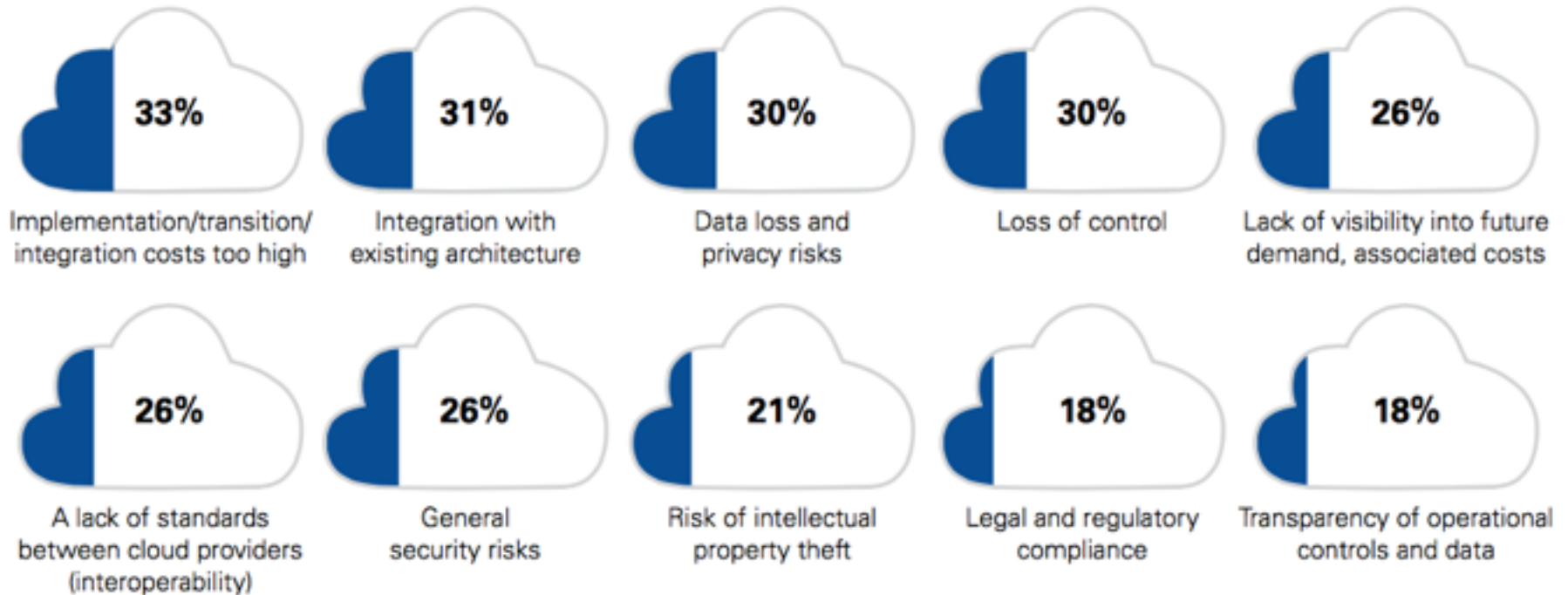
Source: "The Public Cloud Market Is Now In Hypergrowth" <https://www.forrester.com>

Obstacles for Transition

Business Perspective

Which of the following are the key challenges of your approach to cloud adoption?

Total respondents (n = 674)



Source: KPMG International's Global cloud survey: the implementation challenge

Source: KPMG - "The cloud takes shape"

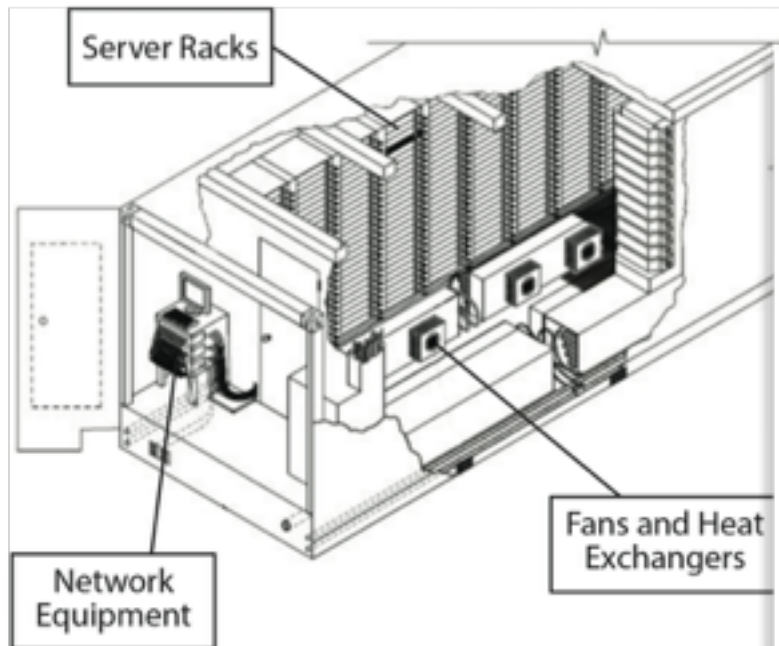
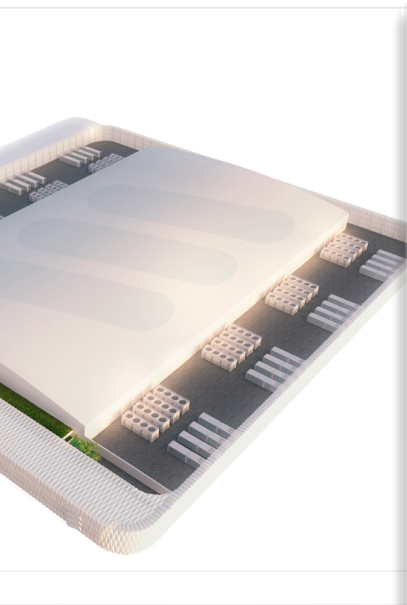
Obstacles for Transition

Technical Perspective

1. Availability
2. Data lock-in
3. Data confidentiality/auditability
4. Data transfer bottlenecks
5. Performance unpredictability
6. Scalable storage
7. Bugs in large-scale distributed systems
8. Scaling quickly
9. Reputation fate sharing
10. Software licensing

