



CLOUD COMPUTING

Intro to Cloud: Background, Terminology, Basics, Data Centers

PAUL TOWNEND

ASSOCIATE PROFESSOR, UMEÅ



A LITTLE BIT ABOUT ME



Originally from Leeds, UK. Joined Umeå in August 2020

PhD in Fault-Tolerant Grid Systems (2006)

General Chair of 13 IEEE conferences, TPC member of 28 IEEE conferences

Co-founded and worked for 3 years as CTO of spin-out company

1500+ citations, 70+ peer-reviewed publications in Distributed Systems

Member of the WASP Graduate School Management board

Examiner for this course and Learning Feature Representations

INDUSTRY COLLABORATIONS



EXTERNAL / INTERNATIONAL ACTIVITY

Leader	University of Leeds Distributed Systems and Services group (2004-2017) UK-China COLAB collaboration (Leeds, Beihang, Chongqing, NUDT) (2006 – present)			
PhD examiner	University of Newcastle, UK (2017-2018)			
Editor	IEEE Transactions on Services Computing 12(1) (<i>Guest editor</i> , 2019) ACM Transactions on Embedded Systems (<i>Late</i> 2015)		Philosophical Transactions of the Royal Society A: Cloud Computing (<i>January</i> 2013) IEEE Distributed Systems Online Journal, Dependability Topic (<i>2004-2007</i>)	
Steering committee	IEEE ISORC	IEEE SOSE	IEEE JCC	
General Chair	IEEE BigDataService 2021 (Online)	IEEE SOSE 2015 (San Francisco, USA)	IEEE SOSE 2016 (Oxford, UK)	
	IEEE ISORC 2018 (Singapore)	IEEE RTCPS 2014 (Reno, USA)	IEEE ISORC 2015 (Auckland, NZ)	IEEE iVCE 2013 (San Francisco, USA)
	IEEE ISORC 2016 (York, UK)	WODSOG 2006 (Leeds, UK)	GCC 2014 (London, UK)	
PC Chair	IEEE BigDataService 2020 (Oxford, UK)	IEEE SOSE 2014 (Oxford, UK)	UK e-Science AHM 2011 (York, UK)	
	IEEE JCC 2020 (Oxford, UK)	IEEE iVCE 2012 (Shenzhen, China)		
Workshop Chair	IEEE ISORC 2014 (Reno, USA)			
Publicity Chair	IEEE ISORC 2013 (Paderborn, Germany)	IEEE SRDS 2007 (Beijing, China)		
Local Chair	RNEC 2008 (Leeds, UK)	IEEE SRDS 2006 (Leeds, UK)		
PC member	IEEE SOSE 2019 (San Francisco, USA)	IEEE EUC 2014 (Milan, Italy)	IEEE SmartIoT 2018 (Xian, China)	IEEE SEUS 2014 (Reno, USA)
	IEEE SOSE 2018 (Bamberg, Germany)	IEEE ICDCS 2013 (Philadelphia, USA)	IEEE SOSE 2017 (San Francisco, USA)	IEEE SOSE 2013 (San Francisco, USA)
	IEEE BDCAT 2018 (Zurich, Switzerland)	IEEE CSE 2013 (Sydney, Australia)	IEEE iThings 2015 (Sydney, Australia)	IEEE ISORC 2012 (Shenzhen, China)
	IEEE BDSEA 2016 (Shanghai, China)	IEEE DSN 2012 (Boston, USA)	IEEE CCBd 2015 (Taipei, Taiwan)	IEEE SOSE 2011 (Los Angeles, USA)
	IEEE iVCE 2015 (San Francisco, USA)	IEEE HASE 2008 (Nanjing, China)	IMTIC 2015 (Jamshoro, Pakistan)	SDMCMM 2012 (Montreal, Canada)
	IEEE DSAA 2015 (Paris, France)	IEEE ICEBE 2019 (Shanghai, China)	IEEE iVCE 2014 (Oxford, UK)	IEEE CCBd 2015 (Taipei, Taiwan)

ABOUT THIS MODULE

Fairly lecture heavy , but hopefully we can have discussion along the way

Light assignment – your major work should be your PhD

Invited keynotes from Ericsson Research and Google USA

A practical session to introduce the Ericsson Research Data Center and OpenStack

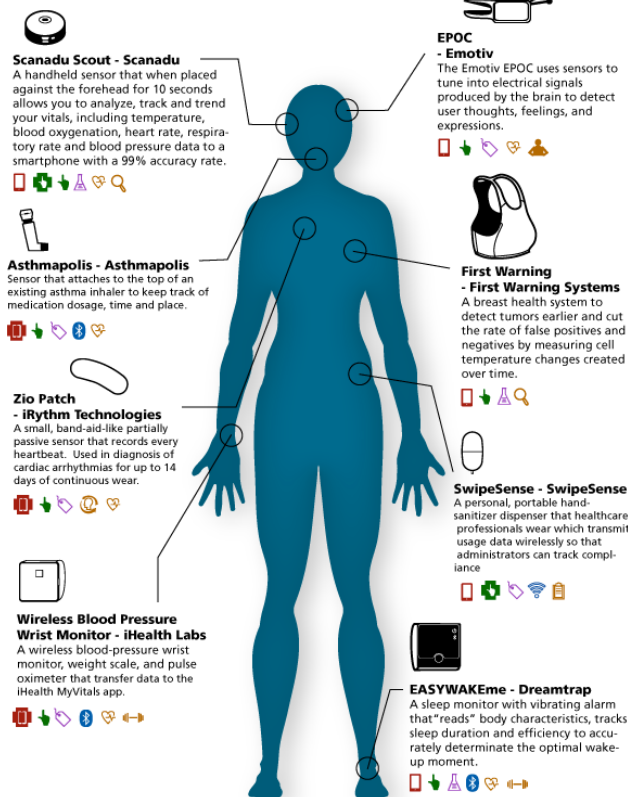
Ultimate goal: learn at least one thing that can inform/be applied to your PhD

SCHEDULE

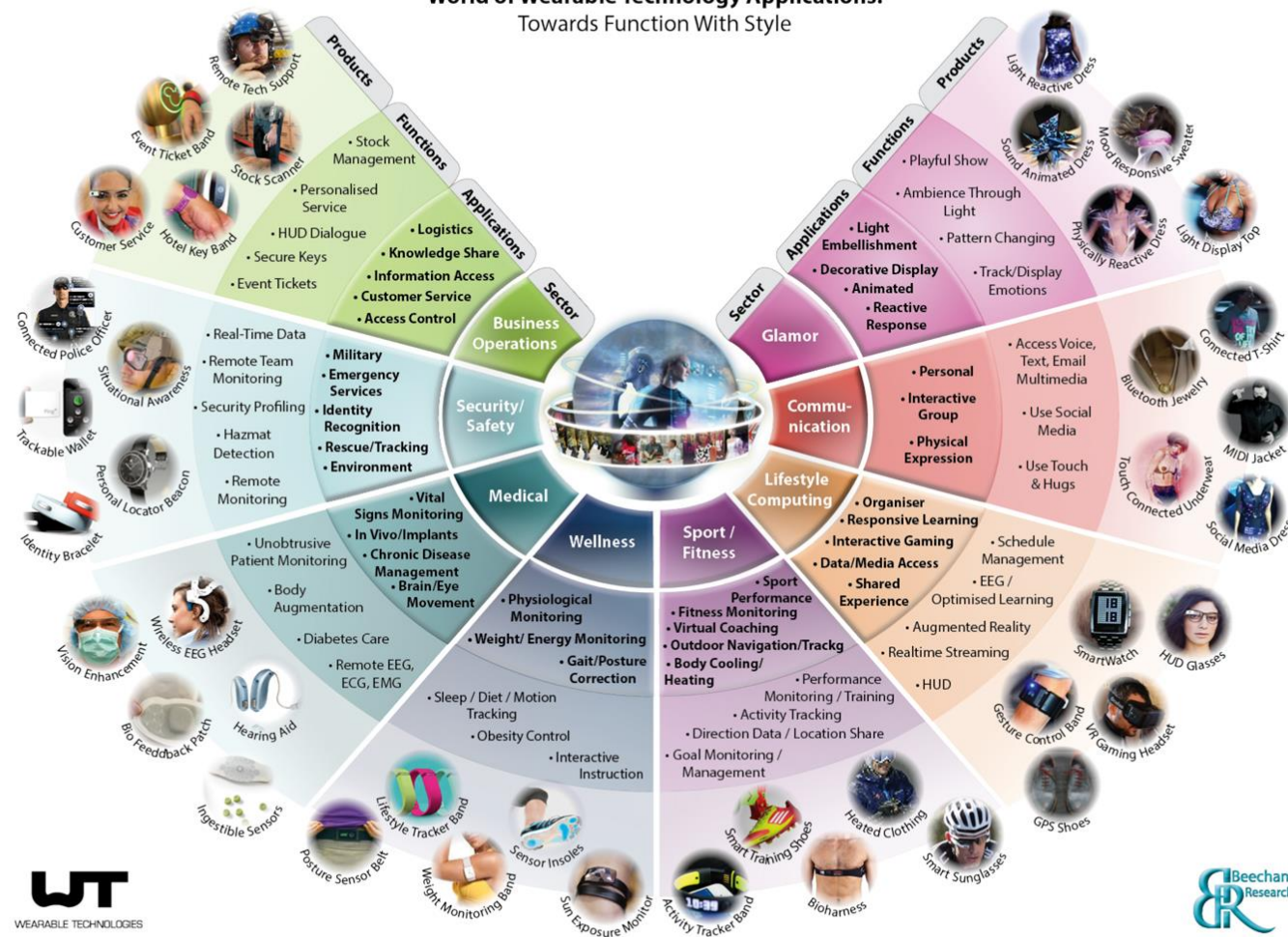
DAY ONE		May 4, 2021
09:30 – 09:45		Welcome to the Course
09:45 – 11:15	L1	Introduction to Cloud: Background, terminology, basics, data centres
11:15 – 11:45		BREAK
11:45 – 12:45	L2	Distributed processing: Hadoop
12:45 – 14:00		BREAK (LUNCH)
14:00 – 15:00	L3	Distributed processing: Apache Spark and Apache Storm
15:00 – 15:30		BREAK
15:30 – 16:45	L4	Edge, Fog, and Serverless Computing

DAY TWO		May 5, 2021
09:30 – 10:45	L5	Cloud Orchestration
10:45 – 11:15		BREAK
11:15 – 11:45	K1	Keynote: Johan Eker (Ericsson)
11:45 – 12:30	P1	Cloud Practical Session (ER DC)
12:30 – 14:00		BREAK (LUNCH)
14:00 – 15:00	P2	Cloud Practical Session (ER DC)
15:00 – 16:00	L6	Cloud Economics
16:00 – 16:30		BREAK
16:30 – 17:30	K2	Industry Keynote: “Challenges & Opportunities in Cloud” - Steve Webster, Google
17:30 – 17:45		Closing remarks and assignment information

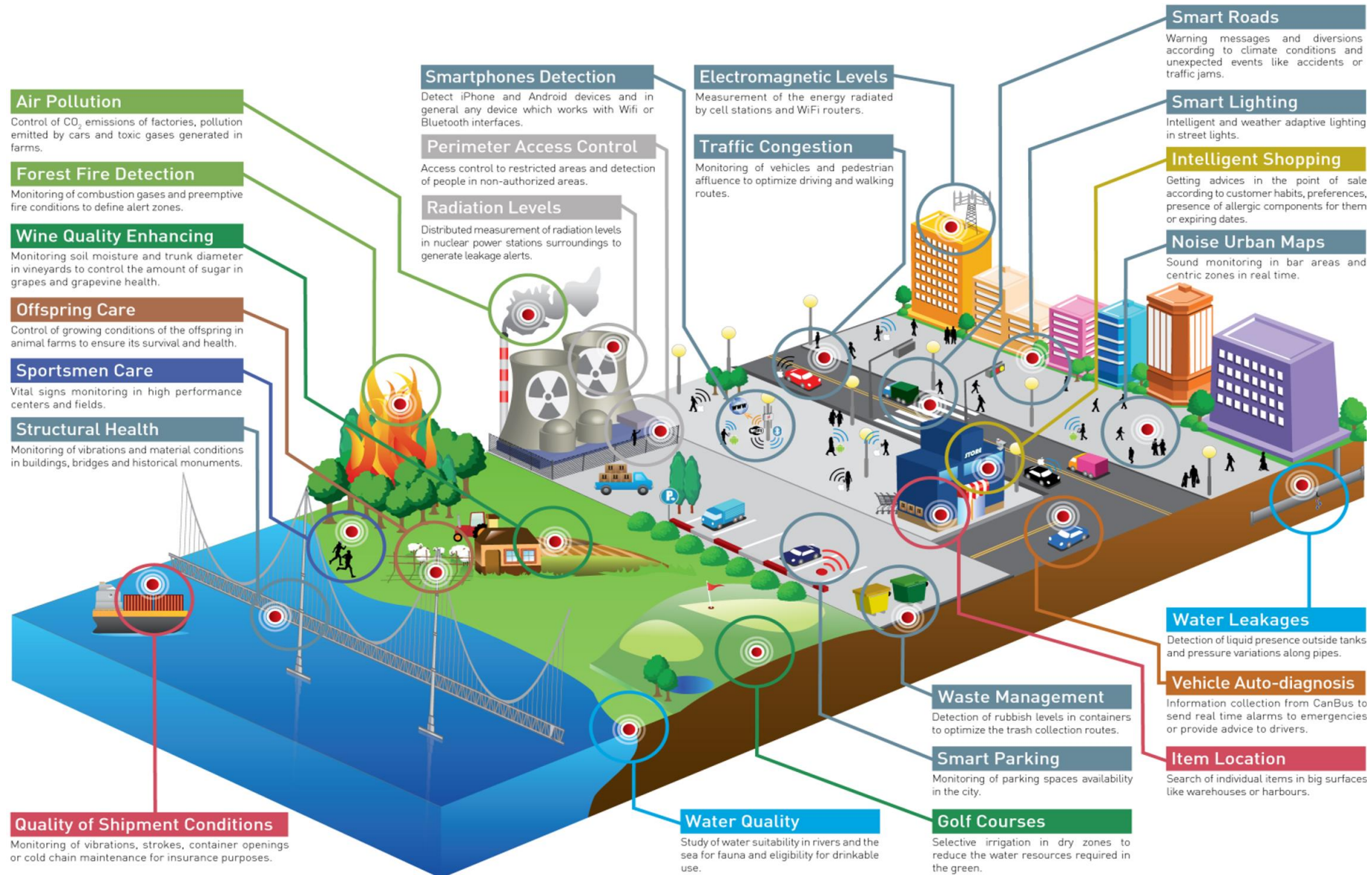
Why study Cloud?



