Imperial College London

LSDS
Large-Scale Data & Systems Group

Cloud Computing and Data-centres

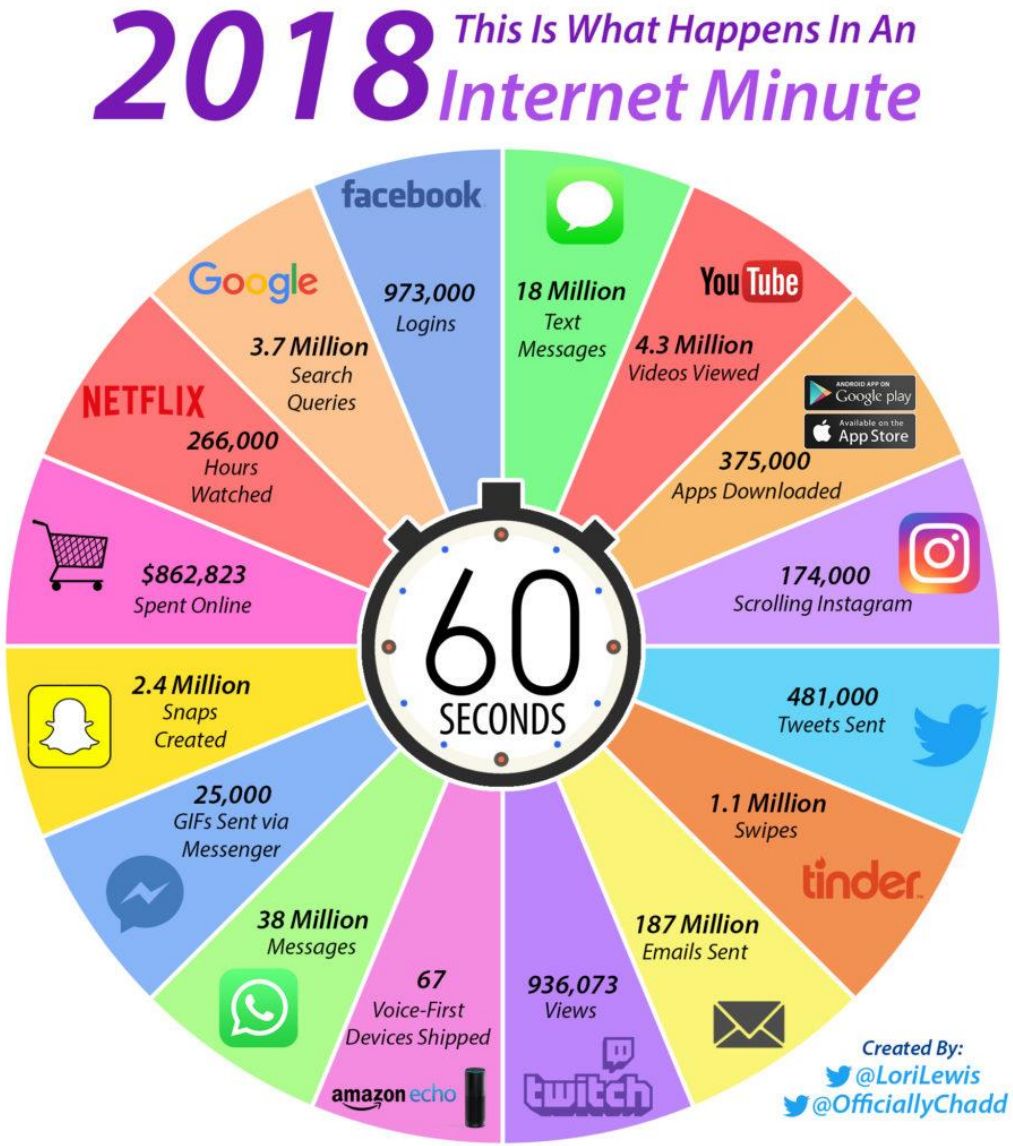
Jana Giceva
jgiceva@doc.ic.ac.uk

Department of Computing
Imperial College London
http://lsds.doc.ic.ac.uk

Fall 2018

Cloud Computing

Big Data and the need for Cloud



This is Big Data!

- Large ***Volume*** of data
- Coming at a high ***Velocity***
- With a large ***Variety***
- What about ***Veracity***

Challenge for traditional IT Systems, so now …

They are supported by moving computing to the Cloud!

What is a Cloud?

Datacentre hardware and software that the vendors use to offer the computing resources and services.

It is a large pool of easily usable ***virtualized*** computing ***resources***, development ***platforms*** and various services and ***applications***.

src image from Google datacenter

What is Cloud Computing?

Cloud computing is the delivery of ***computing as a service***.

Where the ***shared resources, software, and data*** are provided to users by a provider

As a ***metered service over a network.***

Cloud Computing Pros and Cons

User's benefits: ?

- Speed – services provided on demand
- Global scale and elasticity
- Productivity
- Performance and Security
- Customizability

User's concerns: ?

- Dependency on network and internet connectivity
- Security and Privacy
- Cost of migration
- Cost and risk of vendor lock-in

Types of Cloud Computing

Public cloud

- All hardware, software and other supporting infrastructure is owned and operated by cloud vendors (service providers).
- Cloud vendors offer their computing resources over the Internet.
- Example: Amazon AWS, Microsoft Azure, Google Cloud Services

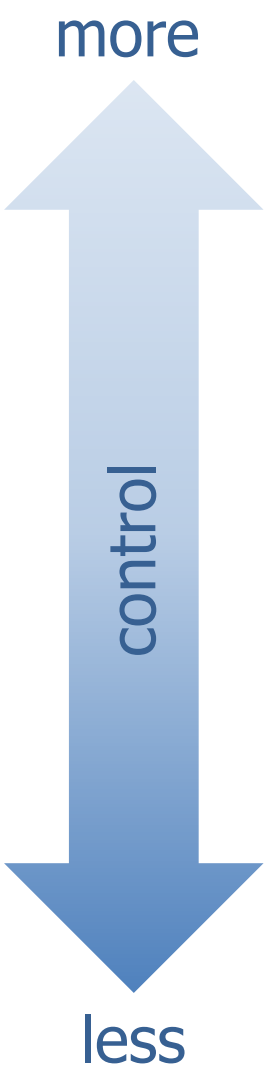
Private cloud

- Cloud computing infrastructure used exclusively by a single business or organization (e.g., physically hosted on the company's on-site data centre).
- Services are infrastructure are maintained on a private network.

Hybrid cloud

- Combines public and private clouds: allows data and applications to be shared between them.
- Gives a business greater flexibility to optimize existing infrastructure, security and compliance.

Cloud Service Models



Infrastructure as a Service (IaaS)

- Rent IT infrastructure – servers and virtual machines (VMs), storage, networks, firewall and security.

Platform as a Service (PaaS)

- Get on-demand environment for development, testing and management of software applications – servers, storage, network, OS, databases, etc.

Serverless (FaaS)

- Overlapping with PaaS, serverless focuses on building app functionality without managing the servers and infrastructure required to do so.
- Cloud vendor provides set-up, capacity planning, and server mgmt.

Software as a Service (SaaS)

- Deliver software applications over the Internet, on demand.
- Cloud vendor handles software application and underlying infrastructure, and handles any maintenance (upgrades, patches, etc.).

Infrastructure as a Service

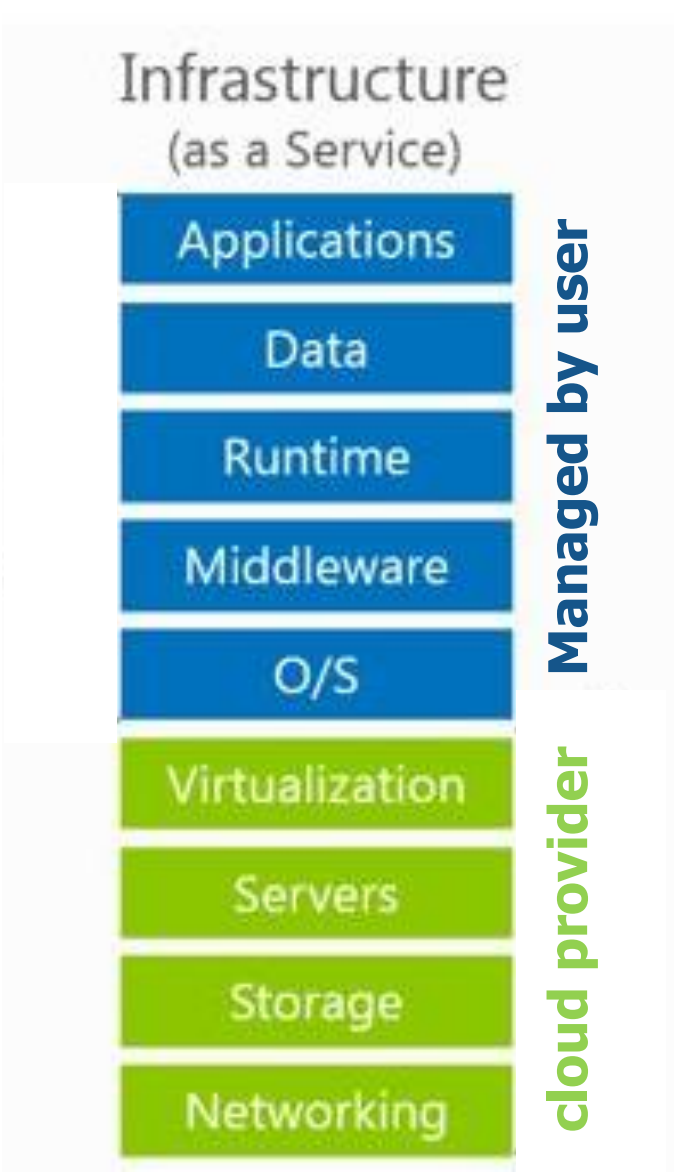
Immediately ***available computing infrastructure***, provisioned and managed by a cloud provider.