Source: "A View of Cloud Computing", Armbrust et al

The Datacenter

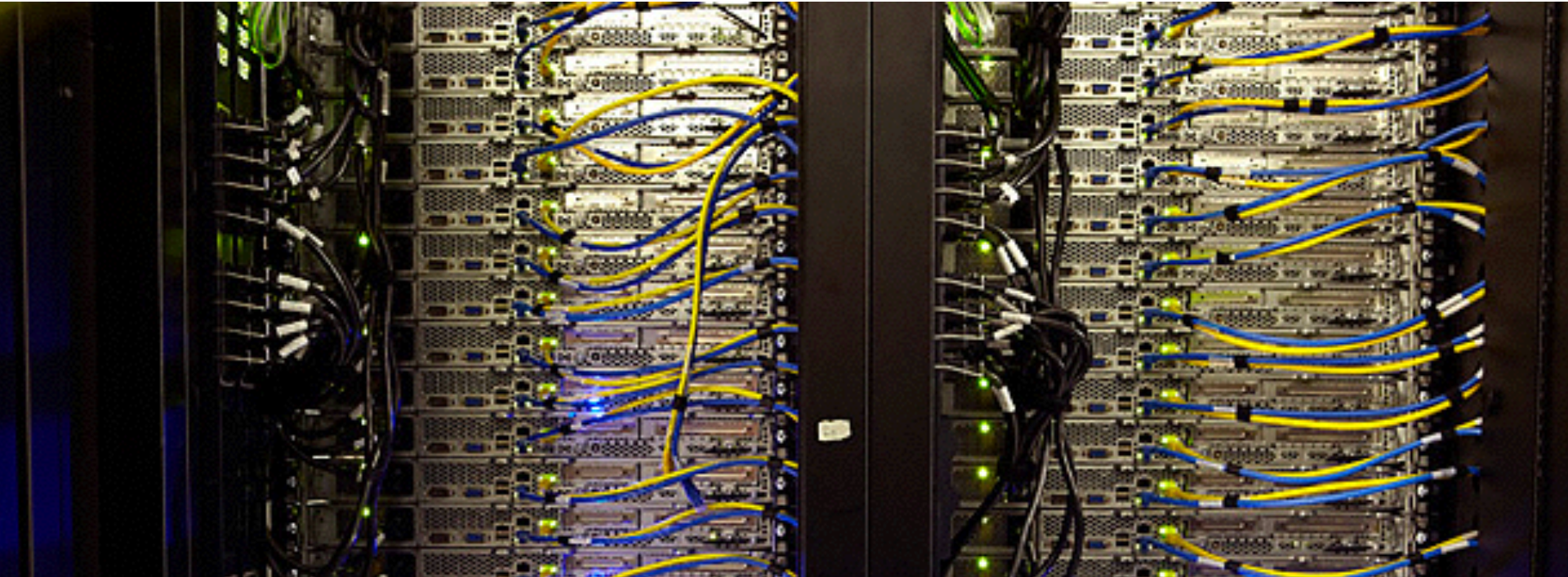


What's inside?



Racks

What's inside?



Networking

What's inside?