World of Wearable Technology Applications: Towards Function With Style



© 2014 Beecham Research Ltd. & Wearable Technologies AG



TRENDS IN MODERN DISTRIBUTED SYSTEMS

Need good
security

Highly distributed
(often mobile)

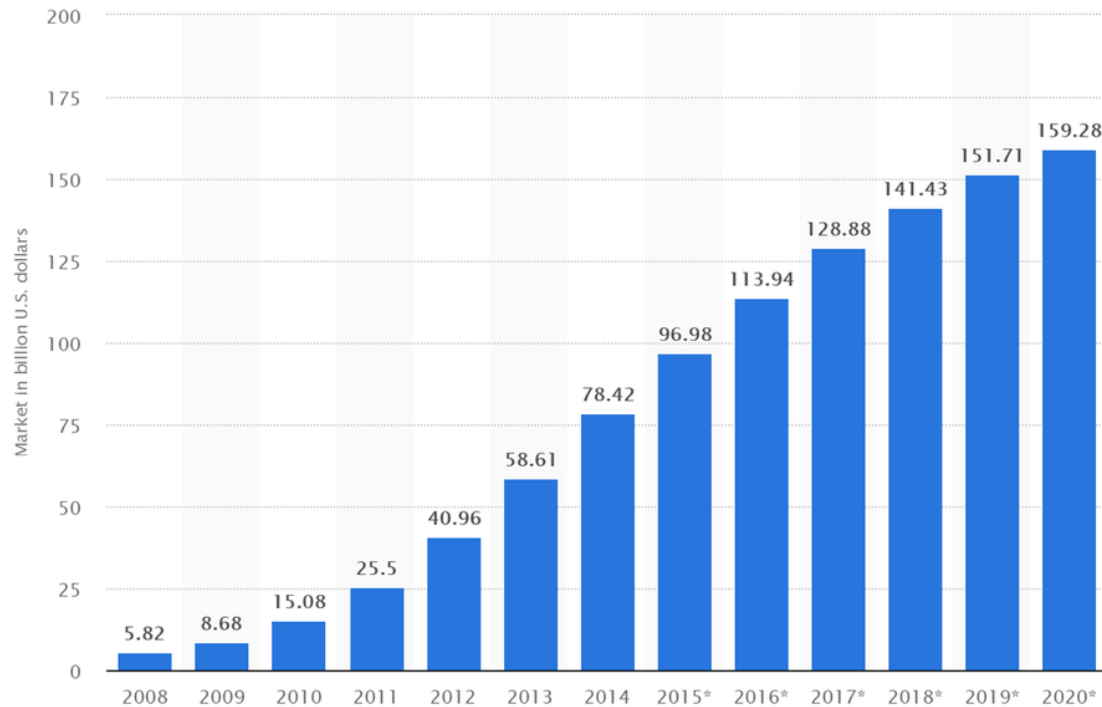
Generate large amounts of
data

Latency can
be critical

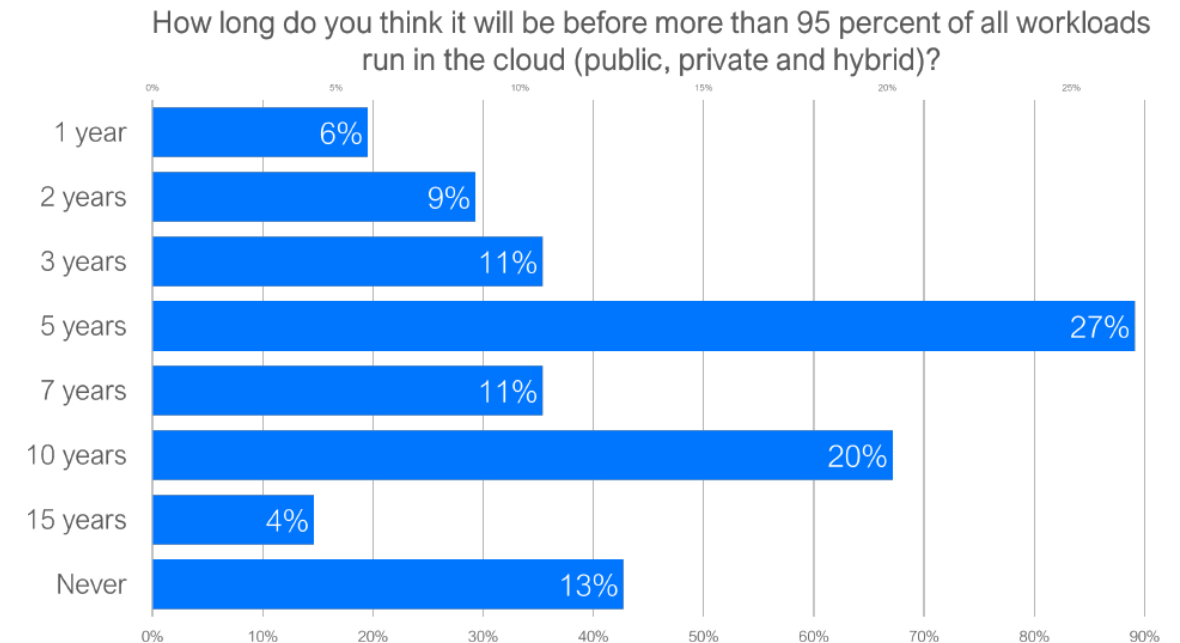
Many small / low powered
devices

Need support for analytics
and decision making

Total size of the public Cloud Computing market from 2008-2020 (Billions USD)

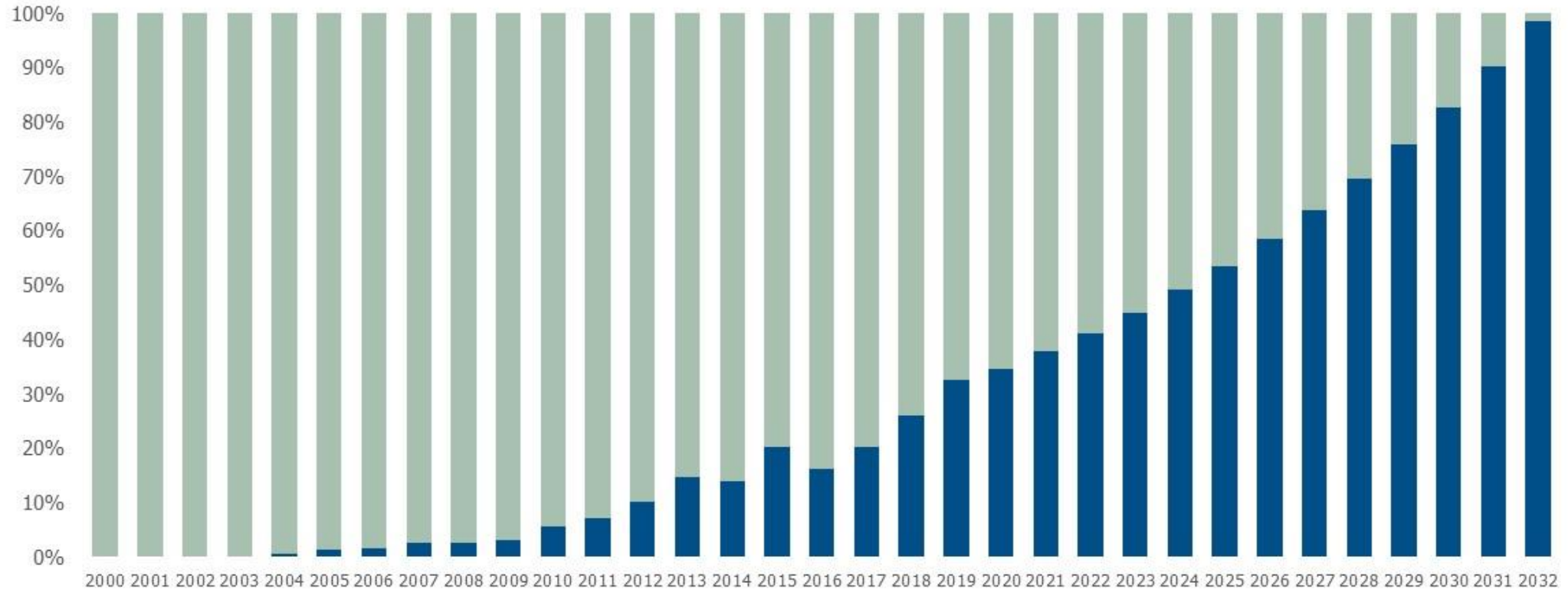


Forbes: 83% of Enterprise Workloads will be in the Cloud by 2020



Cloud is eating software

Cloud will become majority of software market within 5 years



Source: CapIQ; Bessemer Venture Partners analysis;
Cloud CAGR – 20%, Software CAGR – 10%

■ Software ■ Cloud



What is Cloud?



App Definition and Development

Database

Streaming & Messaging

Application Definition & Image Build

Continuous Integration & Delivery

Orchestration & Management

Scheduling & Orchestration

Coordination & Service Discovery

Remote Procedure Call

Service Proxy

API Gateway

Service Mesh

Runtime

Cloud Native Storage

Container Runtime

Cloud Native Network

Provisioning

Automation & Configuration

Container Registry

Security & Compliance

Key Management

Platform

Certified Kubernetes - Distribution

Certified Kubernetes - Hosted

Certified Kubernetes - Installer

PaaS/Container Service

Observability and Analysis

Monitoring

Logging

Tracing

Chaos Engineering

Serverless

CNCF Serverless Landscape

See the serverless interactive display at [s.cncf.io](#)

CLOUD NATIVE Landscape

l.cncf.io

CLOUD NATIVE COMPUTING FOUNDATION

Redpoint

Amplify

This landscape is intended as a map through the previously uncharted terrain of cloud native technologies. There are many routes to deploying a cloud native application, with CNCF Projects representing a particularly well-traveled path.

Special

Kubernetes Certified Service Provider

Kubernetes Training Partner

Members

Members

Members

WASP

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

NIST DEFINITION

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.

You do not own computing infrastructure – you rent it

MAJOR CLOUD CHARACTERISTICS

On-demand self-service

Can provision resources without human intervention

Broad network access

Service is accessible through the network

Resource pooling

Resources are pooled to several customers

Rapid elasticity

Can rapidly increase and decrease capacity

Measured service

Resource usage can be monitored and reported



MAJOR CLOUD CHARACTERISTICS

Virtualization

Physical machines are split into “virtual” components

Multi-tenancy

Multiple jobs / users use the same physical machine

Service model

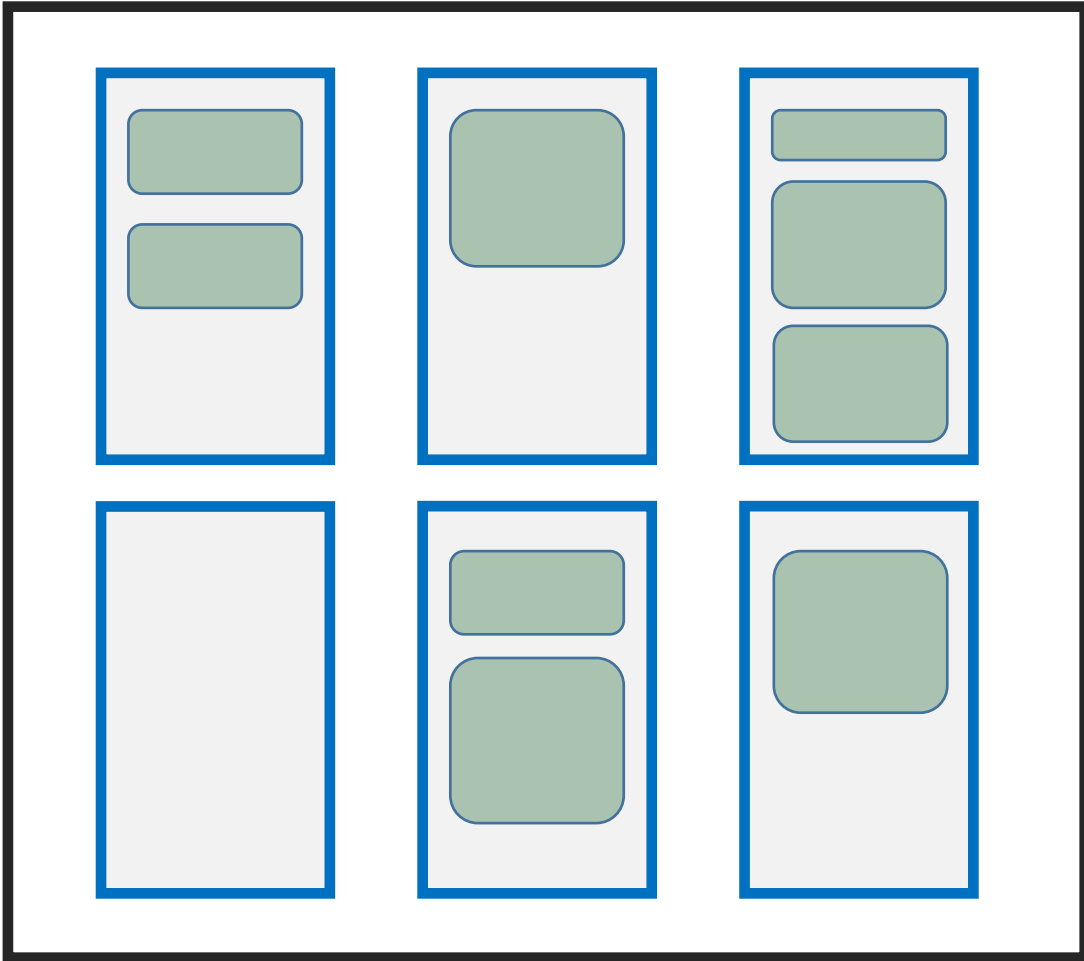
Multiple different ways Clouds can be used