Computing ***resources pooled*** together to serve multiple users/tenants.

Computing resources include: storage, processing, memory, network bandwidth, etc.

What can we use it for?

What are the advantages?



src image from Microsoft Azure

Platform as a Service

Complete development and deployment environment.

Includes system's software (OS, middleware), platforms, DBMSs, BI services, and libraries to assist in development and deployment of cloud-based applications.

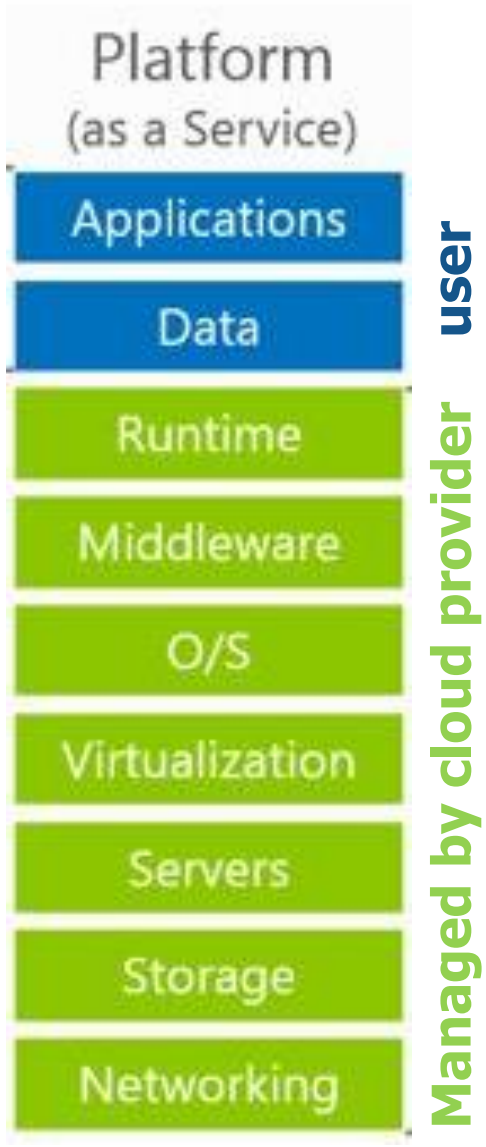
Examples

ORACLE
CLOUD PLATFORM

amazon
web services

What are the advantages?

What is serverless computing then?



src image from Microsoft Azure

Software as a Service

Data Centres

What is a datacenter?

A datacenter (DC) is a physical facility that enterprises use to house computing and storage infrastructure in a variety of networked formats.

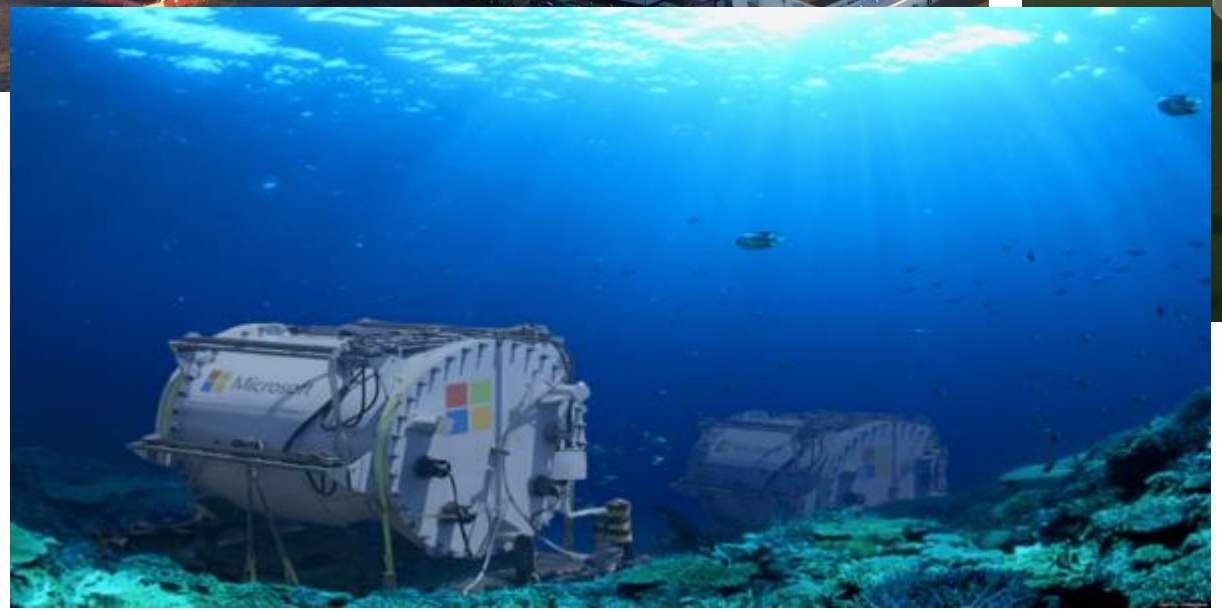
Main function is to deliver utilities needed by the equipment and personnel:

- Power
- Cooling
- Shelter
- Security

Size of datacenters:

- 500-5000 sqm buildings
- 1 MW to 10-20 MW power (on average around 5 MW)

Example data-centers



What you should optimize for?

Does the business require mirrored data centers?

How much geographic diversity is required?

What is the necessary time to recover in the case of an outage?

How much room is it required for expansion?

Should you lease a private data center or a public service?

What are the bandwidth and power requirements?

Is there a preferred carrier?

What kind of physical security is required?

Datacenter standards and classification (ANSI-TIA-942)

Tier	Generators	UPSs	Power Feeds	HVAC	Availability
1	None	N	Single	N	99.671%
2	N	N+1	Single	N+1	99.741%
3	N+1	N+1	Dual, switchable	N+1	99.982%
4	2N	2N	Dual, simultaneous	2N	99.995%