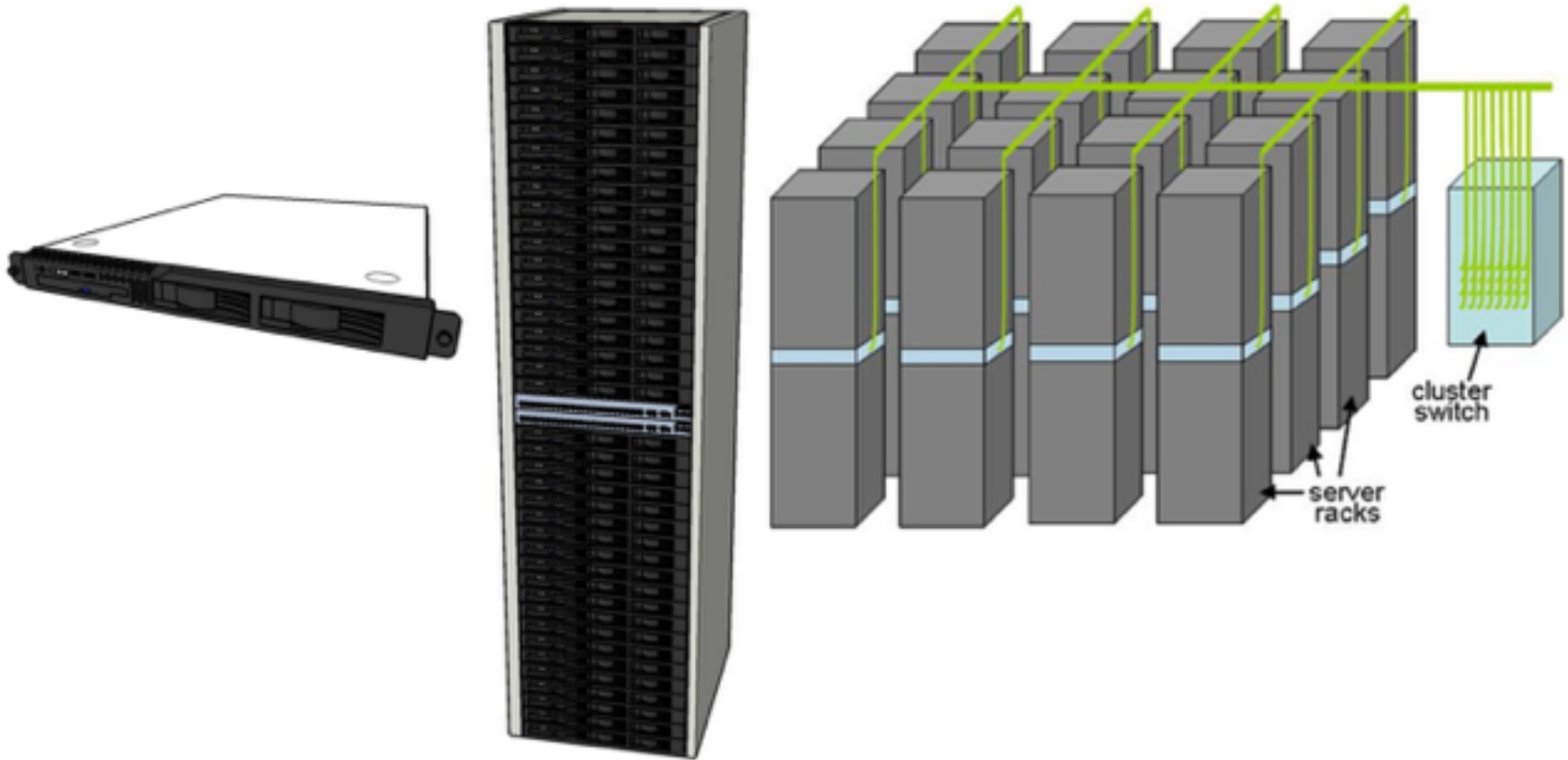
Power supplies

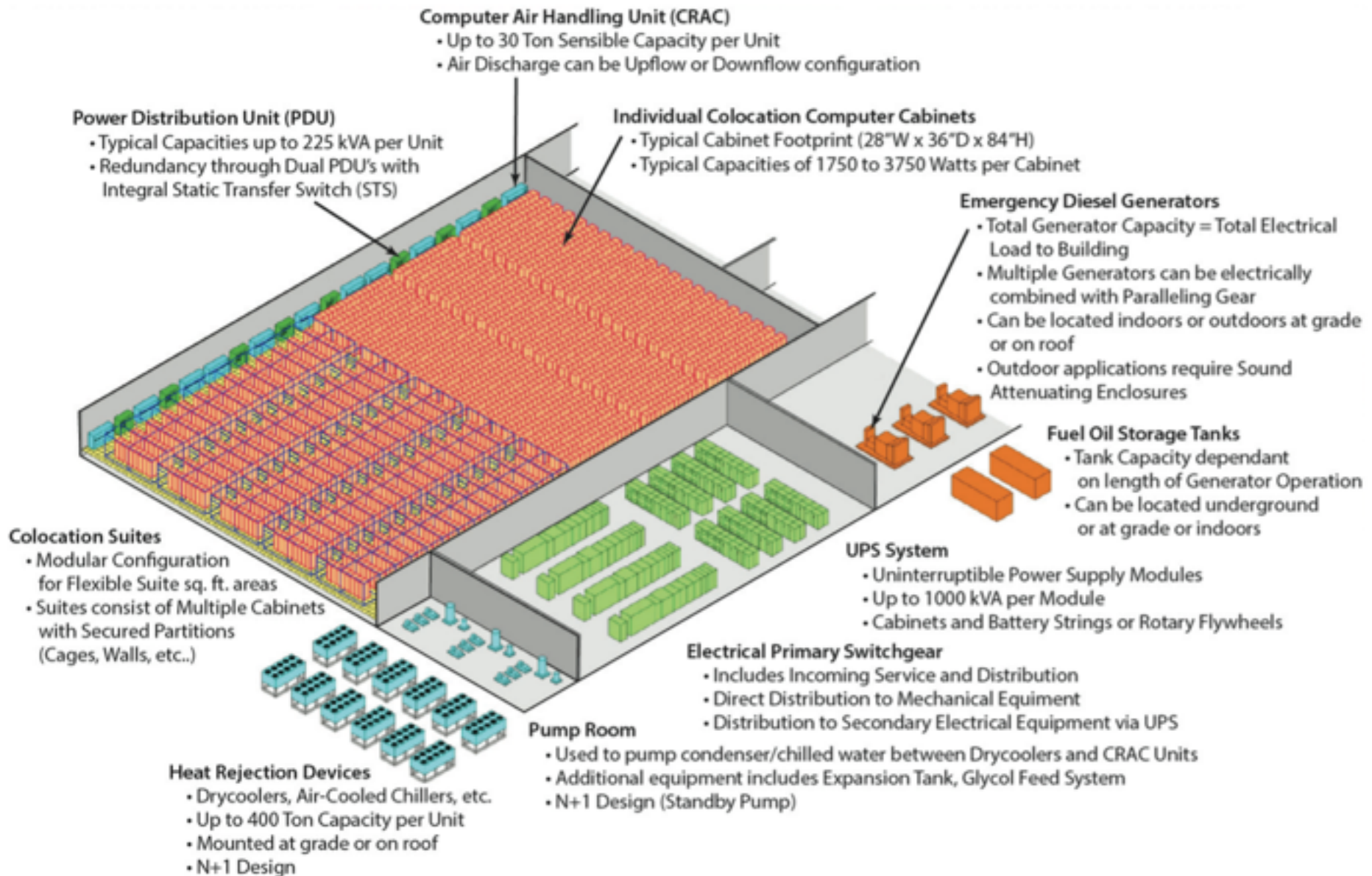
What's inside?