Pay-per-use

Related to on-demand resources: users only pay for the resources that they actually use



VIRTUALISATION



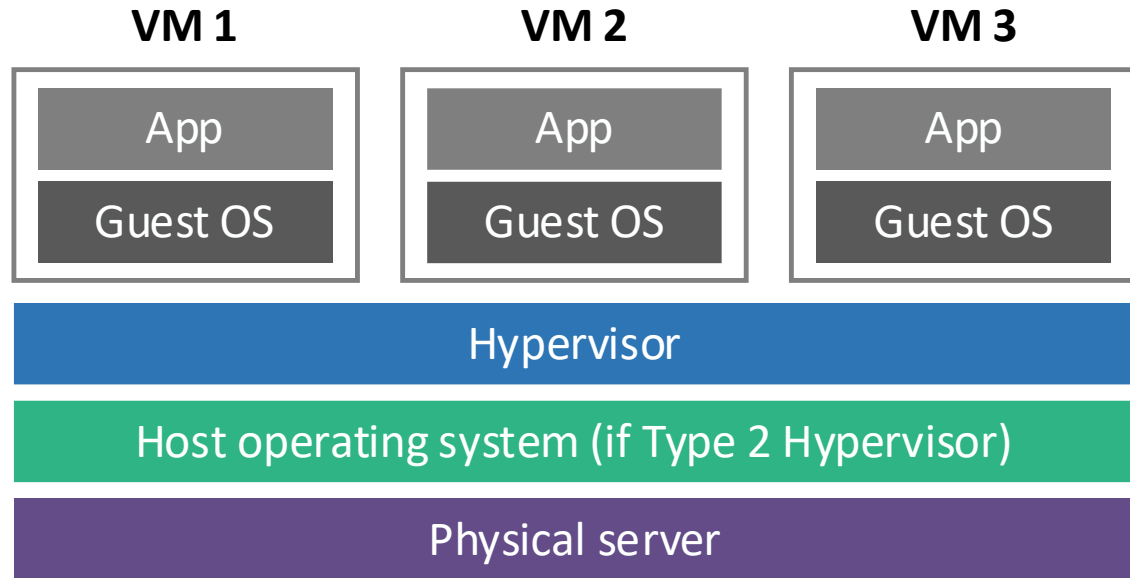
Cloud

Partition physical machines to maximise resource utilisation and elasticity

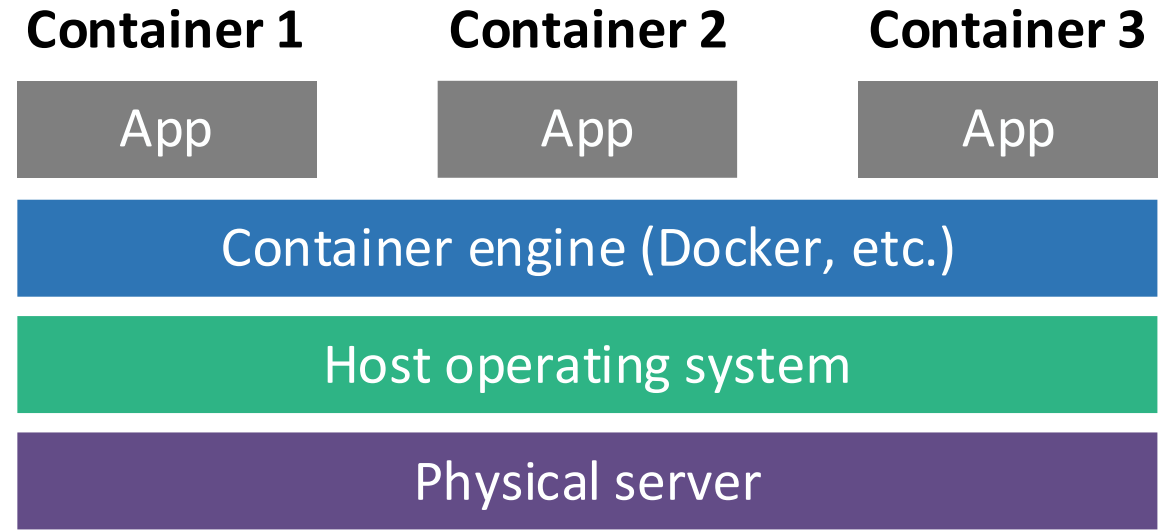
Subject to SLA

Resource management becomes critical

VIRTUAL MACHINES AND CONTAINERS (1)

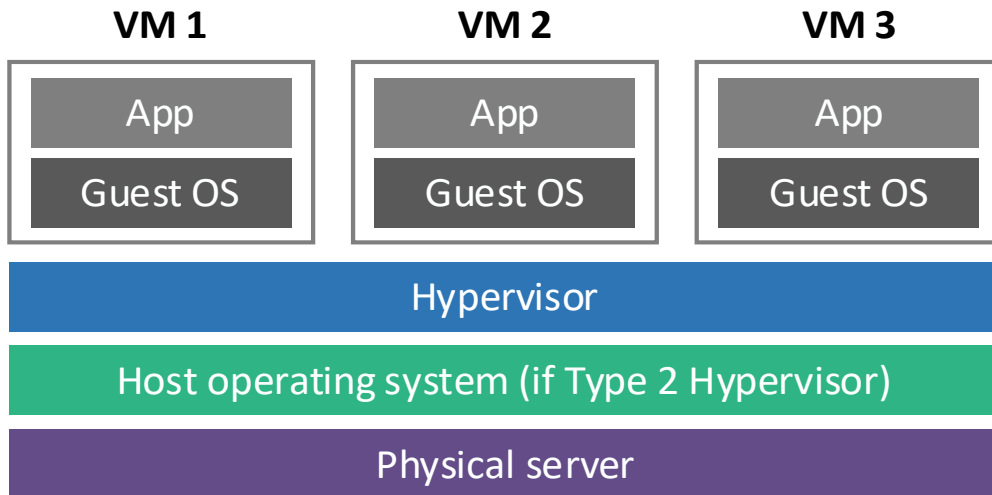
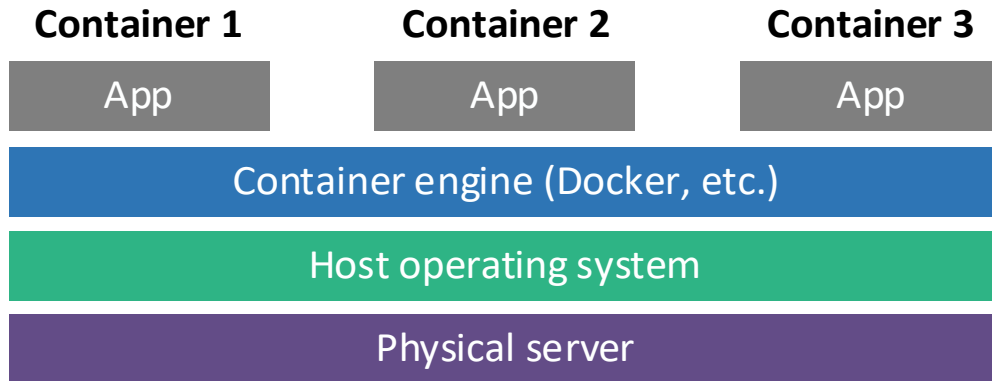


VMs are like neighbours



Containers are like housemates

VIRTUAL MACHINES AND CONTAINERS [2]



“Light”

Fast to create

Do not have migration

Typically short lived and single function

“Heavy”

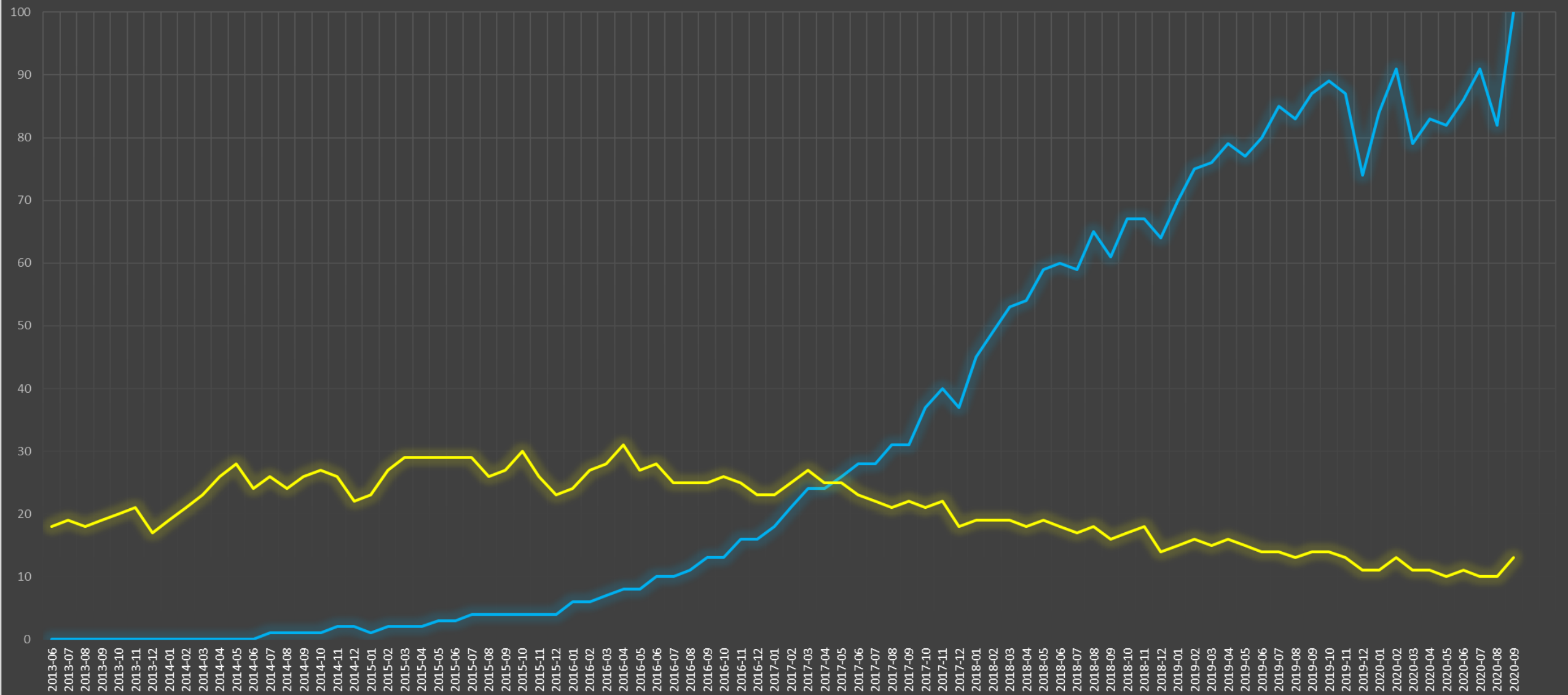
Slow to create

Slow to migrate

Typically long lived and multi-function

Global Search: Kubernetes vs OpenStack 2013-2020

kubernetes: (Worldwide) openstack: (Worldwide)



Cloud and Industry

BENEFITS FROM AN INDUSTRY PERSPECTIVE

Move from capital expenditure to operational expenditure

Pay only for what is used (in some cases)

The **illusion** of having infinite resources

Offloads responsibility for security (physical and virtual) plus uptime, efficiency, maintenance, etc.