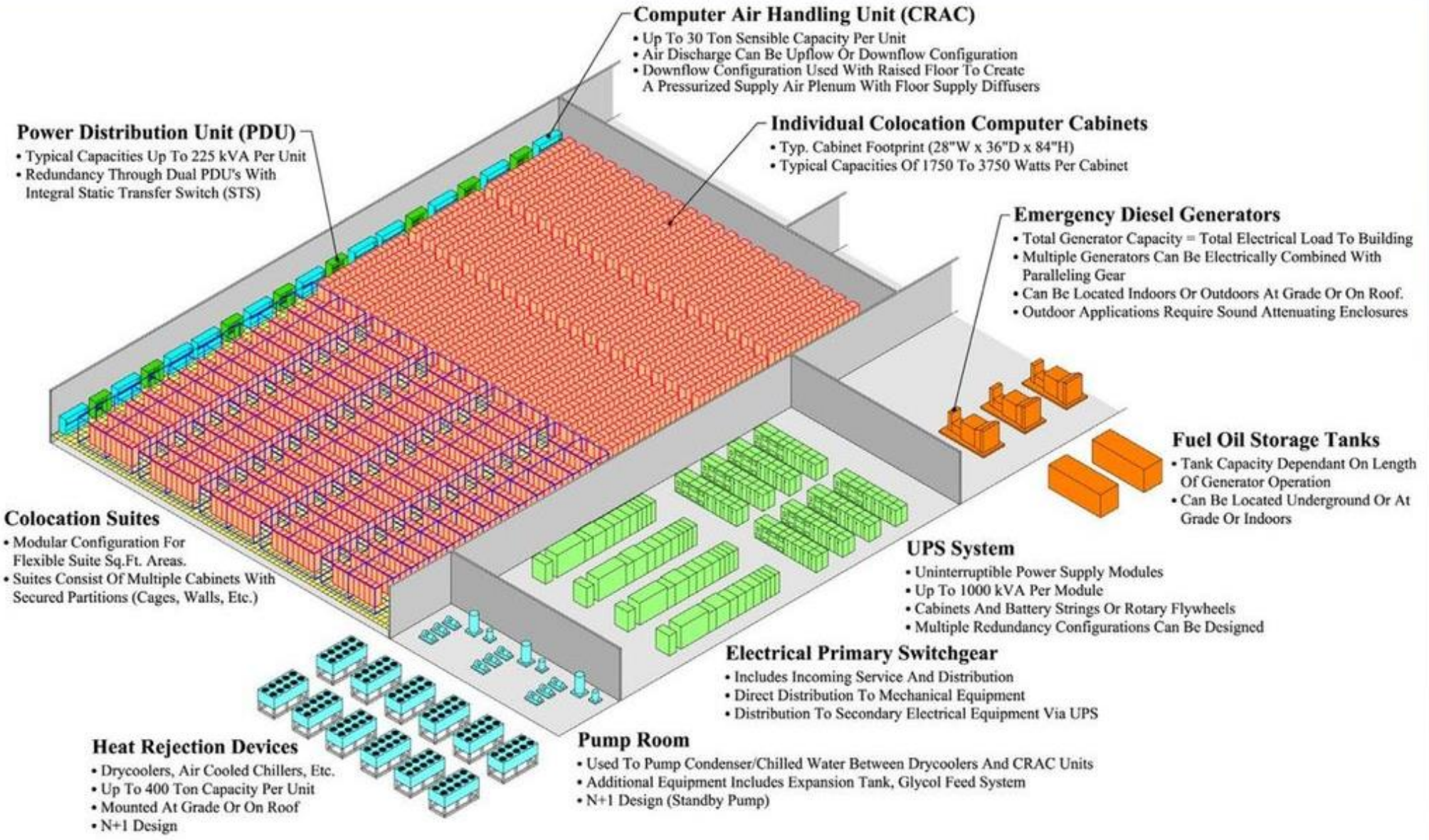
Rate-1: Basic Site Infrastructure

Rate-2: Redundant Capacity Component Site Infrastructure

Rate-3: Concurrently Maintainable Site Infrastructure

Rate-4: Fault-Tolerant Site Infrastructure

What are the main components of a datacenter?



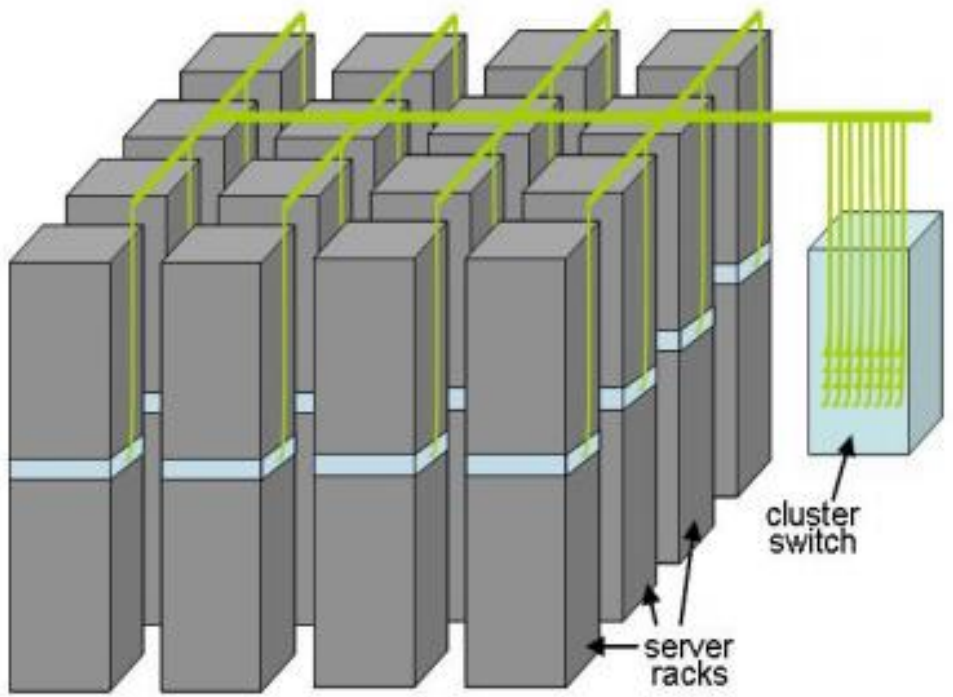
src: The Datacenter as a Computer – Barroso, Clidaras, Holzle

What's inside a data center?

Servers mounted on 19" rack cabinets

Servers
- CPUs
- DRAM
- Disks

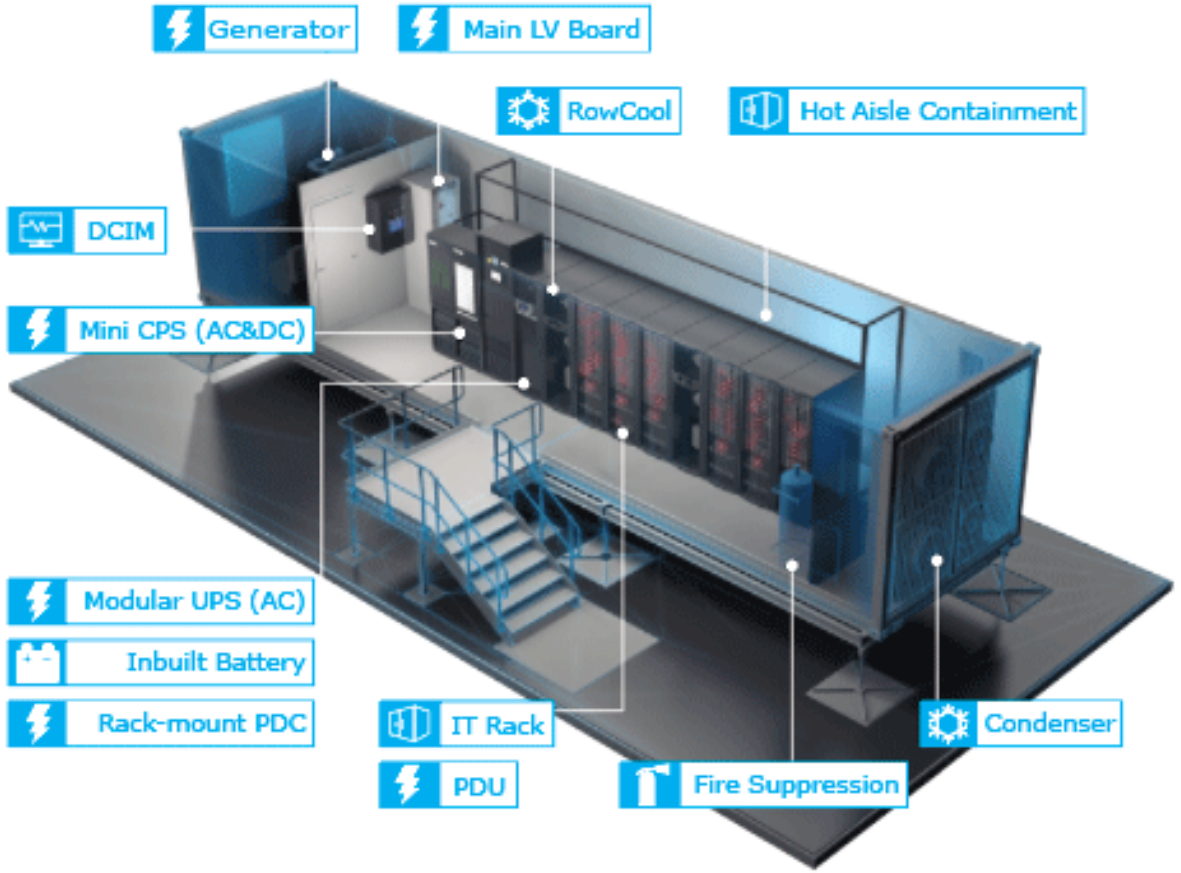
Racks
- 40-80 servers
- Ethernet switch



Clusters

Racks are placed in single rows forming corridors between them.

What's inside a data center?