Cooling

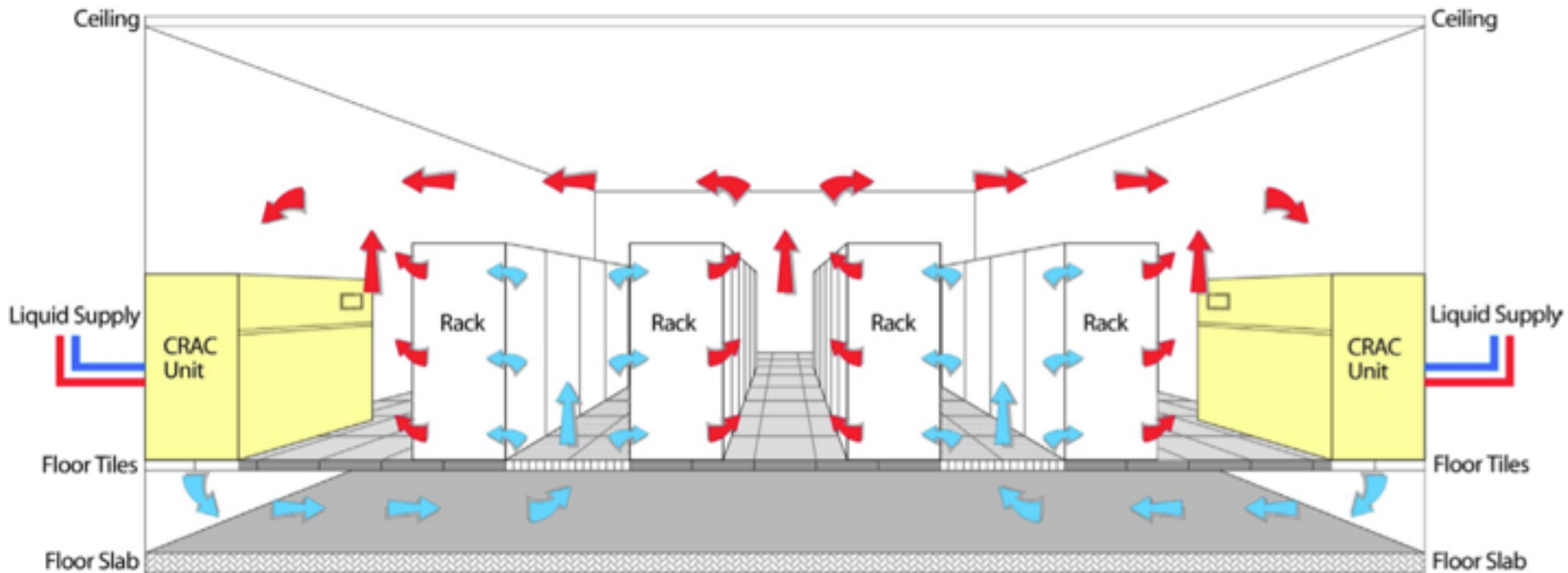
Datacenter Elements





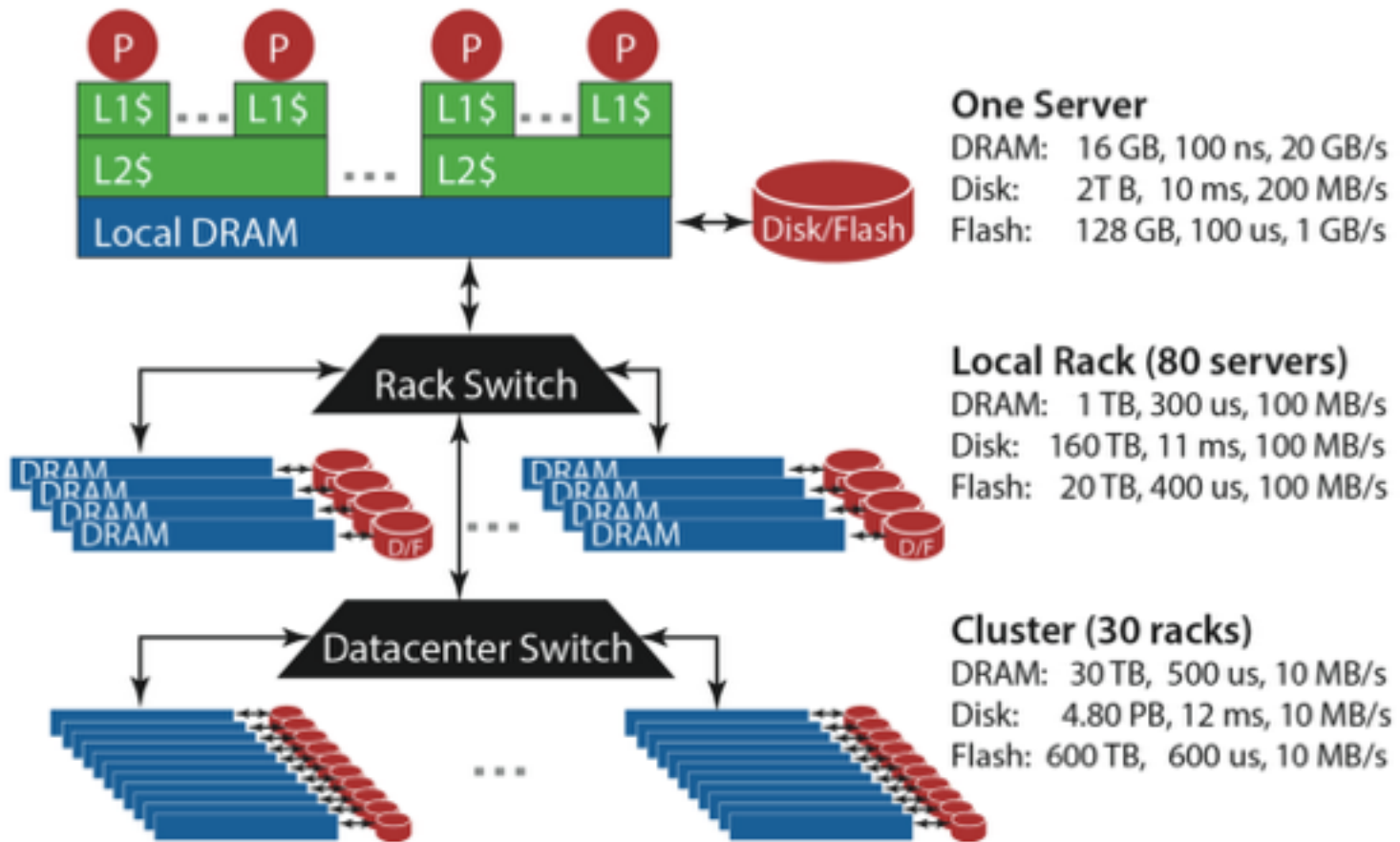
Source: "The Datacenter as a computer", Barroso et al