AlfaPeople

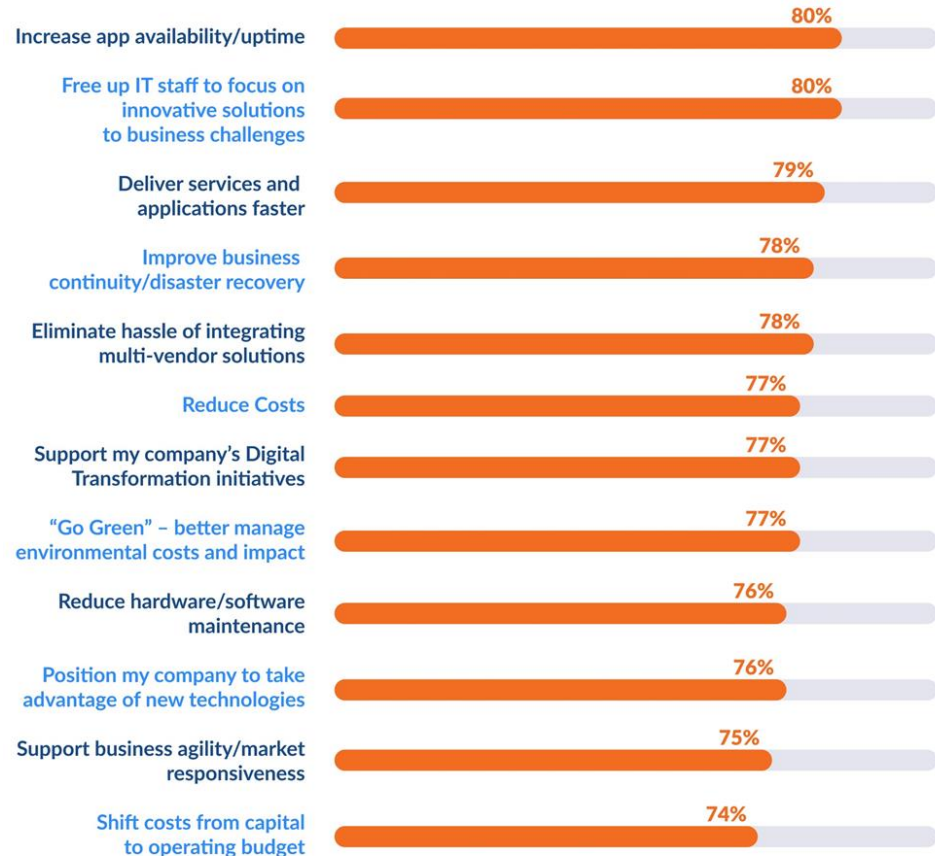
Moving your Business to the Cloud

- 01. Anywhere, Anytime**
Shift from traditional in-house hardware to internet-based functions, providing your staff with instant access to data, inventory, and back-office functions.
- 02. Eliminate Disconnect**
With Microsoft Dynamics 365 cloud-based system, your ERP and CRM work in unison, giving you a complete, accurate view of your business — from inventory to quarterly reports — all in one place.
- 03. Reduce Overhead**
Avoid upfront investments in, for example, data center infrastructure and the ongoing maintenance of traditional IT hardware.
- 04. User-Friendly Features**
Access to an easy-to-use dashboard and customizable features, similar to the email and social media platforms you use every day.

Directly connect to one of our Dynamics cloud specialists to learn more.
Contact us at 855-732-6484 or email info.us@alfapeople.com

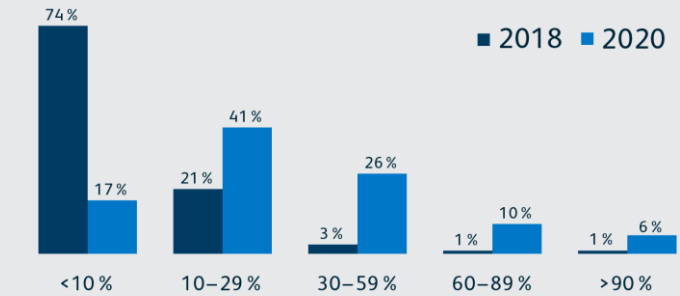
CLOUD INDUSTRY DRIVERS

Business Drivers for Implementing Cloud Solutions, Global 2019

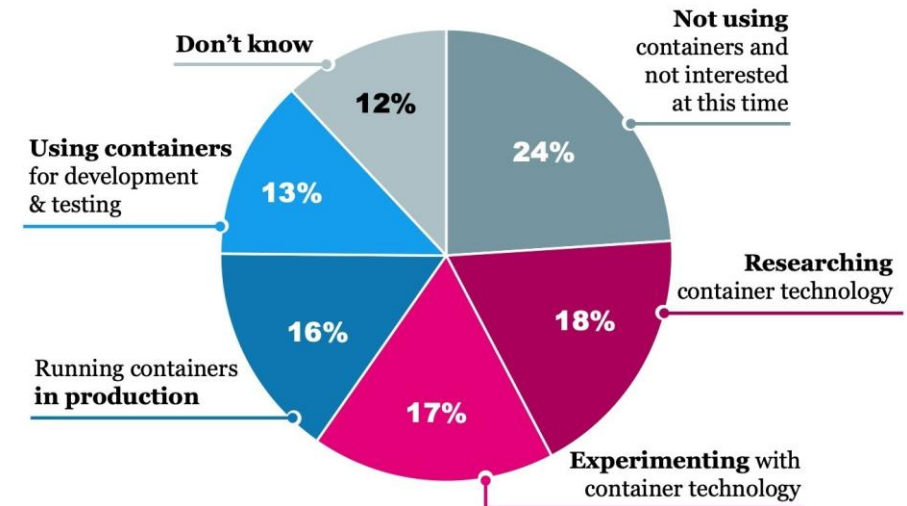


Reference: Succeed in the Digital Era with Cloud Communications, Frost & Sullivan, 2019

How many of your applications come from the cloud?

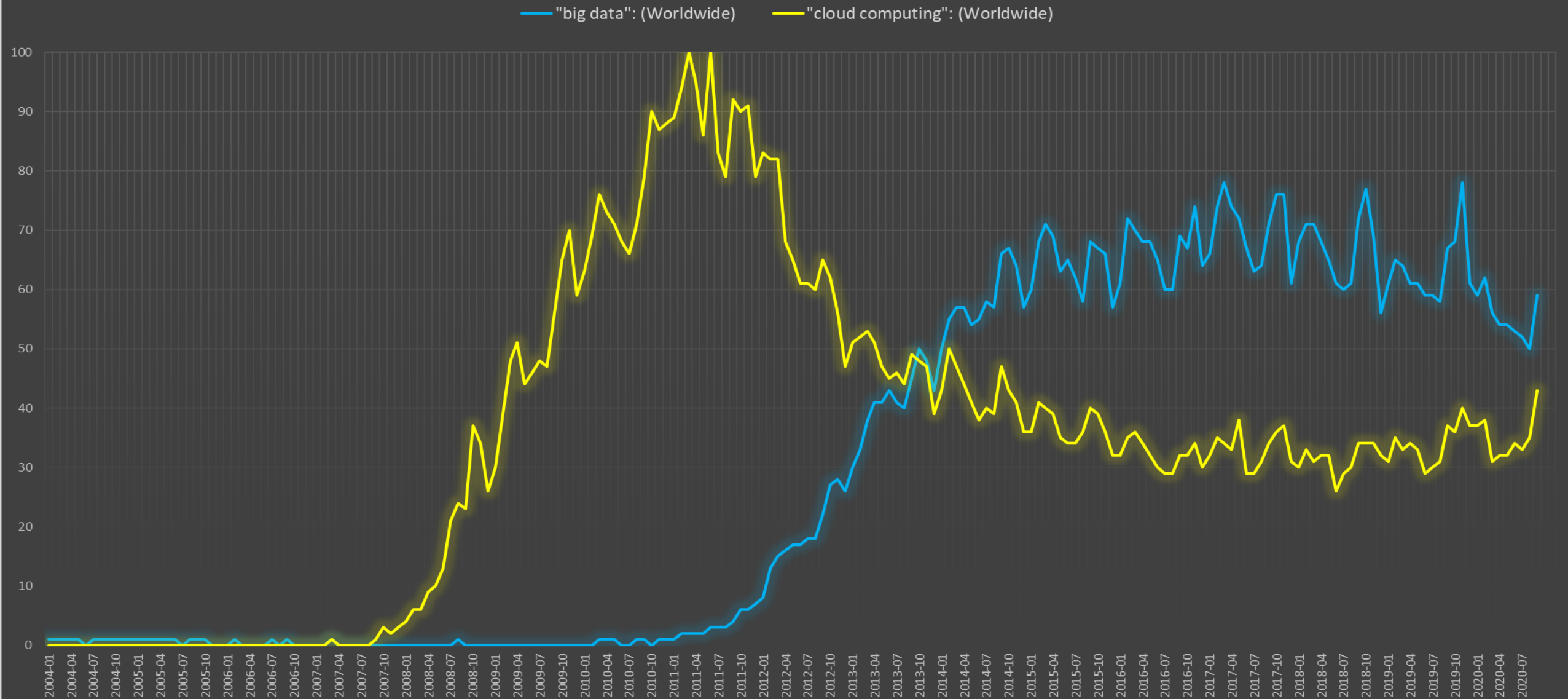


Maturity Survey: 2018 Cloud Computing



Q. Which of the following best describes your organization's use of container technology today?

Big Data vs Cloud Computing



EXAMPLES OF CLOUD PROVIDERS



Google Cloud Platform



 Alibaba Cloud



ORACLE®
CLOUD



Hyperscalers

Smaller, often co-locational

PROBLEMS WITH USING THE CLOUD

Customer perspective

Privacy and security: less control

Legal problems: who does the data belong to? etc

Designing scalable applications

Communication latency

Risk of lock-in: difficult to migrate to other providers

Ability to negotiate and manage the SLA

Provider perspective

SLAs: how much does it cost to respect them?

Cloud interoperability

Multi-tenancy management: interference

Energy efficiency management

Uncertainty and variability of service requests

Deployment and Service Models

DEPLOYMENT MODELS (WHERE IS IT, WHO OWNS IT?)

Private Cloud

- Internal resources managed in a “Cloud-like” fashion
- Greater level of security and personalization
- Less scalability and higher costs

(owned internally)

Public Cloud

- Resources rented to anyone who will pay

(owned by someone else)

Hybrid Cloud

- A combination of the above
- Typically links two or more cloud infrastructures (public or private) via a standard technology

(owned by multiple orgs)

SERVICE MODELS (WHAT DOES IT PROVIDE?)

Software-as-a-Service (SaaS)

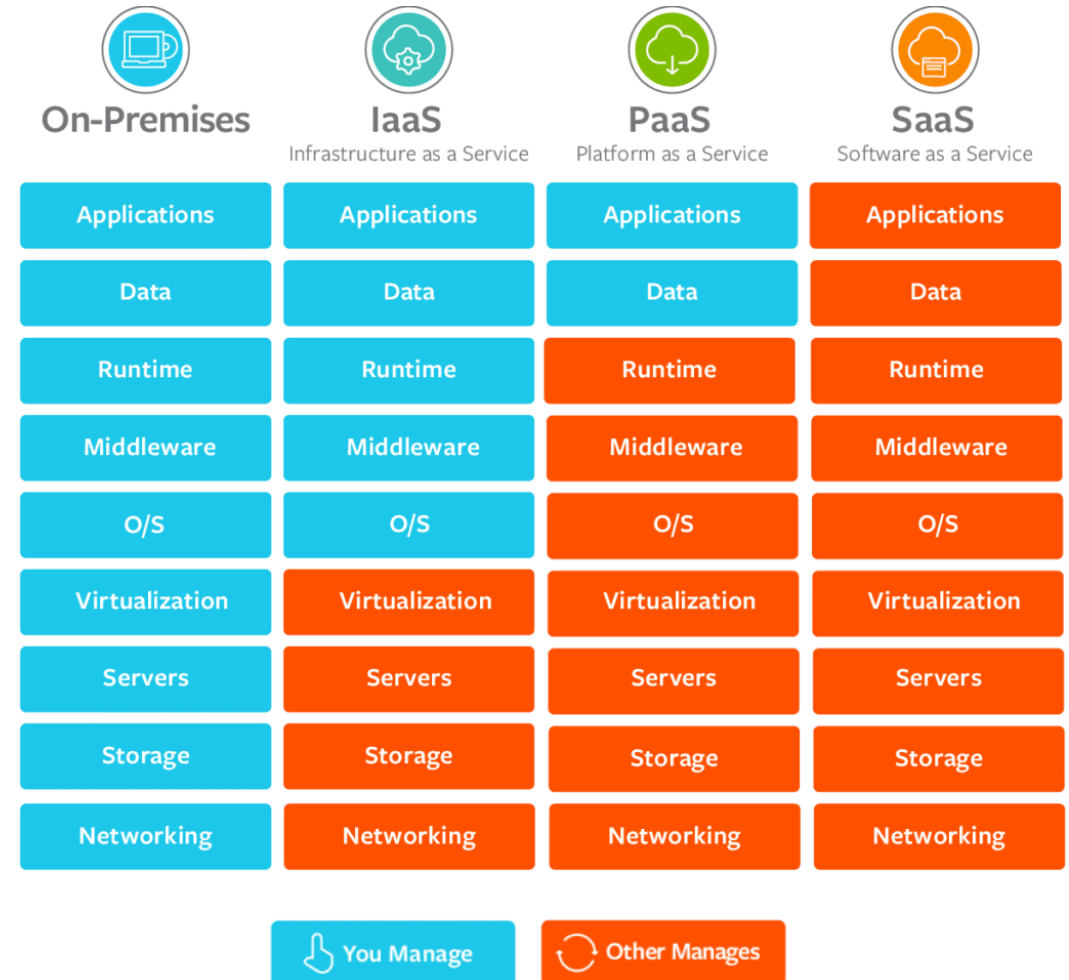
Ready-to-use applications
(O365, Spotify, Dropbox, etc)

Platform-as-a-Service (PaaS)

Ready-to-use platform
(Windows, Google Apps Engine, etc)

Infrastructure-as-a-Service (IaaS)

Full access to a hosted machine / VM
(Amazon EC2, Windows Azure, etc)



SERVICE MODELS: ON PREMISES (“ON PREM”)

Traditional approach - organisation owns and manages its own infrastructure
(buy hardware, install / configure / run applications, etc.)

**What is good
about this?**

Total control of the infrastructure

**What is bad
about this?**

Capacity management: how many machines?
Hardware management: buy, install, manage, upgrade
Security: updates, hardware decommission (hard drives)
Networking management: how to connect everything?
Storage management: redundancy, disaster recovery



On-Premises

Applications

Data

Runtime

Middleware

O/S

Virtualization

Servers

Storage

Networking

SERVICE MODELS: INFRASTRUCTURE-AS-A-SERVICE

Service as a commodity (e.g. electricity)

Cloud providers own and manage the infrastructure, customers rent resources

Provide hardware, VMs, storage, servers, monitoring, alerts, etc.