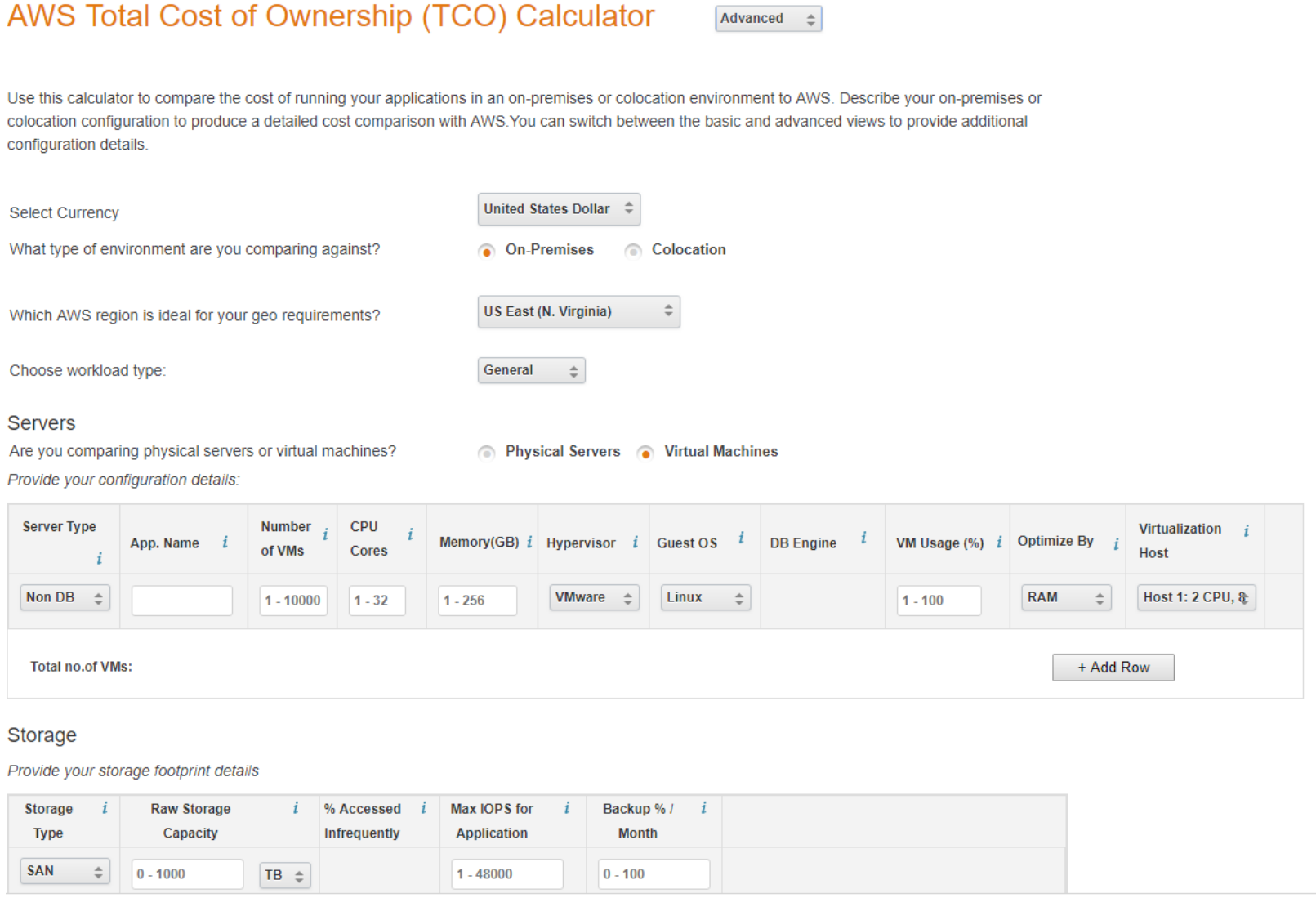
Today's DCs use shipping containers packed with 1000s servers each.

For repairs, whole containers are replaced.

Costs for running a data-center

TCO = CapEx + OpEx

- CapEx – capital expenses, investments that must be made upfront
- OpEx – operational expenses, monthly costs of running the equipment: electricity, maintenance, etc.



AWS Total Cost of Ownership (TCO) Calculator

Advanced

Use this calculator to compare the cost of running your applications in an on-premises or colocation environment to AWS. Describe your on-premises or colocation configuration to produce a detailed cost comparison with AWS.You can switch between the basic and advanced views to provide additional configuration details.

Select Currency: United States Dollar

What type of environment are you comparing against? On-Premises Colocation

Which AWS region is ideal for your geo requirements? US East (N. Virginia)

Choose workload type: General

Servers

Are you comparing physical servers or virtual machines? Physical Servers Virtual Machines

Provide your configuration details:

Server Type	App. Name	Number of VMs	CPU Cores	Memory(GB)	Hypervisor	Guest OS	DB Engine	VM Usage (%)	Optimize By	Virtualization Host
Non DB		1 - 10000	1 - 32	1 - 256	VMware	Linux		1 - 100	RAM	Host 1: 2 CPU, 8

Total no.of VMs: + Add Row

Storage

Provide your storage footprint details

Storage Type	Raw Storage Capacity		% Accessed Infrequently	Max IOPS for Application	Backup % / Month
SAN	0 - 1000	TB		1 - 48000	0 - 100

The cost for operating a Data-center

DCs consume 3% of global electricity supply

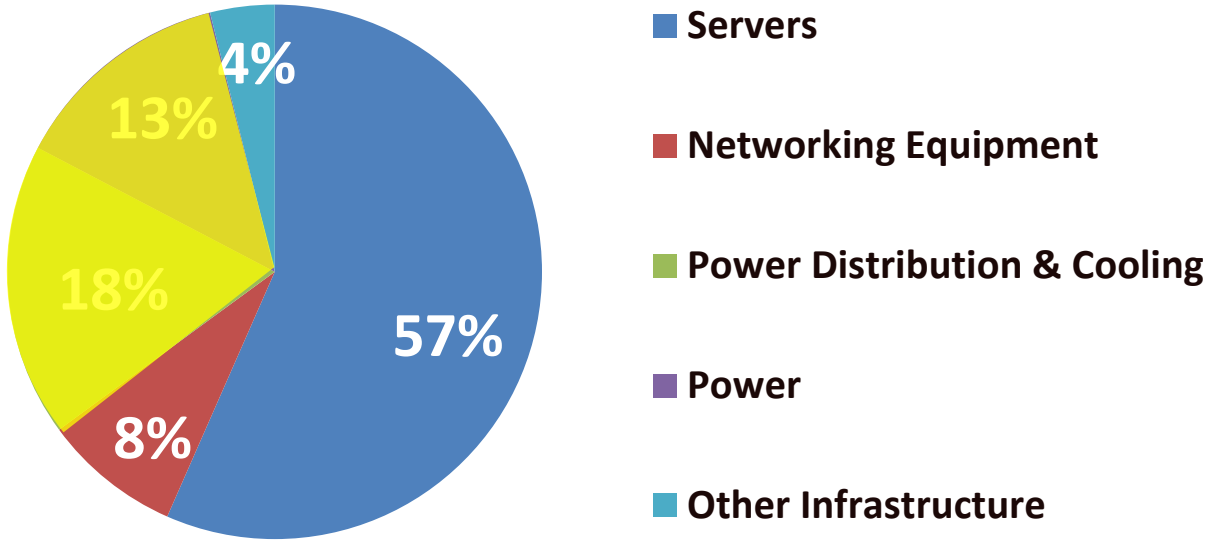
(416.2 TWh > UK's 300 TWh)

DCs produce 2% of total greenhouse gas emissions

DCs produce as much CO2 as The Netherlands or Argentina

31 % Power

Monthly costs = $3,530,920



45,978 servers, 3yr server & 10 yr infrastructure amortization

Power Usage Effectiveness (PUE)

PUE is the ***ratio*** of

- total amount of energy used by a DC facility
- to the energy delivered to the computing equipment.

PUE is the inverse of data center infrastructure efficiency.

Total facility power = covers IT systems (servers, network, storage) + other equipment (cooling, UPS, switch gear, generators, PDUs, batteries, lights, fans, etc.)

How can DC Operators Reduce Costs?

Location of the DC – cooling and power load factor.