Computer Architecture



Source: "The Datacenter as a computer", Barroso et al

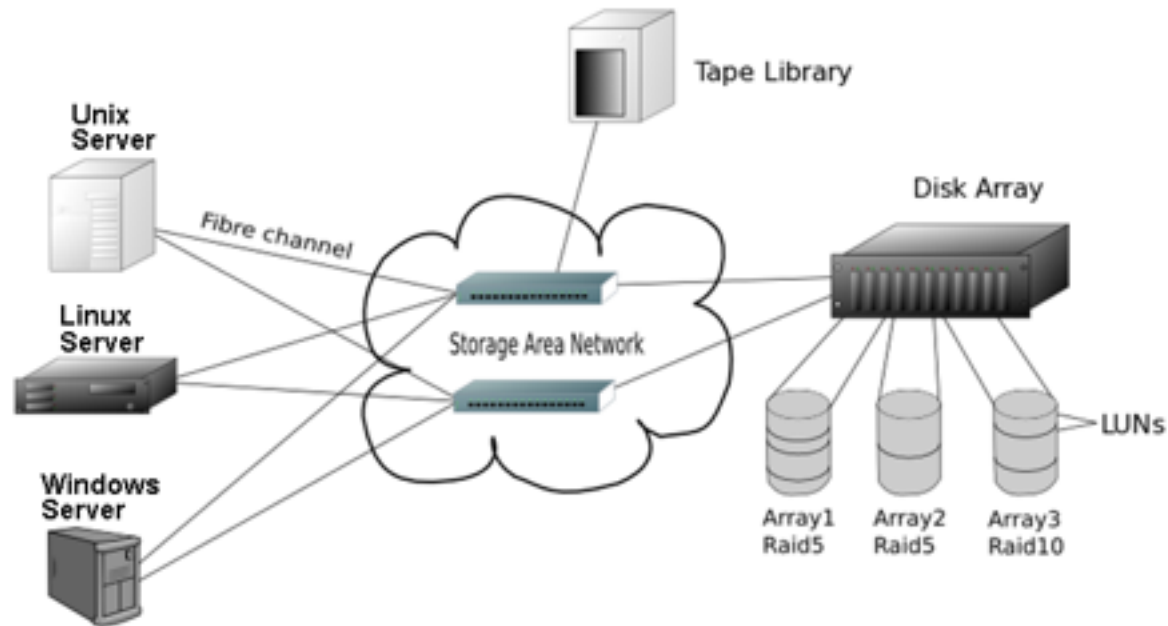
Storage at Google



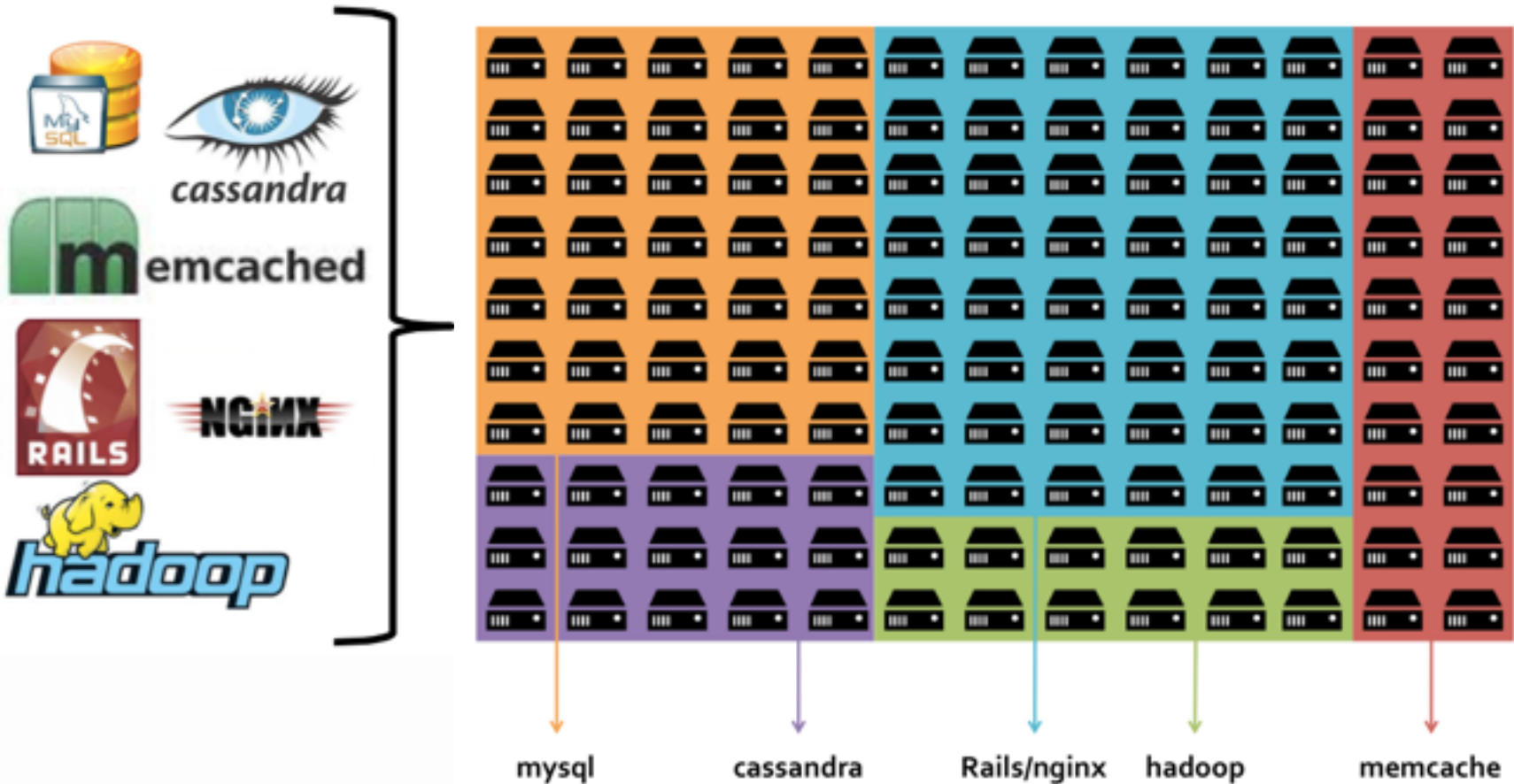
Source: "The Datacenter as a computer", Barroso et al

Storage Area Network (SAN)

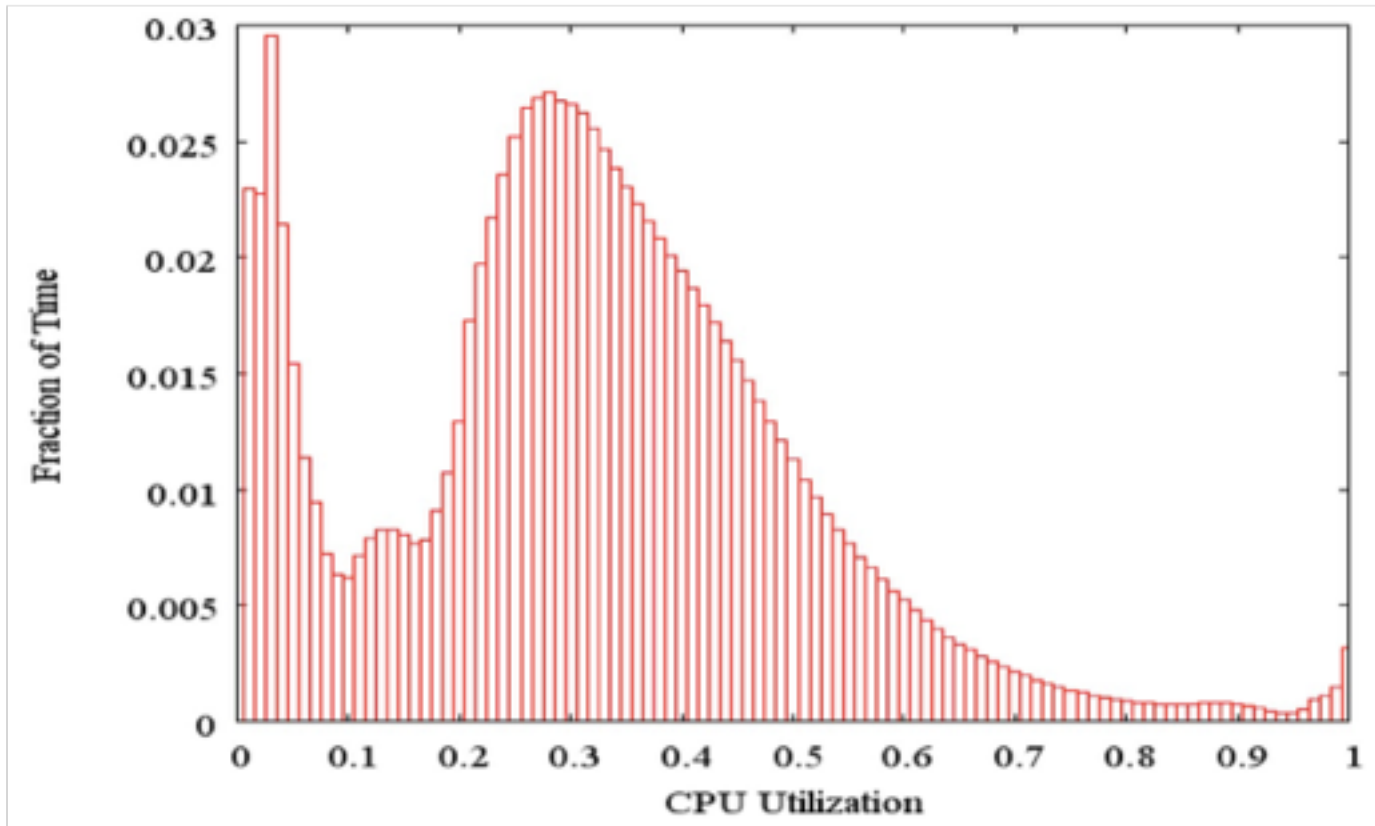
- A dedicated network that provides access to consolidated, block level data storage.
- Separation between compute and storage
- A NAS is a single storage device that operate on data files, while a SAN is a local network of multiple devices that operate on disk