What is good
about this?

No need to buy hardware (as a customer)
Elastic and pay-per-use

What is bad
about this?

Lost control on the infrastructure
Possible higher cost in the long run (see: cloud economics)
Legal issues (how to manage data?)
Applications may require re-design for cloud deployment



IaaS

Infrastructure as a Service

Applications

Data

Runtime

Middleware

O/S

Virtualization

Servers

Storage

Networking

SERVICE MODELS: INFRASTRUCTURE-AS-A-SERVICE

Challenges

How to deploy an application?

Manage availability and scalability, load balancing

Manage the operating system

Selection and configuration of the hardware



IaaS

Infrastructure as a Service

Applications

Data

Runtime

Middleware

O/S

Virtualization

Servers

Storage

Networking

SERVICE MODELS: PLATFORM-AS-A-SERVICE

“Rent the operating system”

Customer uploads + controls applications, and can configure some of the environment

What is good about this?

Infrastructure is managed by the cloud provider

What is bad about this?

Limitation in the deployment environment

Challenges

The customer is in charge of managing load balancing and networking



PaaS

Platform as a Service

Applications

Data

Runtime

Middleware

O/S

Virtualization

Servers

Storage

Networking

SERVICE MODELS: SOFTWARE-AS-A-SERVICE

Use an existing application that is provided on a Cloud infrastructure

Same application is shared on different customer devices by a thin client

What is
good about
this?

Can use with no deployment or configuration decisions

What is bad
about this?

Limited configuration decisions and control



SaaS

Software as a Service

Applications

Data

Runtime

Middleware

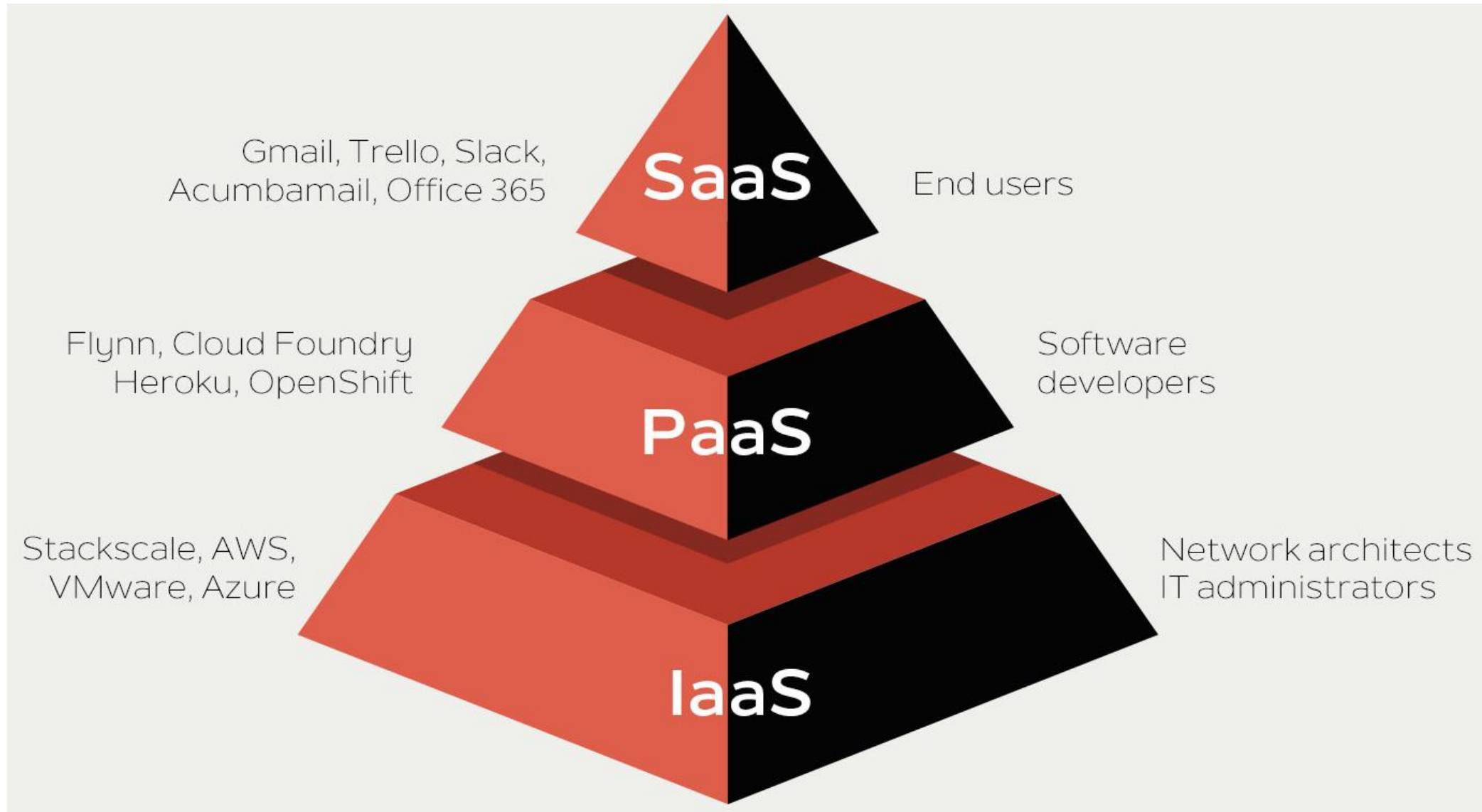
O/S

Virtualization

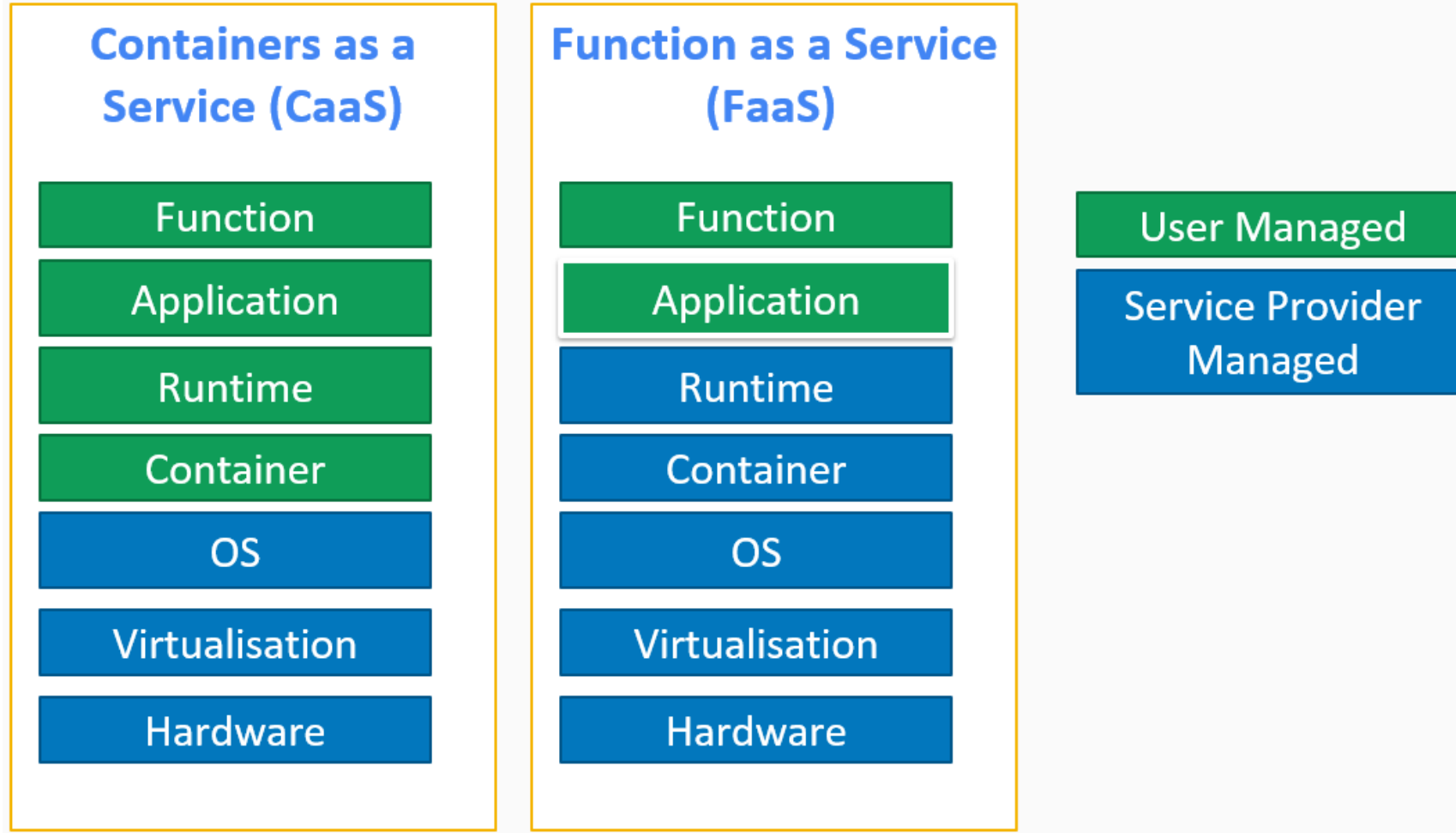
Servers

Storage

Networking



NEWER SERVICE MODELS



SERVICE MODELS: CONTAINER-AS-A-SERVICE [CAAS]

Containers pack applications and dependencies into an image

Same isolation as virtualisation, lighter than VMs

Requires a container orchestrator
Manages the infrastructure and runs containerised applications
Manages the interactions between applications

What is good
about this?

Declarative deployment (desired final state of infrastructure)
Avoids vendor lock-in (multi-cloud environment)
Multiple containers on a single machine

Containers as a Service (CaaS)

Function

Application

Runtime

Container

OS

Virtualisation

Hardware

SERVICE MODELS: FUNCTION-AS-A-SERVICE (FAAS)

Serverless – a new trend. Pay per request (no idle time)

Auto-scaling and availability provided out-of-the-box

Computation is implemented as functions and execution is event driven

Customers define functions

Users select functions and specify the events triggering them

Examples include AWS Lambda, Google Cloud Functions, etc.

Function as a Service (FaaS)