Raise temperature of aisles

- usually 18-20 C; Google at 27C
- possibly up to 35 C (trade-off failures vs. cooling costs)

Price per Kilo Watt Hour	Where?	Possible Reason Why
3.6 cents	Idaho	Hydroelectric Power; Not Sent Long Distance
10.0 cents	California	Electricity Transmitted Long Distance over the Grid; Limited Transmission Lines in the Bay Area; No Coal Fired Electricity Allowed in California.
18.0 cents	Hawaii	Must Ship Fuel to Generate Electricity

Reduce conversion of energy

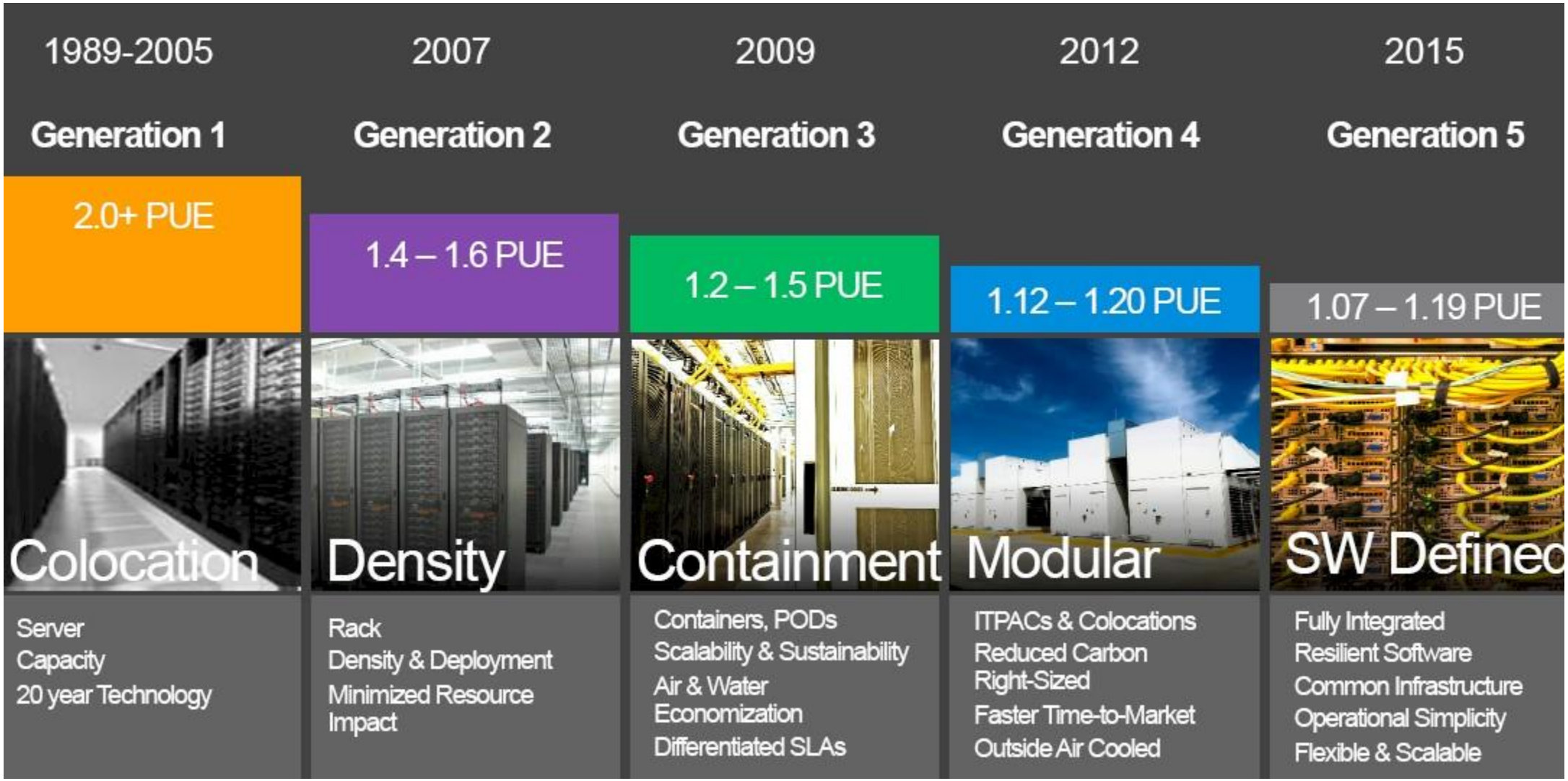
- eg Google motherboards work at 12V rather than 3.3/5V
- distributed UPS more efficient than centralised one

Go to extreme environments

- Arctic circle (Facebook)
- Floating boats (Google)
- Underwater DC (Microsoft)

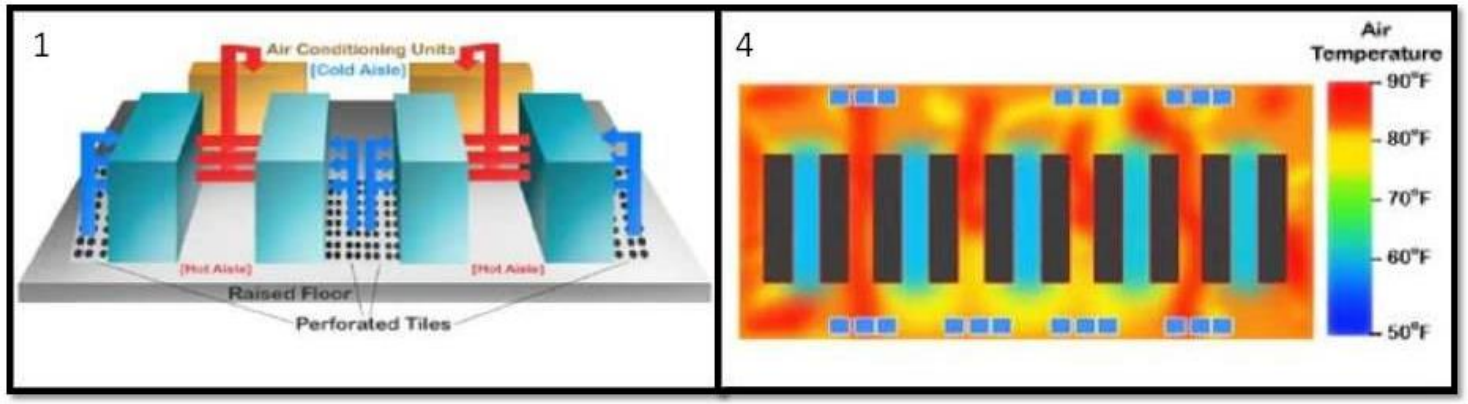
Reuse dissipated heat

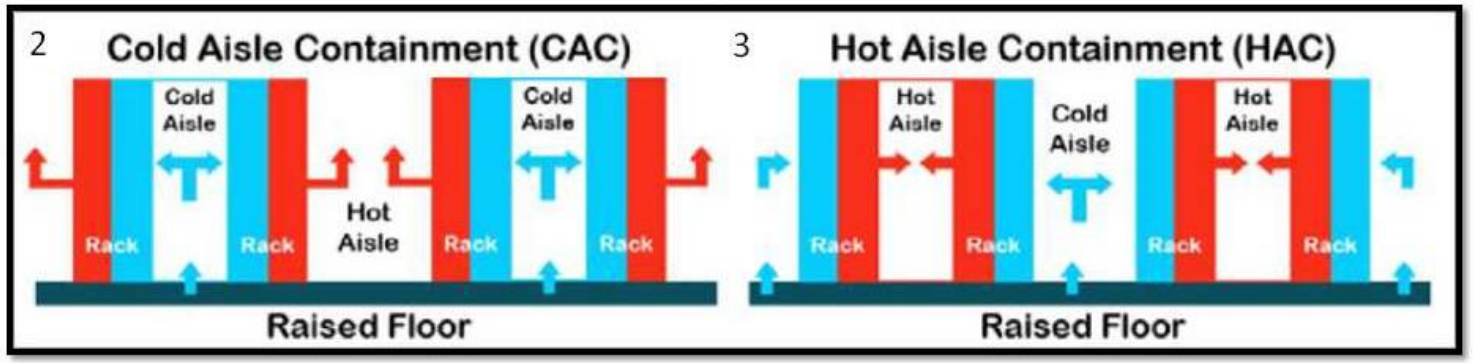
Evolution of data center design (case study Microsoft)



https://www.nextplatform.com/2016/09/26/rare-tour-microsofts-hyperscale-datacenters/

Challenge: Cooling data-centers





1- Conventional cooling
2- Cold Aisle Containment (CAC)
3- Hot Aisle Containment (HAC)
4- Thermal modelling

Cooling plant at a Google DC in Oregon

Challenge: Energy Proportional Computing

Average real-world DC and servers are too inefficient.

- The average DC wastes 2/3 or more of its energy.