Resource Management

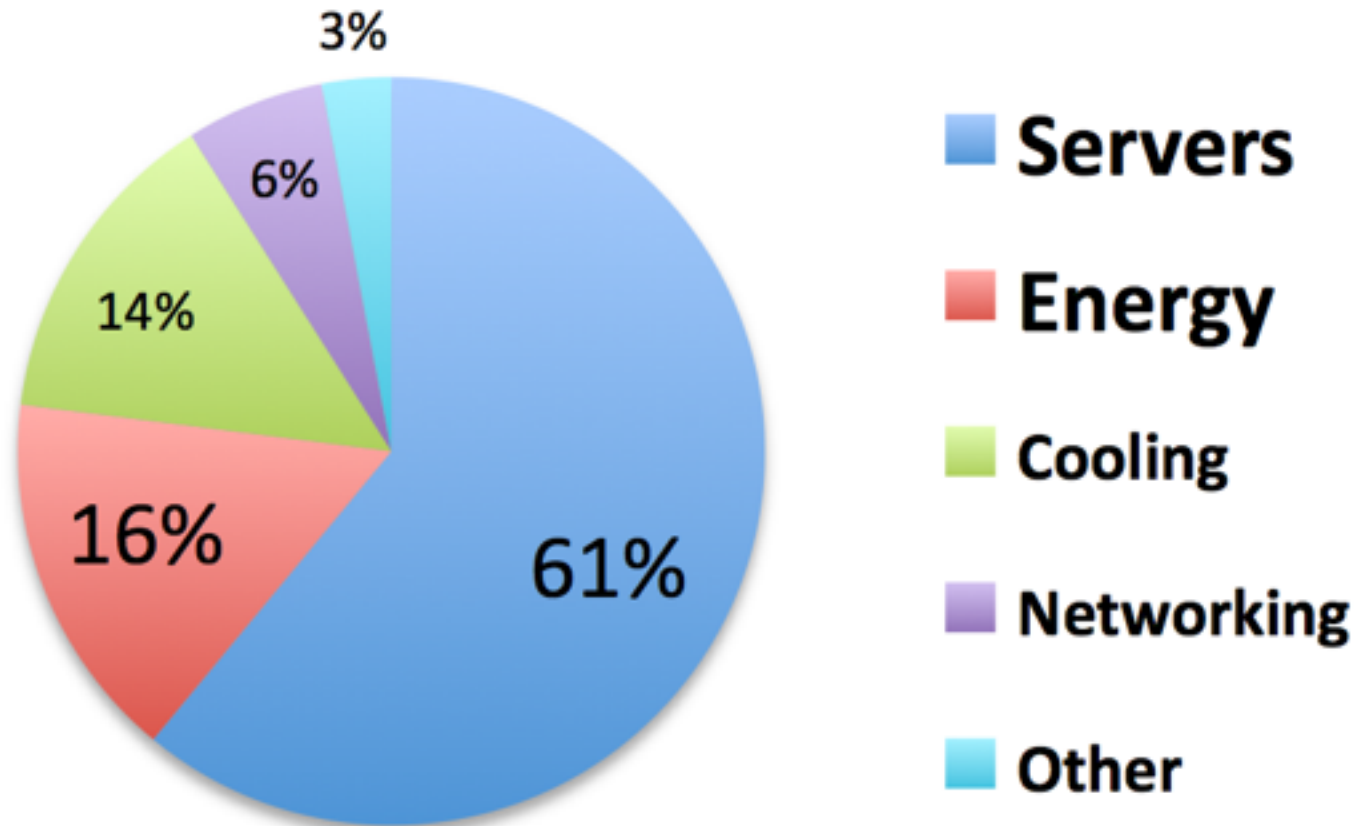


Datacenter Utilization



5000 google servers over 6 months
Source: *"The Datacenter as a computer"*, Barroso et al

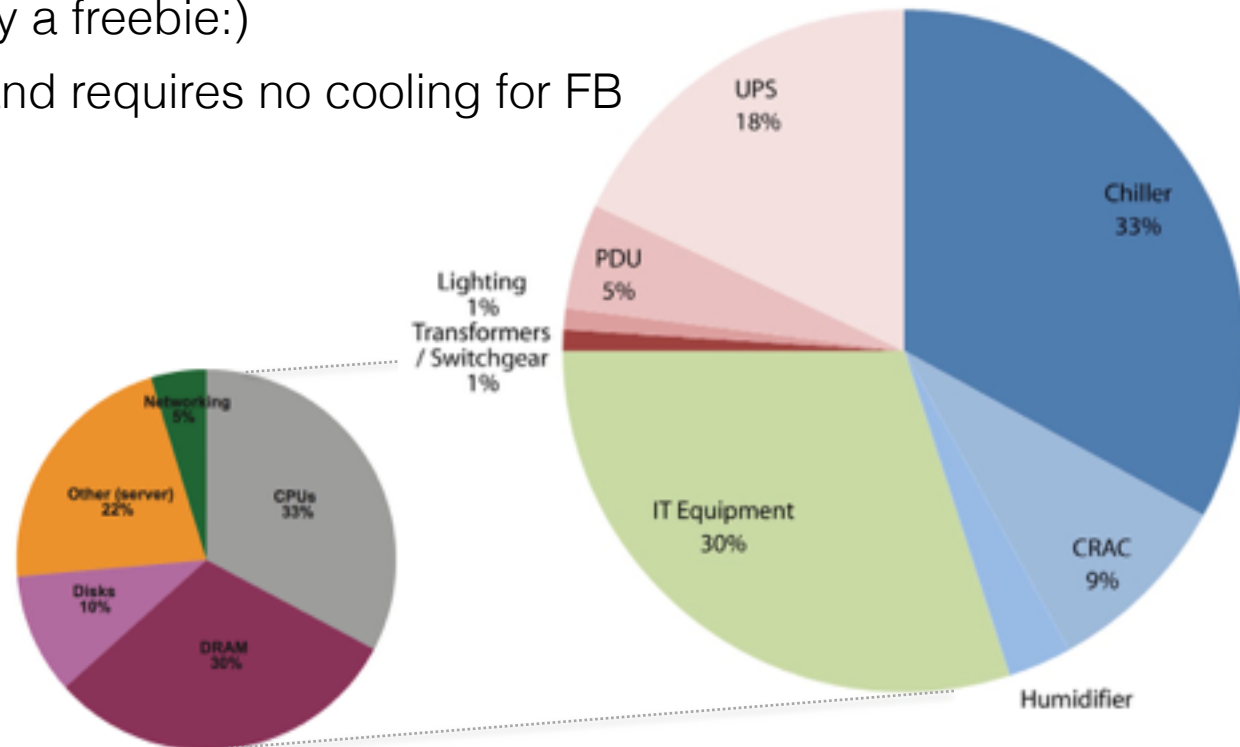
Total Cost of Ownership



[J. Hamilton, <http://mvdirona.com>]