Function

Application

Runtime

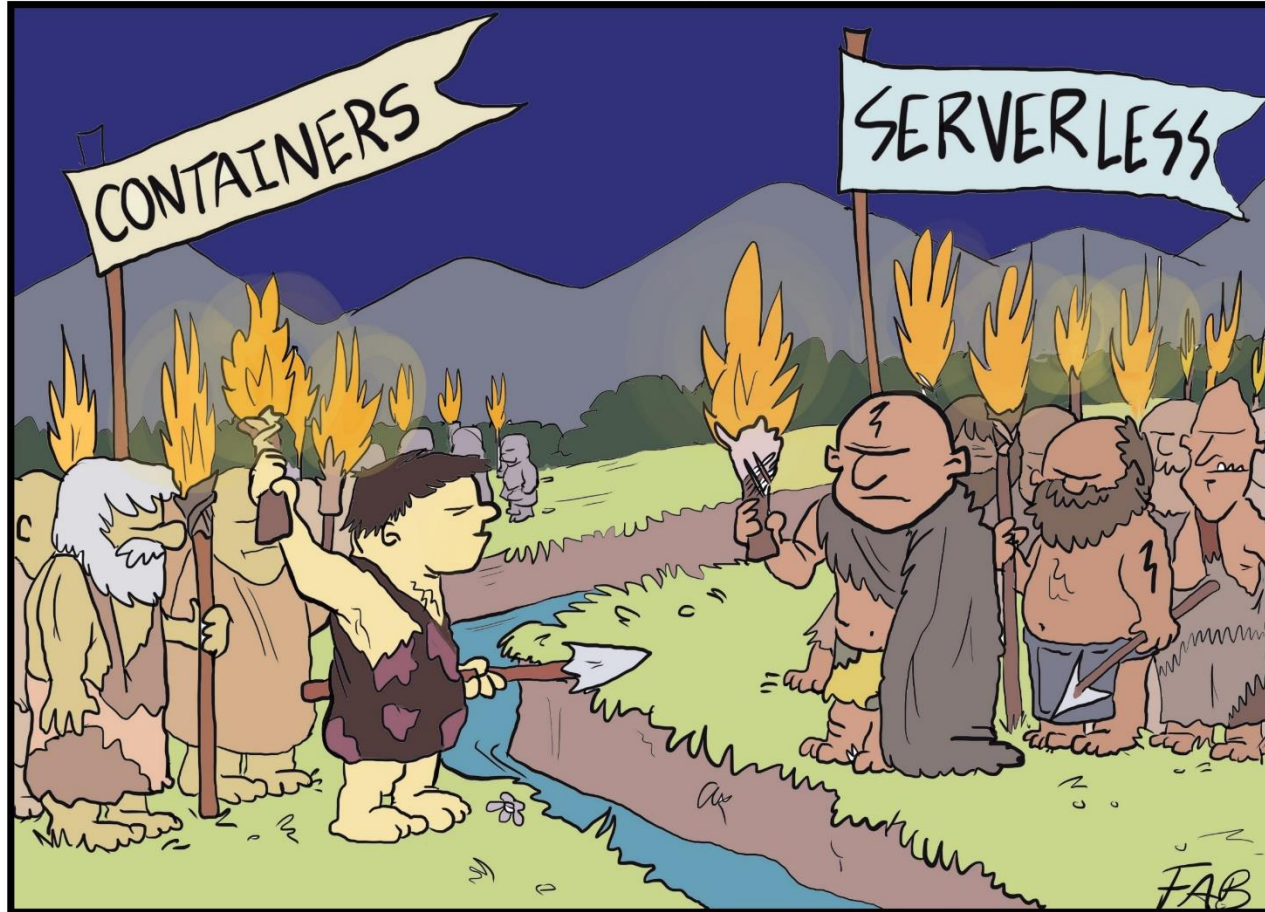
Container

OS

Virtualisation

Hardware

FaaS and Furious by Forrest Brazeal A CLOUD GURU



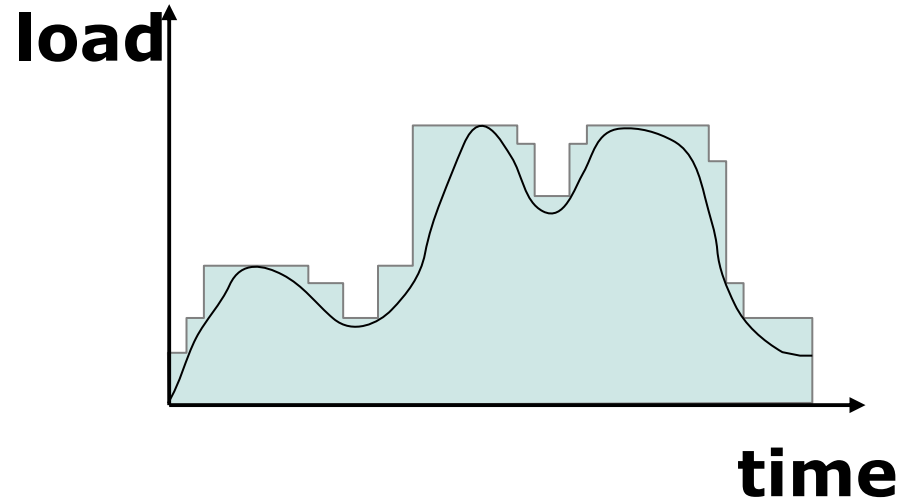
The two tribes regarded each other suspiciously
in the glow of their brightly blazing production environments.

Cloud Native Applications

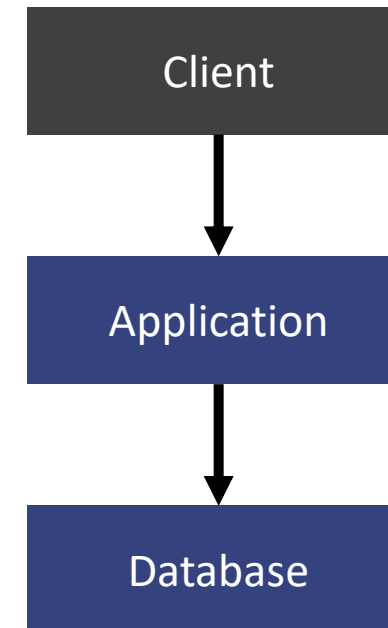
MAKING AN APPLICATION “CLOUD READY”

Almost any application can run in the Cloud, but they are **not** “cloud-ready”

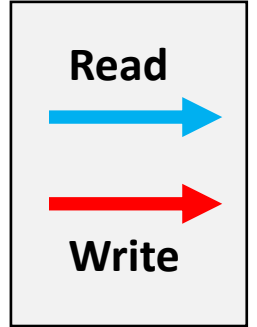
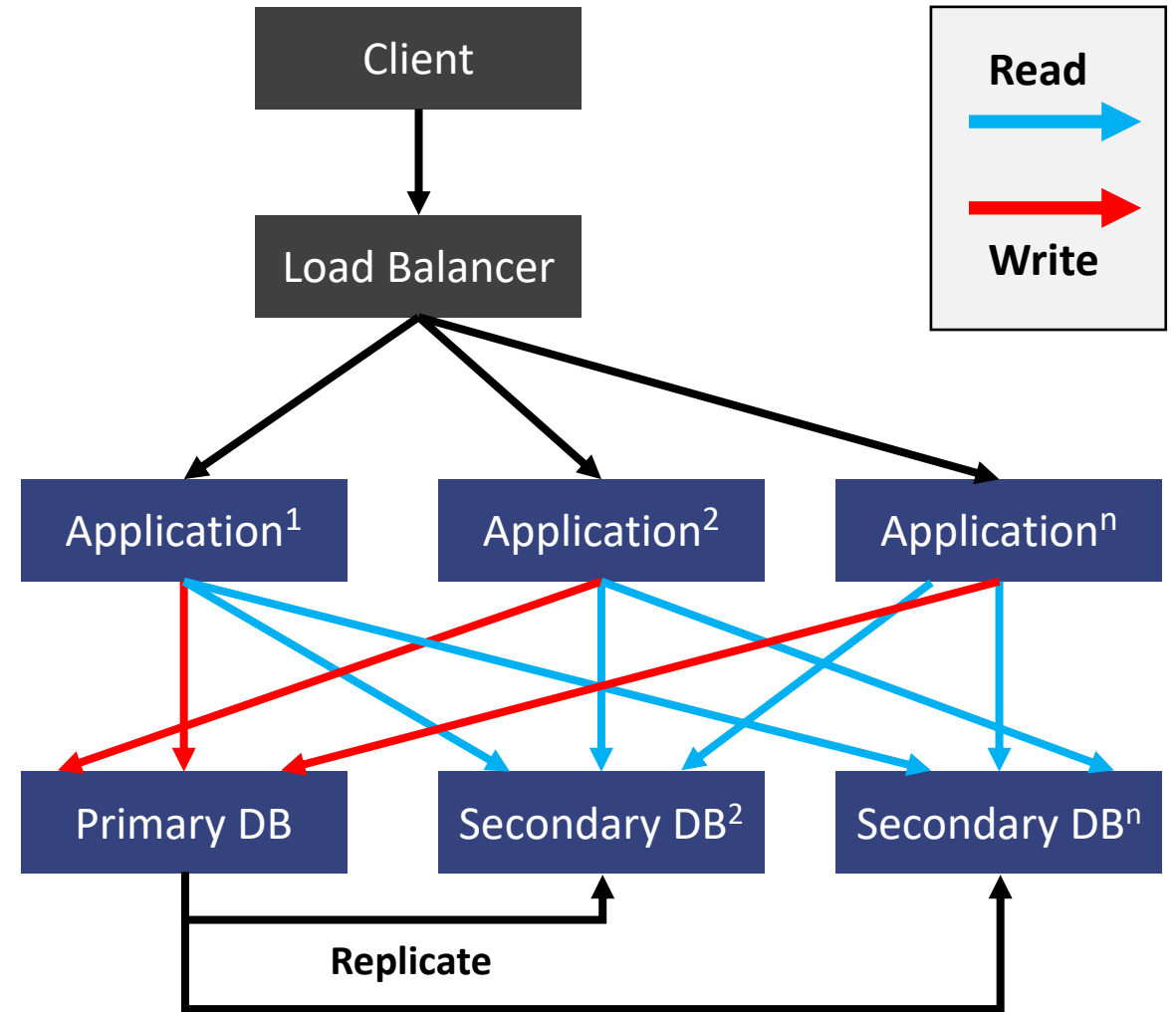
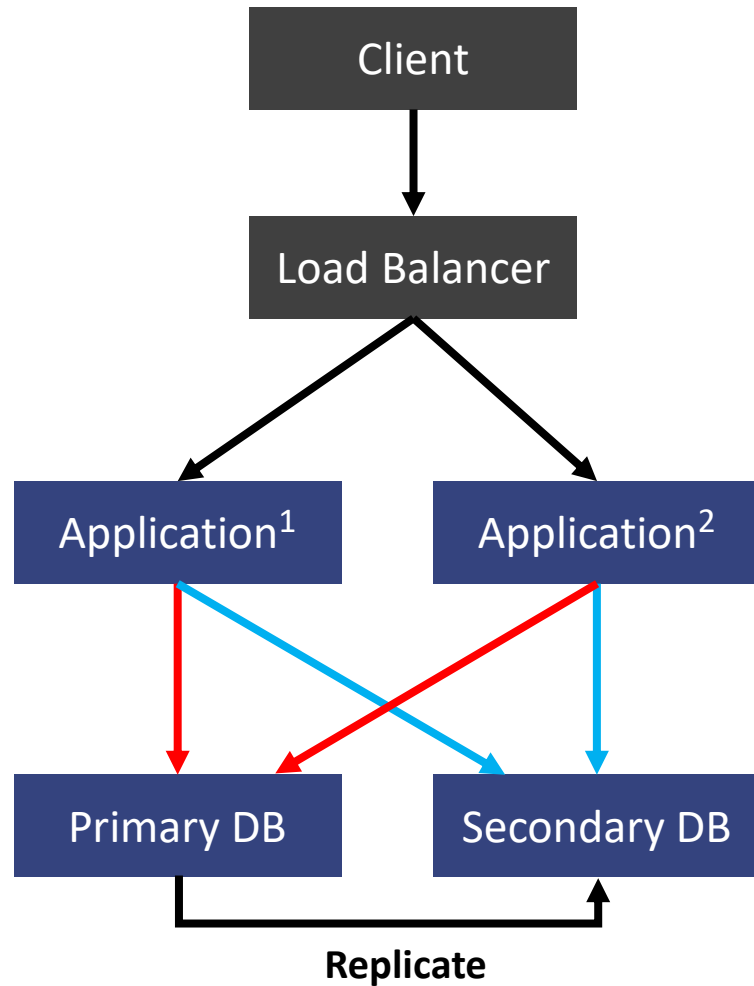
A key advantage to cloud is the ability to **auto-scale**: an “elastic” application can scale with load



Traditional application: 1 VM per tier



AUTO-SCALING CLOUD APPLICATION ARCHITECTURES



Physical Cloud Infrastructure

[Data Centers]

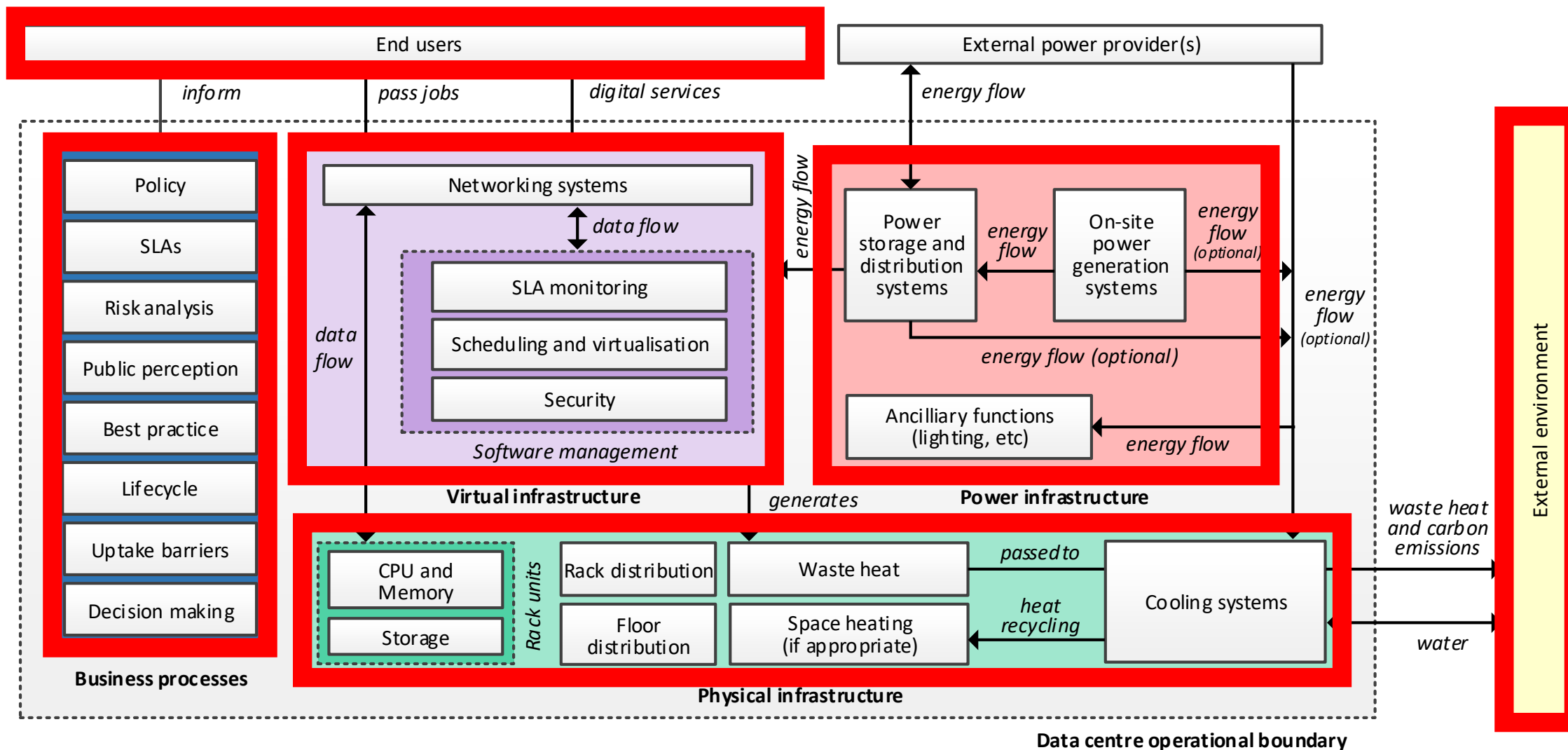
DATA CENTERS

The background image shows a vast, dimly lit data center. In the foreground and middle ground, there are several tall, dark server racks. Some of these racks have their doors open, revealing internal components that are illuminated with a warm, yellowish light. The floor is made of large, light-colored square tiles. The ceiling is high and features a complex network of pipes, conduits, and structural beams. Some of the ceiling lights are glowing with a cool, blue light. The overall atmosphere is one of a high-tech, industrial environment.