Energy consumption not proportional to load

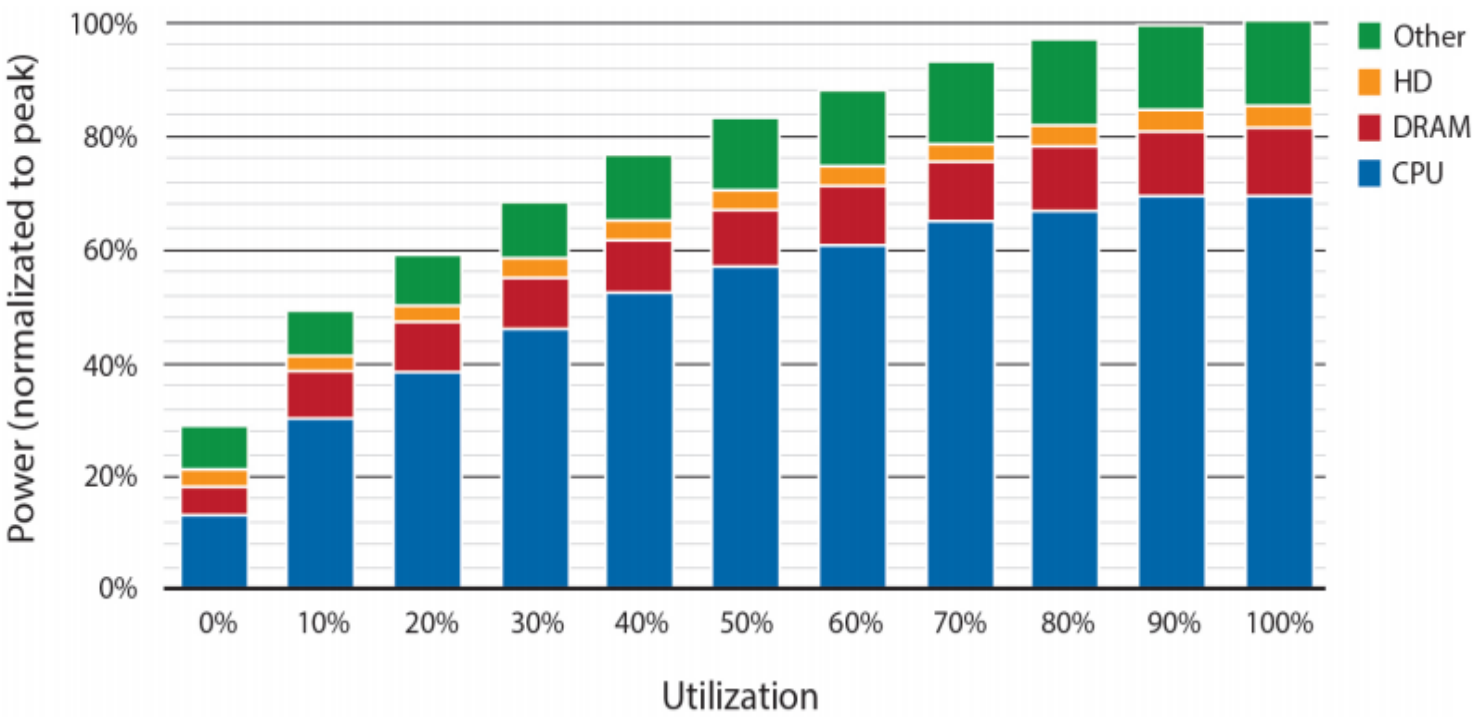
- CPUs not so bad but other components are
- CPU is the dominant energy consumer in servers – using 2/3 of energy when active/idle.

Try to optimise workloads

- On is better than off

Virtualisation to consolidate service on fewer servers

Sub-system power usage in an x86 server as the compute load varies from idle to full (reported in 2012).

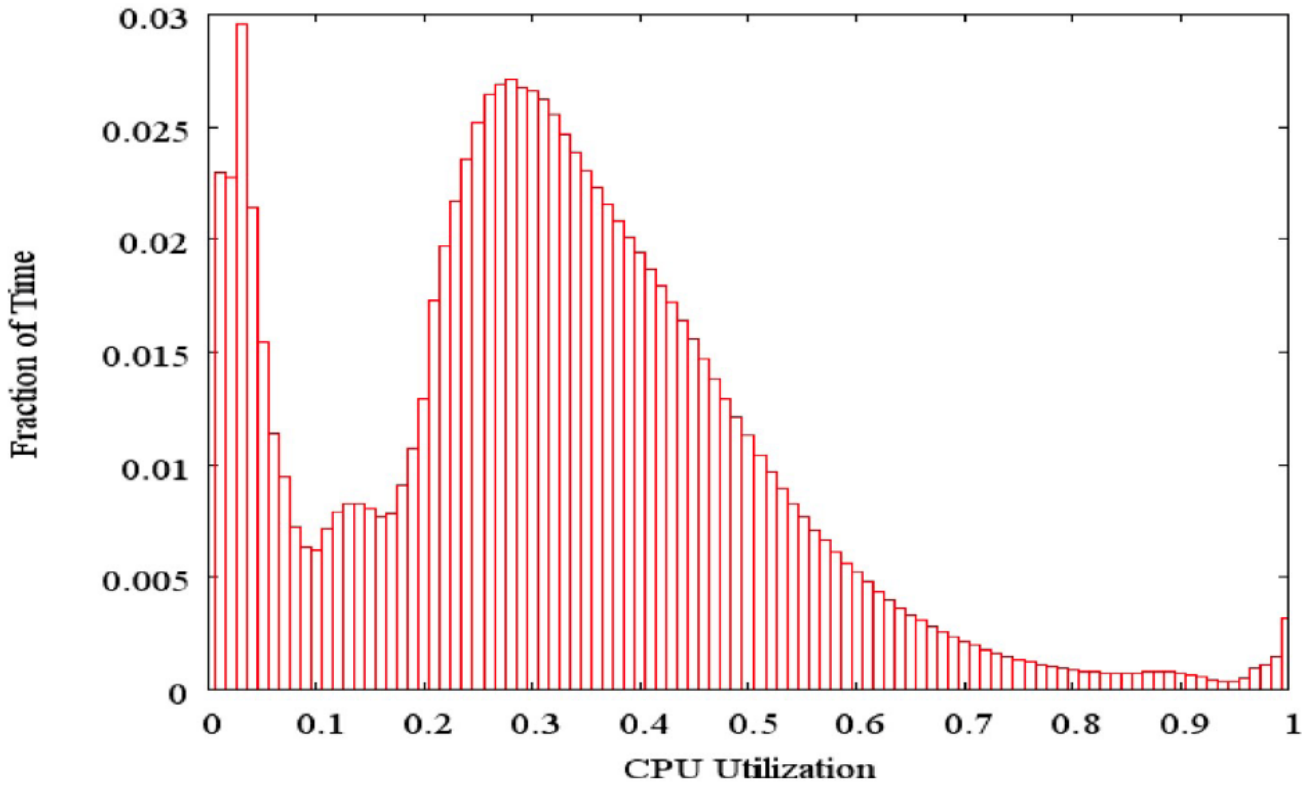


src: "The Datacenter as a Warehouse Computer"

Challenge: Managing a data-center and its resources

Servers idle most of the time

- For non-virtualized servers 6-15% utilization.
- Virtualization can increase it to an average utilization ~30%