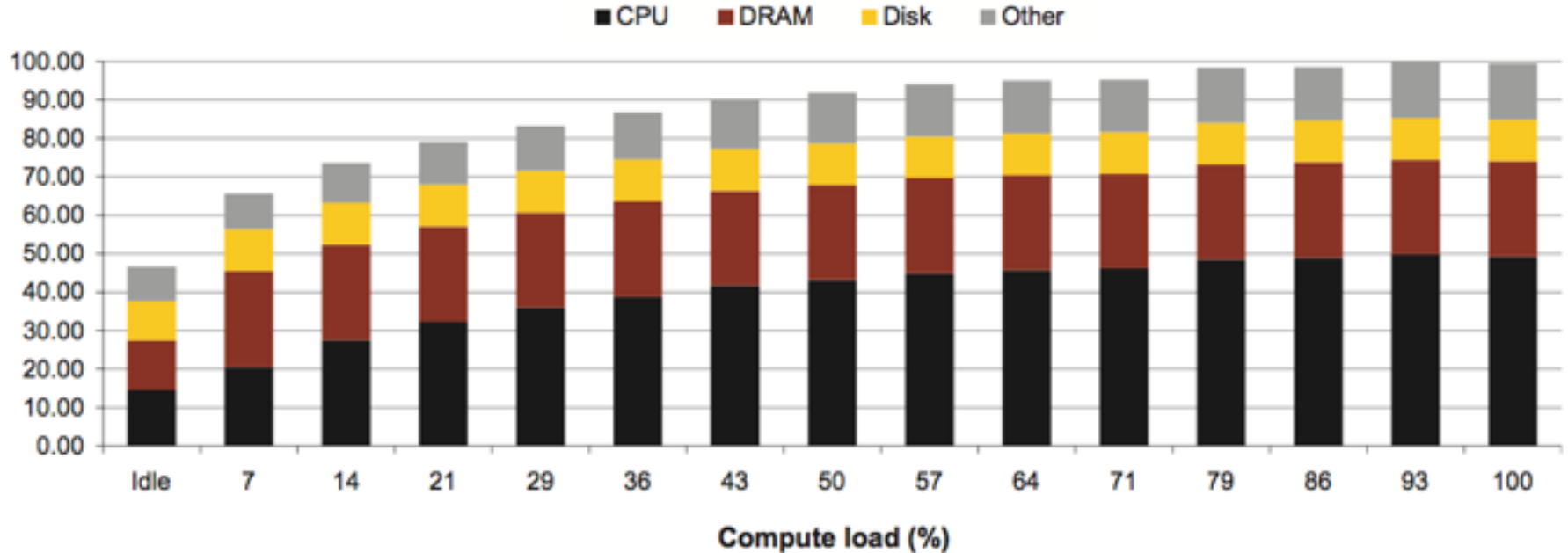
Power Usage Classic DC

- Power availability drives datacenter deployment decisions
 - Facebook locates DC in Luleå
 - Google builds one in Finland
- The bad weather is merely a freebie:)
- Below 27° celsius is OK and requires no cooling for FB

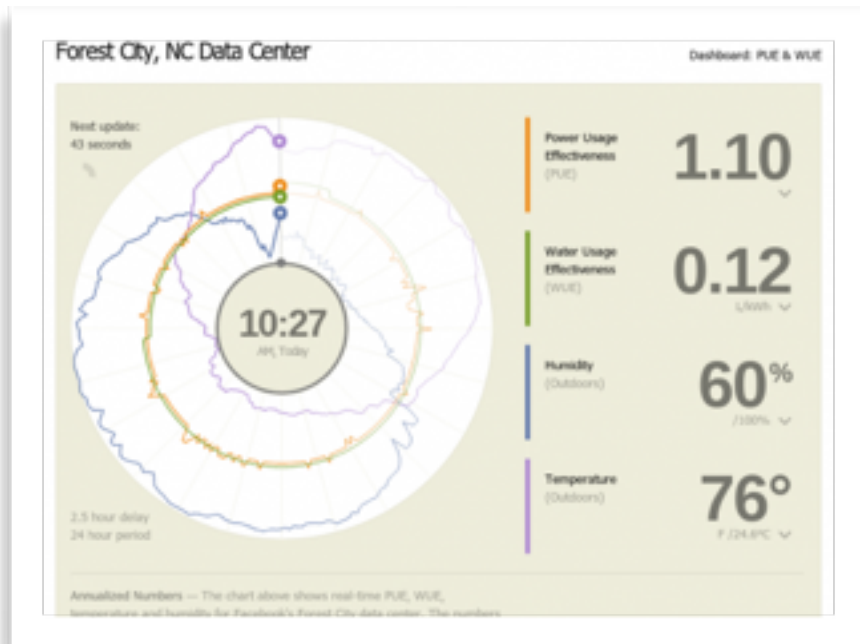


Source: "The Datacenter as a computer", Barroso et al

The Hardware is Not Energy Proportional



Facebook DC Efficiency



$$\text{WUE} = \frac{\text{Annual Water Usage}}{\text{IT Equipment Energy}}$$

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

Infrastructure-as-a-Service

- Provide a virtual datacenter
 - Compute/storage/network
 - Often some basic services also
 - The user is responsible for making the application run correctly, i.e. fault tolerance, timing, handling crashes, scaling, authentication, redundancy, etc.
- Pay per usage
- Amazon Web Services, Google Compute, Rackspace