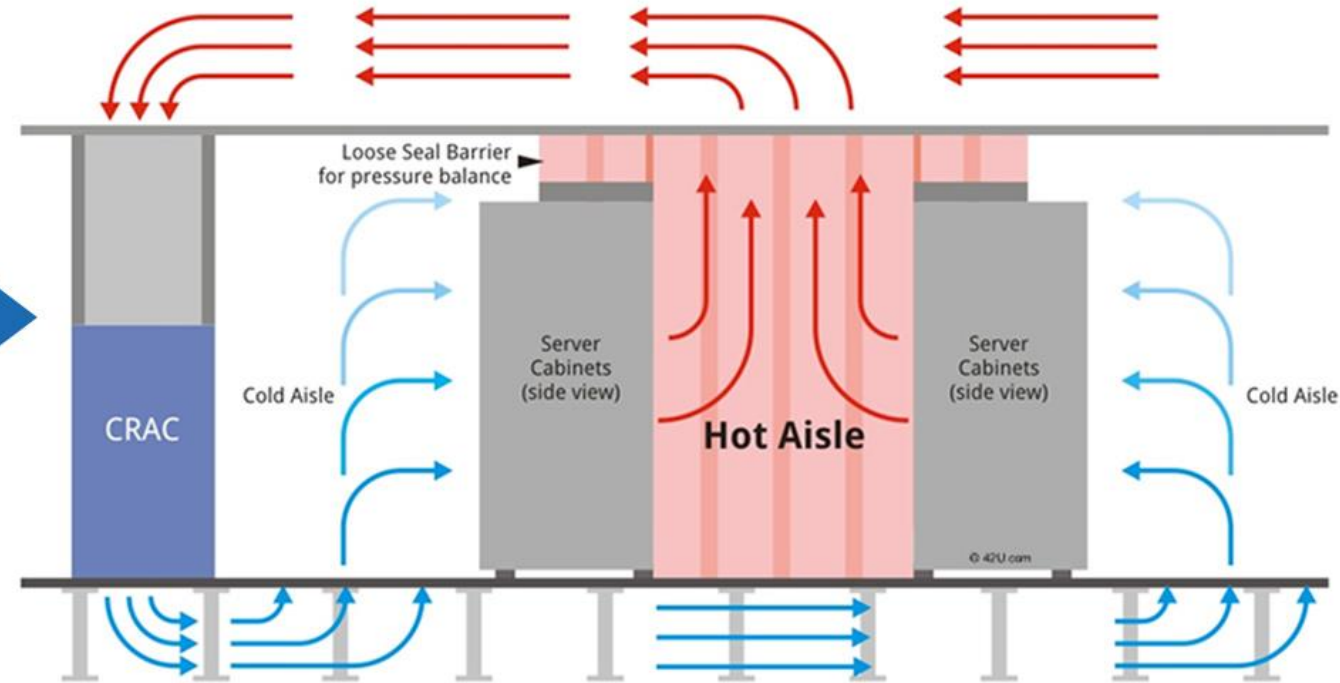
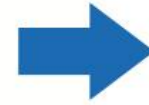
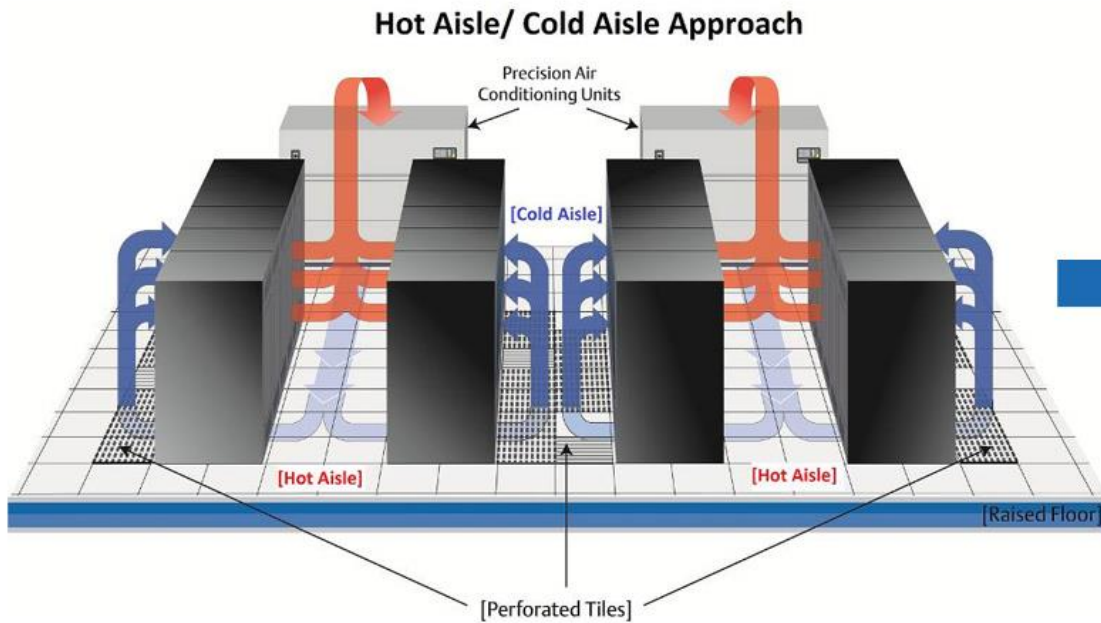
A **data center** is a facility composed of networked computers and storage that businesses or other organizations use to organize, process, store and disseminate large amounts of data