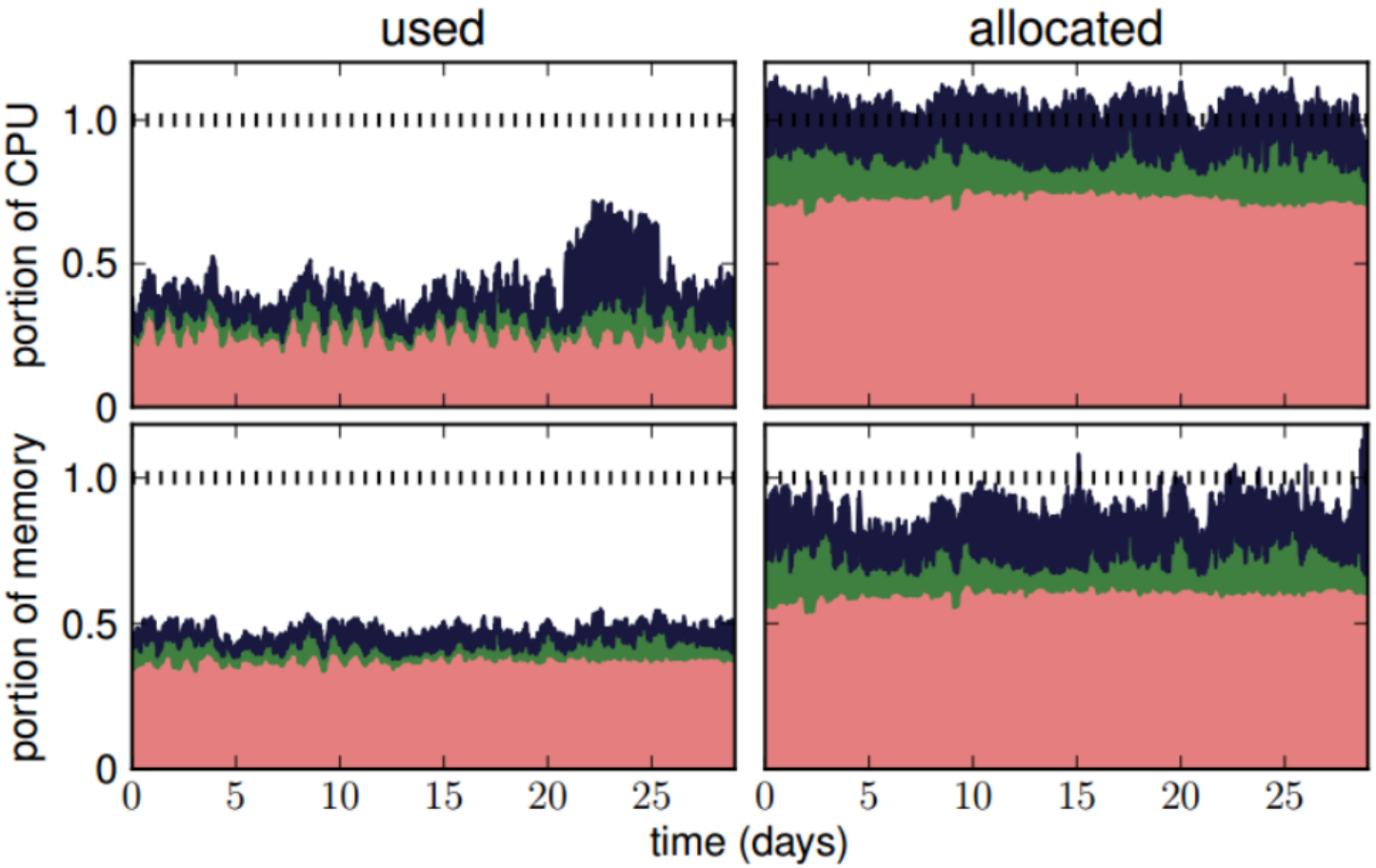


src: Luiz Barroso, Urs Hölzle "The Datacenter as a Computer"

Need for resource pooling and application and server consolidation

Need for resource virtualization

Challenge: Managing a data-center and its resources



src: "Heterogeneity and dynamicity of clouds at scale: Google trace analysis" SoCC'12

Even with virtualization and software defined DC, resource utilization can be poor.

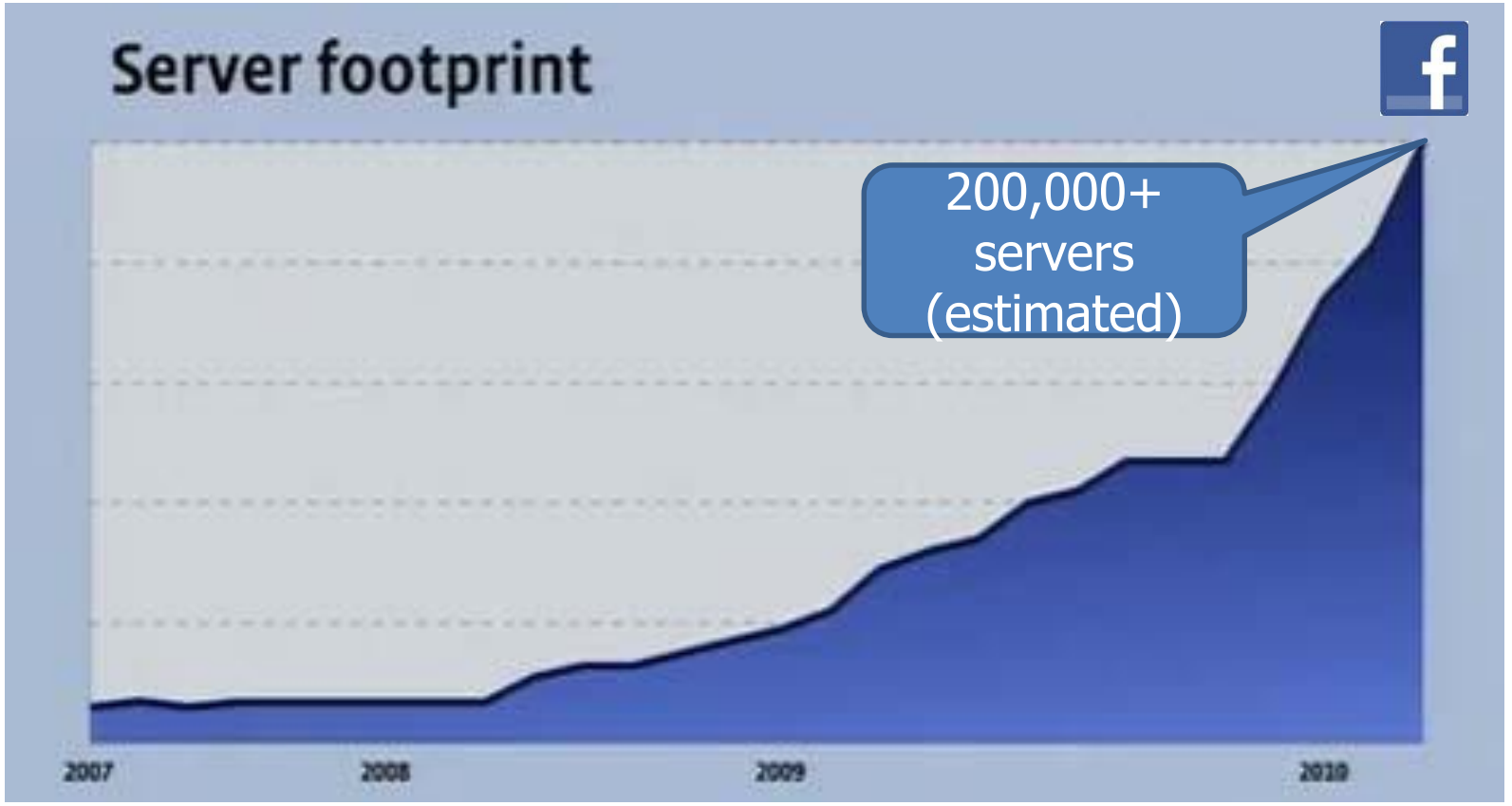
Need for efficient monitoring (measurement) and cluster management.

Goal to meet SLOs and SLIs. Job's tail latency matters!

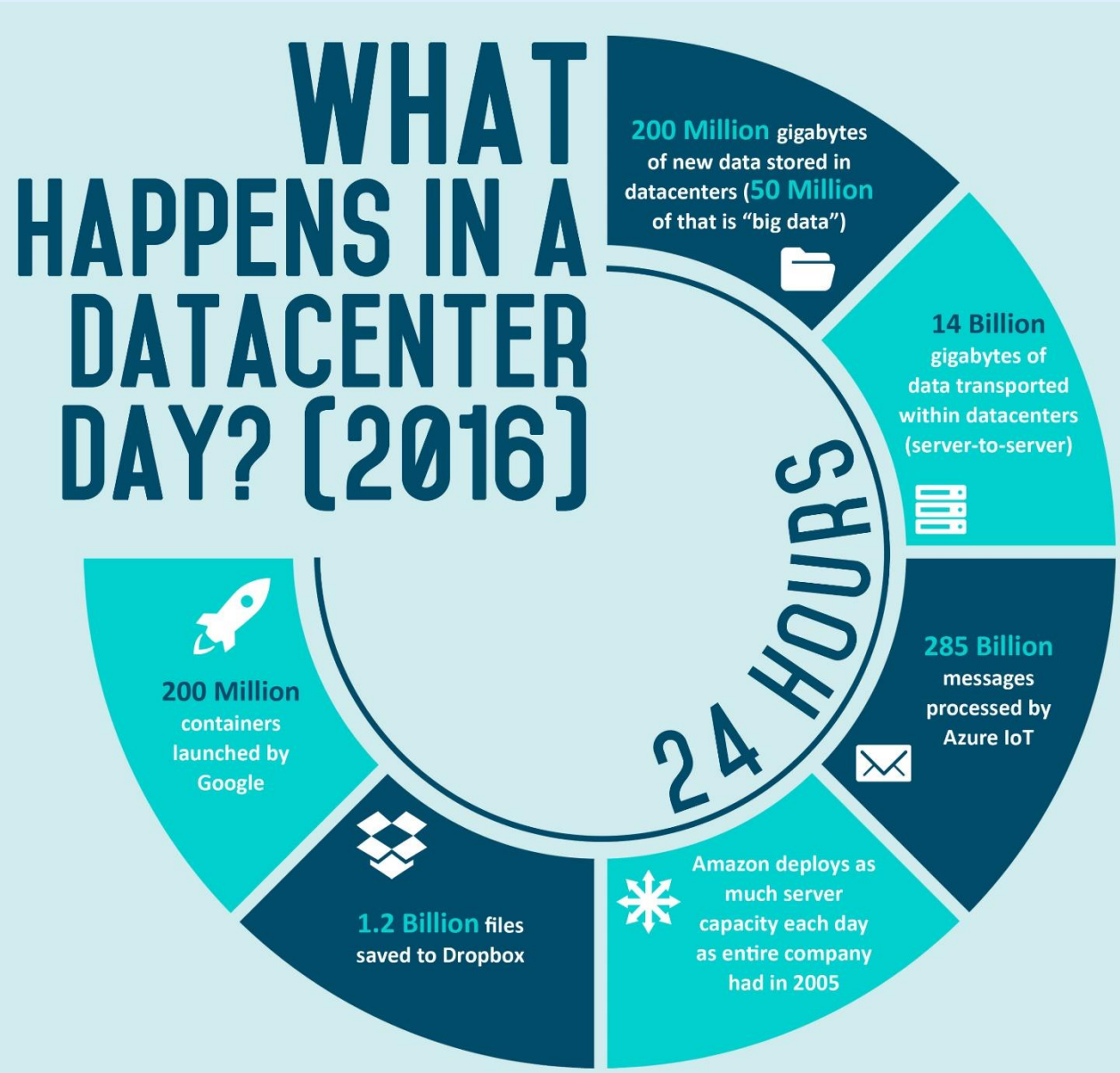
Challenge: Managing the scale and growth

In 2016, Gartner estimated that Google has 2.5 million servers.