Application

Runtime

Databases

Security

OS

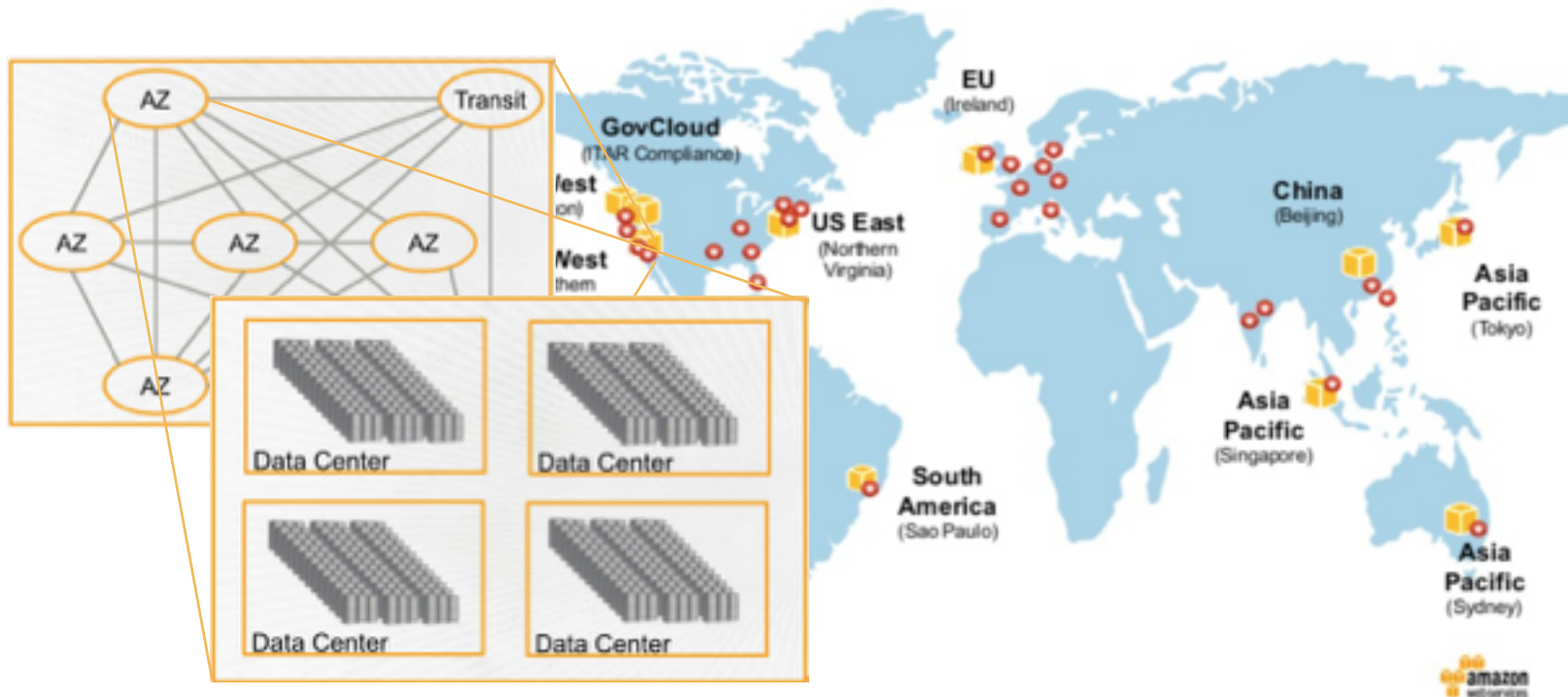
Virtualization

Servers

Storage

Network

Amazon Infrastructure



- 11 Regions, connected by private fiber
- Regions consists of 2 or more AZs
- 28 Az
- AZ < 2 ms apart and usually <1ms
- AZ one or more DCs
- DC consists of 50-80.000 machines
- Inter AZ DCs < 1/4 ms apart

source: Jame Hamilton, ReInvent 2014

AWS

- EC2 (Elastic Compute Cloud)
 - Linux or Windows VM
 - On Demand instances
 - Reserved instances - long term, low price
 - Spot instances” - you bid on a VM
 - Dedicates instances - single tenant HW
- AMI (Amazon Machine Image)
 - A large # of prefabs + make-your-own



	Linux/UNIX Usage
Standard On-Demand Instances	
m1.small	\$0.048 per Hour
m1.medium	\$0.096 per Hour
m1.large	\$0.193 per Hour
m1.xlarge	\$0.385 per Hour
Second Generation Standard On-Demand Instances	
m3.medium	\$0.077 per Hour
m3.large	\$0.154 per Hour
m3.xlarge	\$0.308 per Hour
m3.2xlarge	\$0.616 per Hour

Service Commitment

AWS will use commercially reasonable efforts to make Amazon EC2 and Amazon EBS each available with a Monthly Uptime Percentage (defined below) of at least 99.95%, in each case during any monthly billing cycle (the "Service Commitment"). In the event Amazon EC2 or Amazon EBS does not meet the Service Commitment, you will be eligible to receive a Service Credit as described below.

4.3 hours downtime/year

More AWS Services

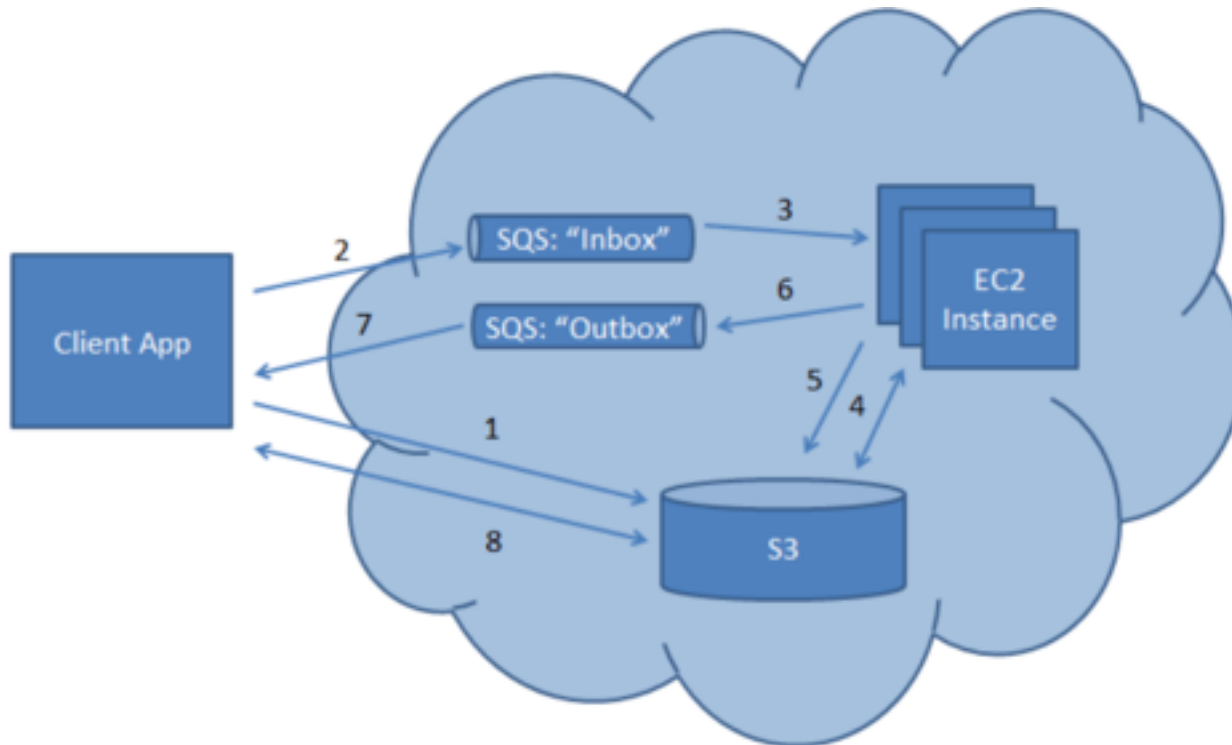
- EBS (Elastic Block Storage)
 - Raw, unformatted device, can create file systems, etc. on top
 - Can be mounted like device/file system by EC2 instances
 - Persistent storage for EC2 images
- S3 (Simple Storage Service)
 - Bucket – object container with unique name
 - Key – unique name in bucket (like file name)
 - Objects are indexed by bucket+key, <http://s3.amazonaws.com/bucket/key>
 - Operations (RESTful)
 - Create/Delete bucket
 - Put/Get/Delete data from bucket
- SQS (Simple Queuing System)

Enabling Technologies

- Distributed computing
- Networking
- Virtualization
- Storage
- Datacenter OS & Application architecture
- Programming models

Home Assignment #1

Build a small application on AWS



Course Project

- Build your own search engine
- Build your own Facebook
- Make a PaaS for *your* research area
 - Eg. control system simulator in the cloud
- More to come...