In 2017, Microsoft Azure was reported to have more than 3 million servers.



Size and growth of Data Centers (2016 – 2020)



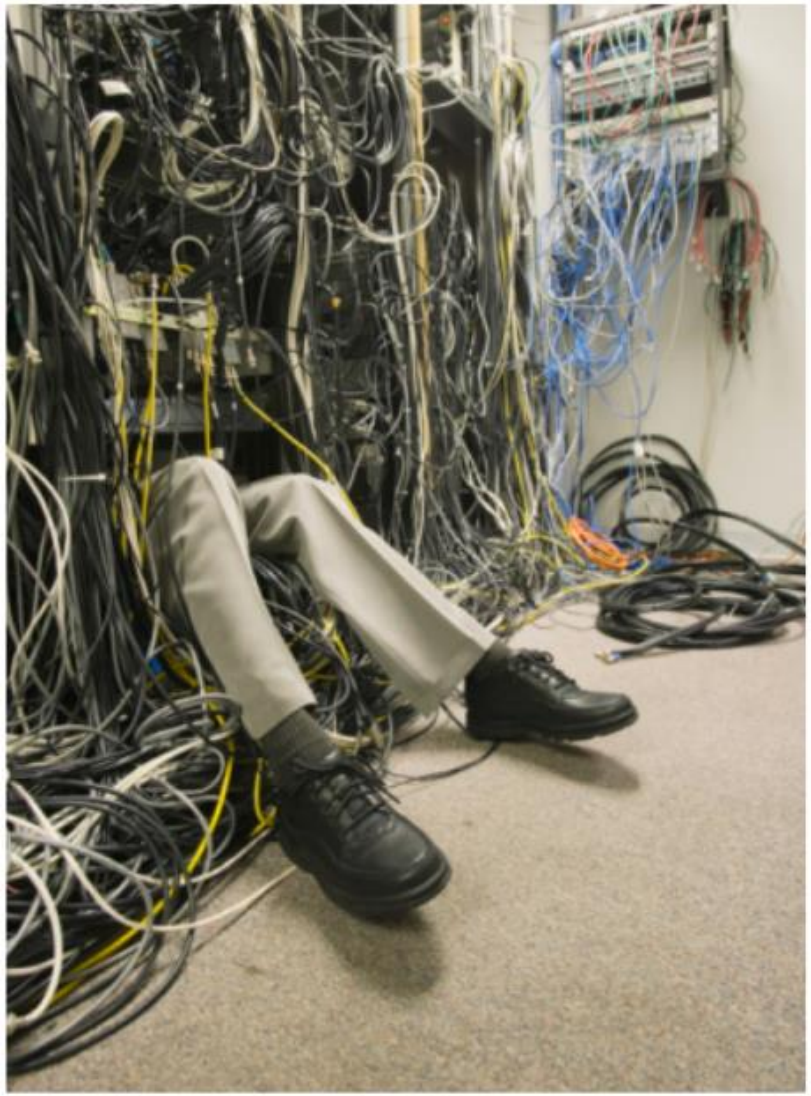
Source: Cisco Global Cloud Index and Public Sources, 2016

The scale and complexity of DC operations grows constantly.

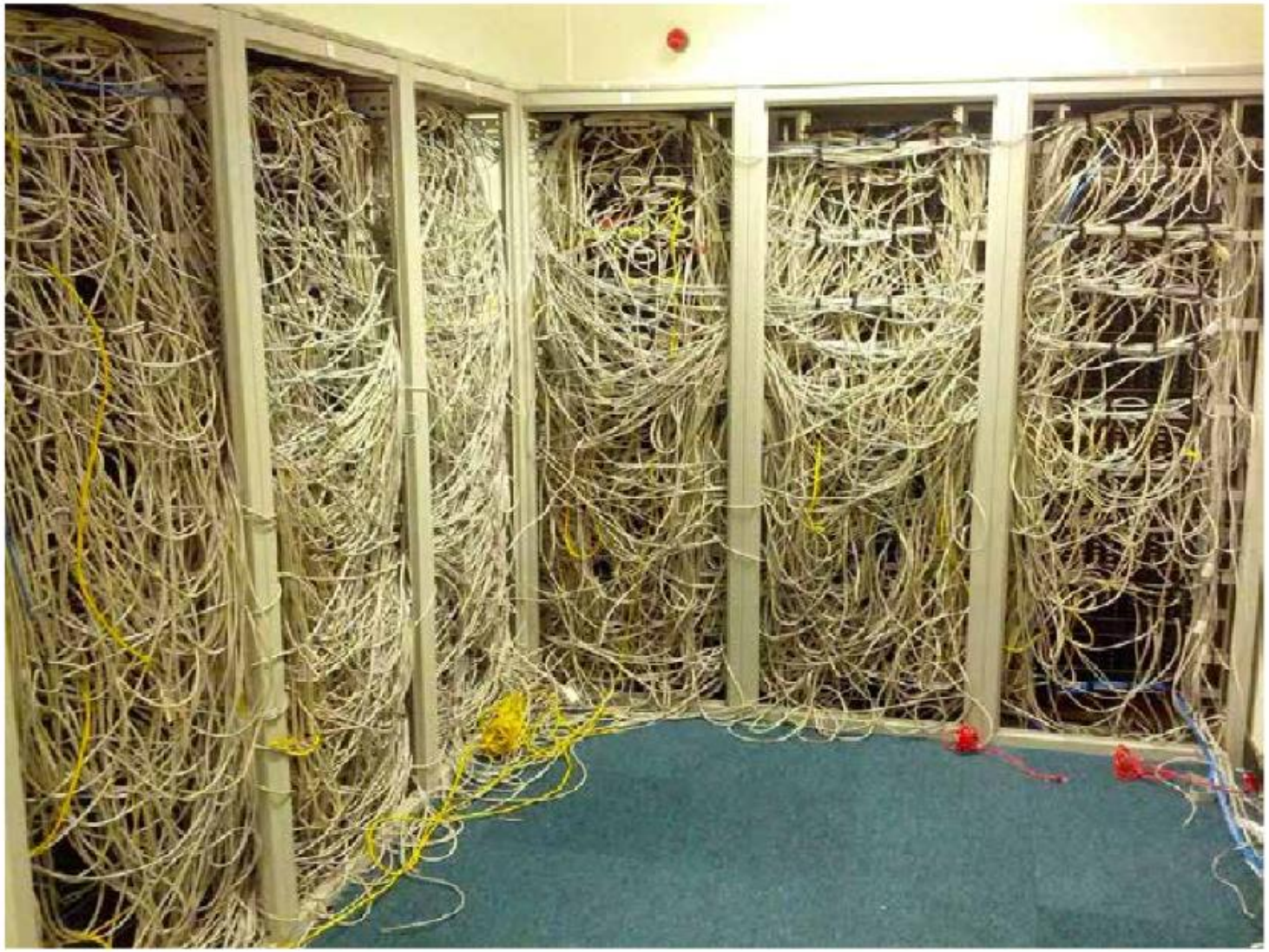
By 2020, we expect to have 600 million GB of new data saved each day (200m GB big data)

→ the volume of big data by 2020 will be as much as all of the stored data today!

Challenge: Networking at Scale



[David Samuel Robbins, gettyimages.ch]



[@AlexCWheeler, Twitter]

Challenge: Networking at scale (cont.)

Building the right abstractions to work for a range of workloads at hyperscale

Software Defined Networking (SDN)

Within DC, 32 billion GBs will be transported by 2020.

src: Cisco report 2016-2026

Google's "machine-to-machine" traffic is several orders of magnitude larger than what goes out to the Internet.

src: "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network" (ACM SIGCOMM'15).