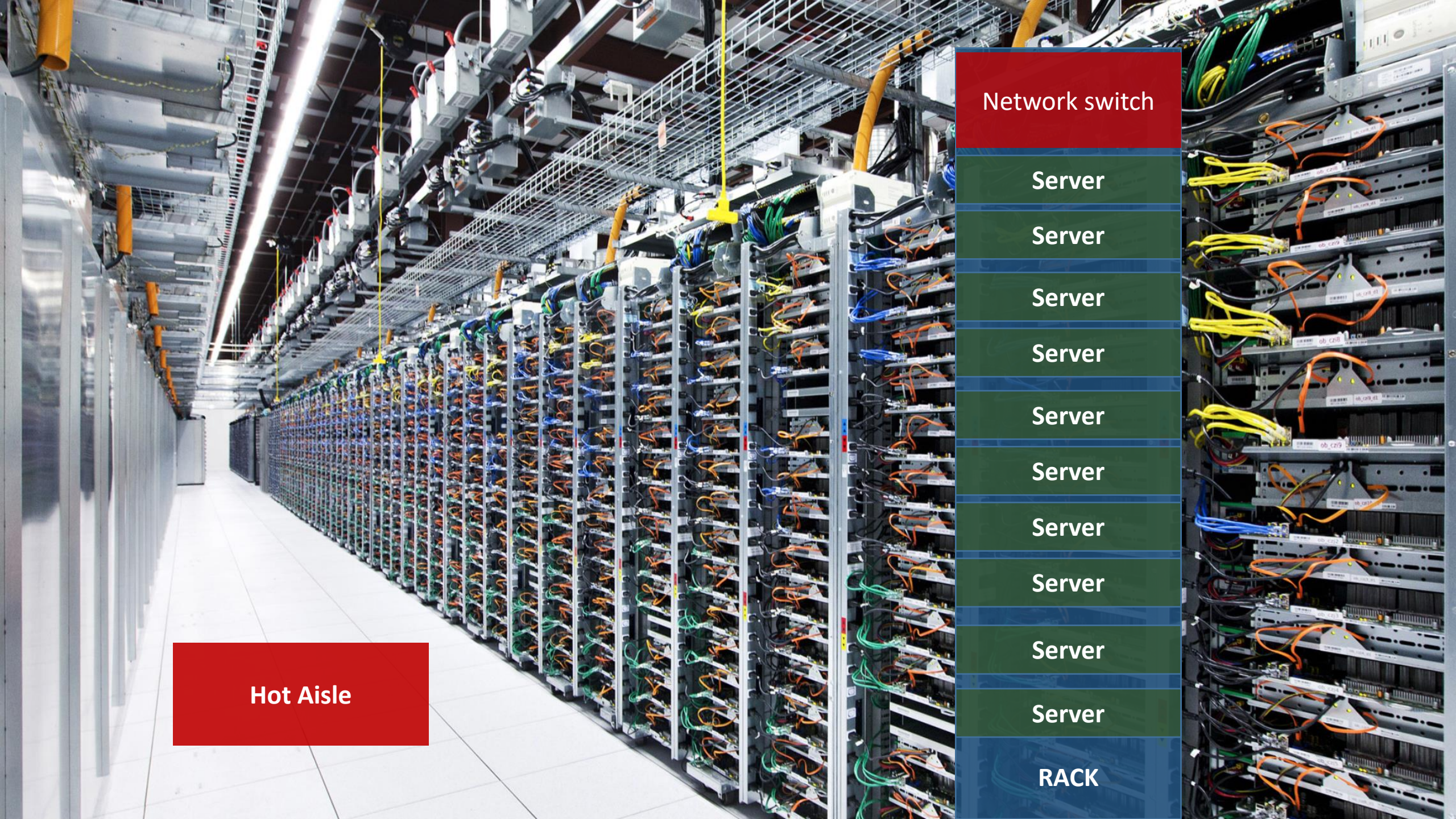
GOOGLE HAMINA DATA CENTRE, FINLAND

DATA CENTERS AS SYSTEMS OF SYSTEMS



HOT AISLE / COLD AISLE CONTAINMENT





Hot Aisle

Network switch

Server

Server

Server

Server

Server

Server

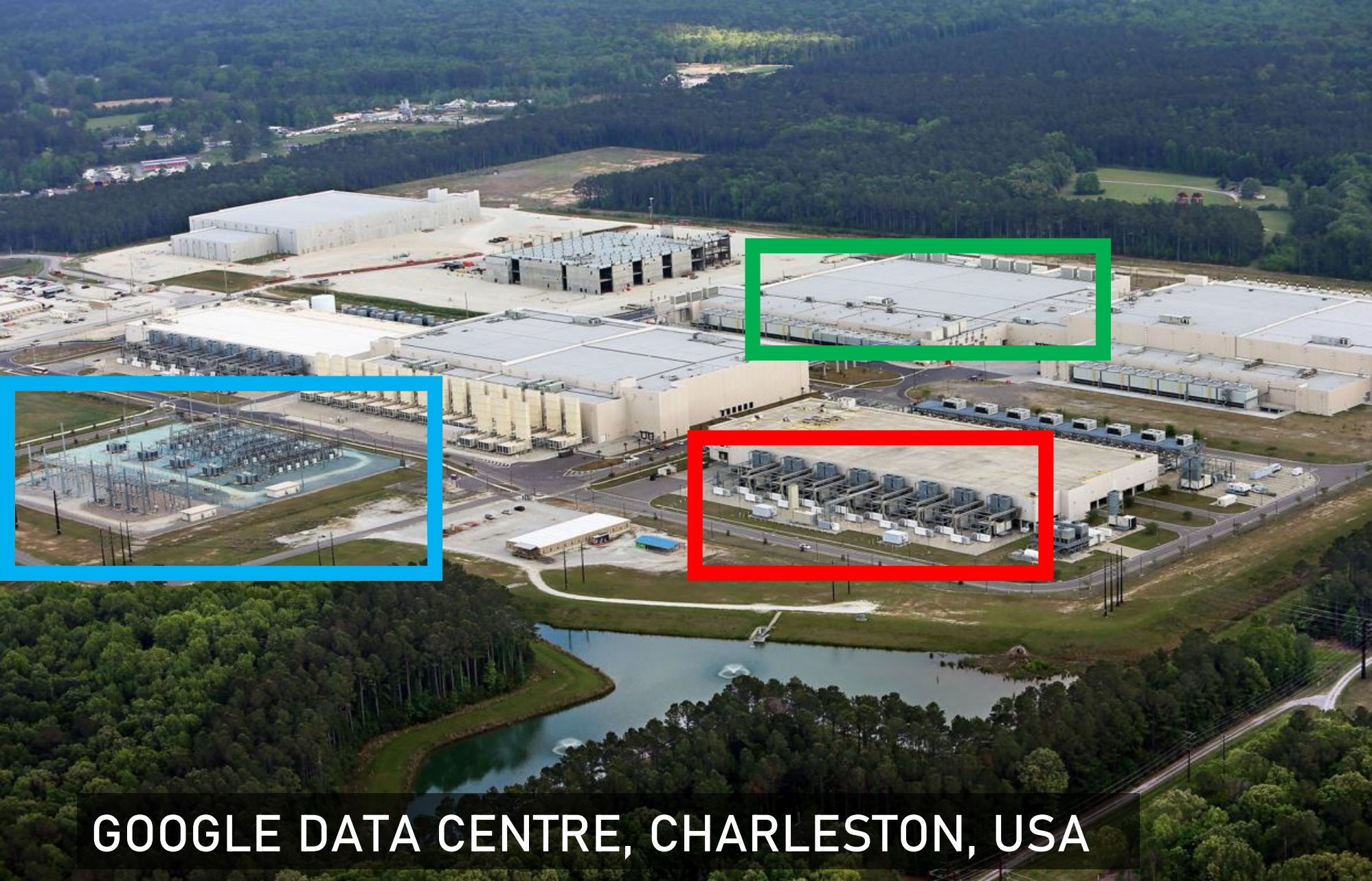
Server

Server

Server

Server

RACK



21 Billion SEK facility
(twenty one)

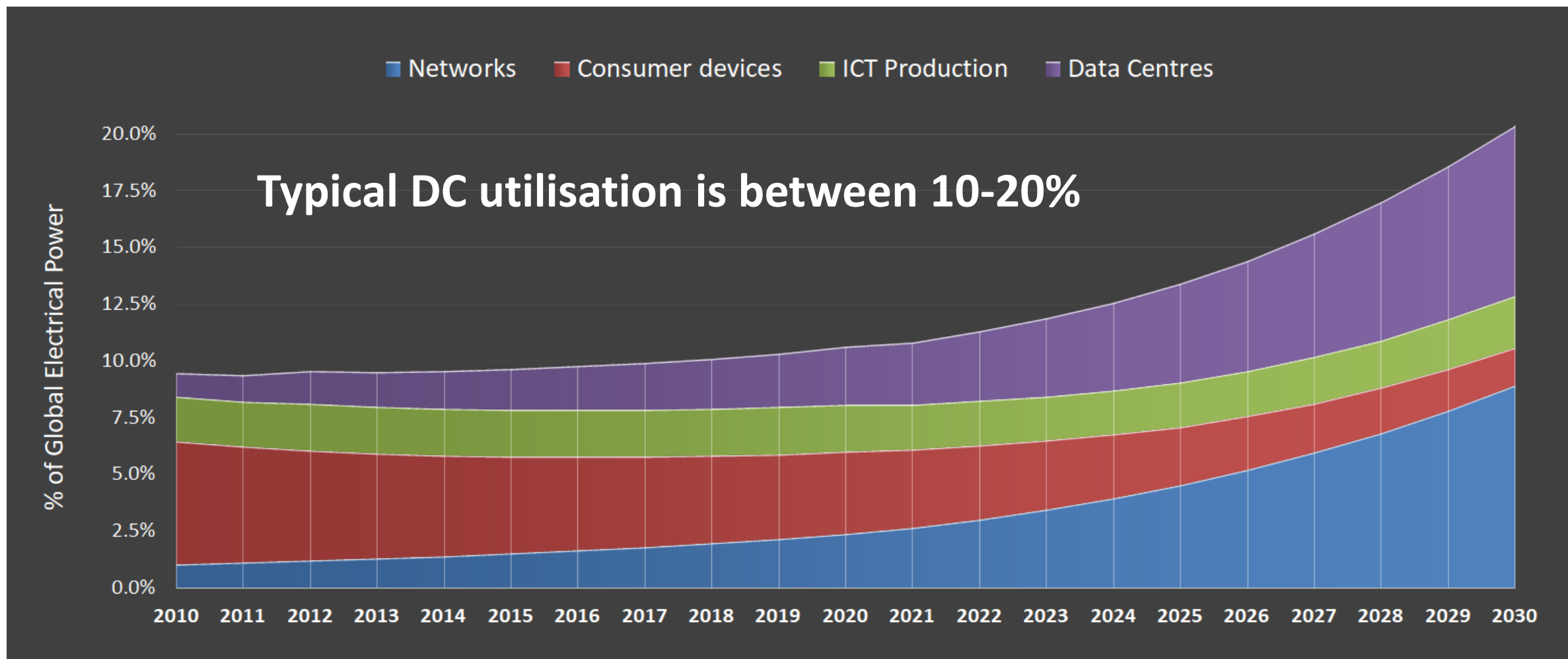
Server buildings

Cooling

Power facility

GOOGLE DATA CENTRE, CHARLESTON, USA

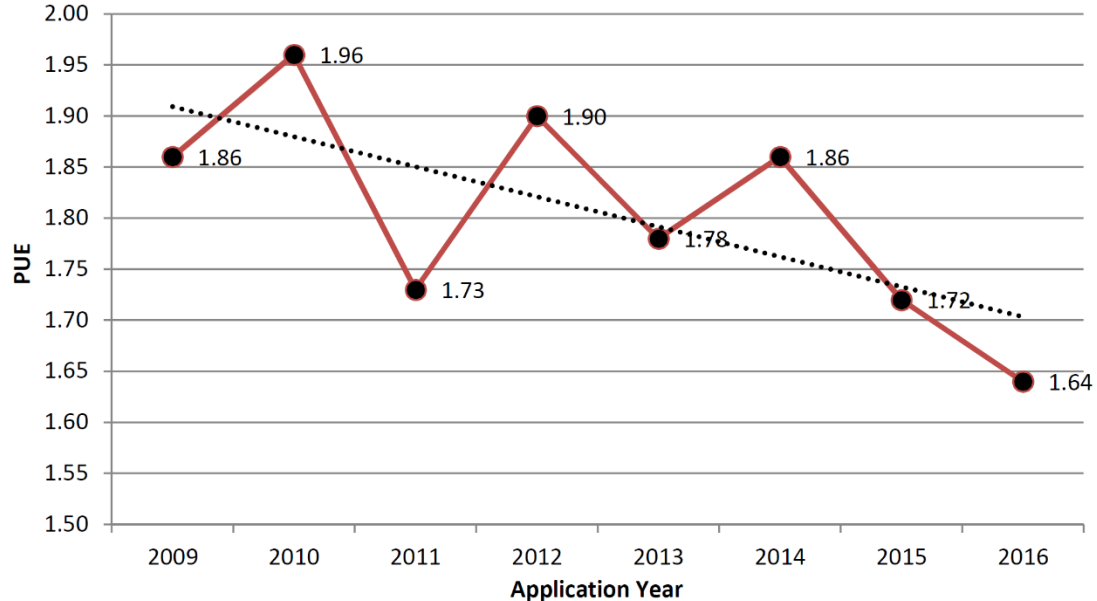
PREDICTED GLOBAL POWER CONSUMPTION



A. Anders, T. Edler, "On global electricity usage of communication technology: trends to 2030.", Challenges 6, no. 1 (2015): 117-157

POWER USAGE EFFECTIVENESS

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

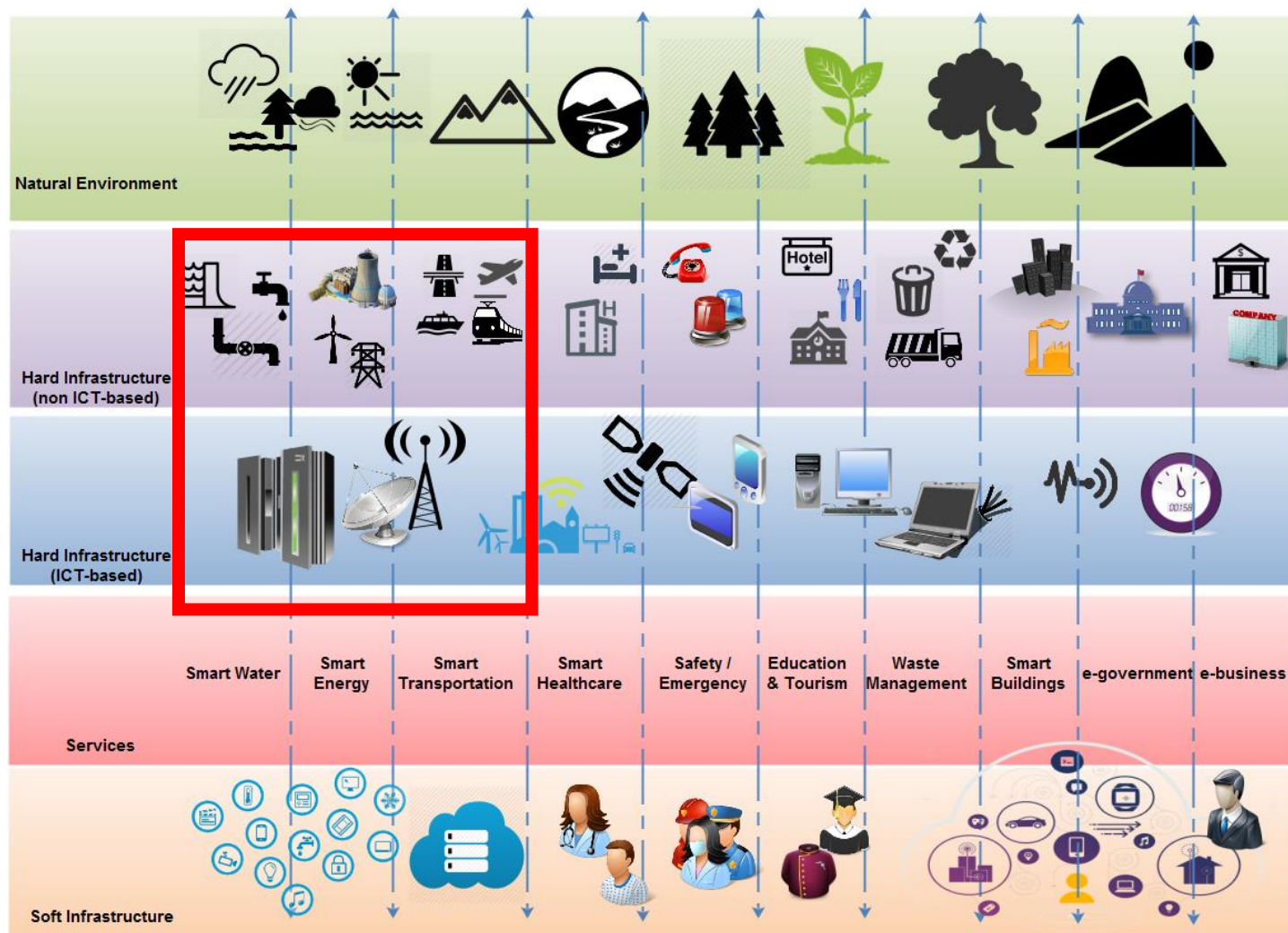


PUE is the data centre industry standard metric for efficiency

What are some of the problems with PUE?

As PUE approaches 1, why does IT power become increasingly important?

SMART INFRASTRUCTURE STACK: IMPLICATIONS



IMPROVED ENVIRONMENT

REDUCTION IN POWER
CONSUMPTION AND EMISSIONS

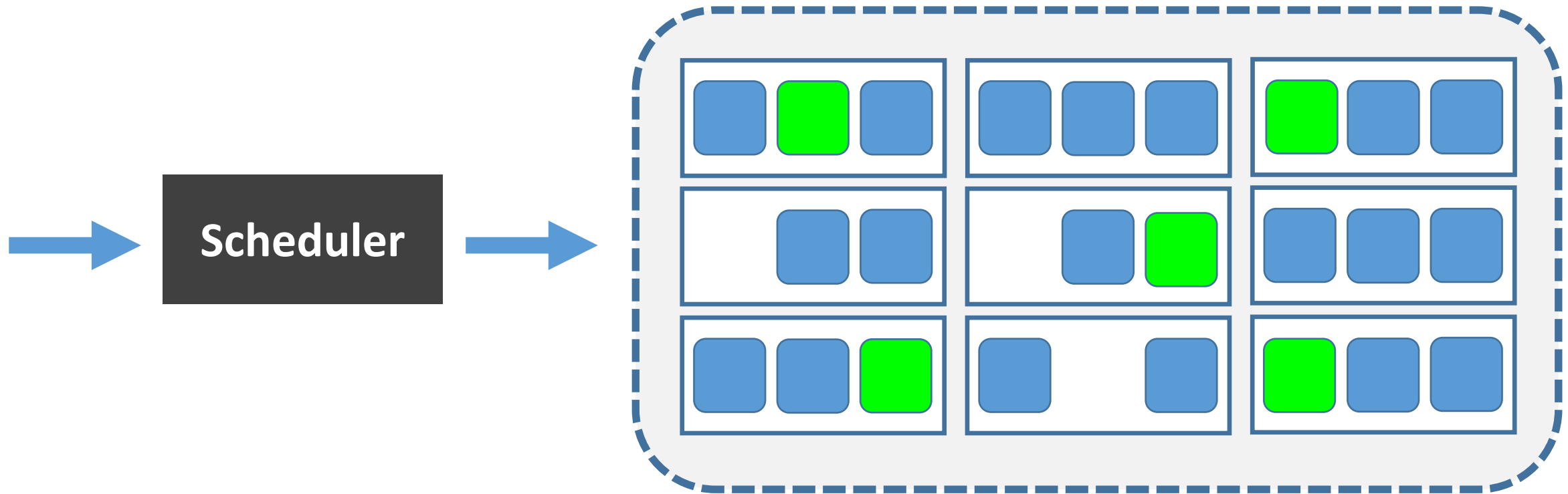
IMPROVEMENT IN DC EFFICIENCY

REDUCED LATENCY

IMPROVED QOS

SOFTWARE BASED EFFICIENCY IMPROVEMENTS

Schedule virtual workloads in a more effective manner



WHAT CAN WE DO WITH SCHEDULING?

Over-allocation

Allocate more work on the same nodes

Use less machines

Interference

Avoid contention between co-located workloads

Reduce power, improve performance

Optimise hardware

Allocate work until nodes are at “optimum” efficiency

Reduce power, improve performance

TYPES OF DATA CENTER

Co-locational

Require customers to supply their own hardware

Telecoms

Typically feature high connectivity requirements and run specialised software services

Dedicated hosting

Provide server capacity to single customers with no sharing of machines (bare metal hosting)

Managed hosting

Provide servers and storage systems for customers, often as PaaS, IaaS or SaaS

Shared hosting

Cloud data centres are an example. Provide virtualised multi-tenant resources.

Edge

Typically smaller than traditional data centres, and located closer to where data is generated

Hybrid

A data centre facility with more than one of the above service models

DATA CENTER TIERS

Tiers classify the structure of a data centre

Tier I

- Single power supply system and single cooling system
- No backup policies or redundant components

Tier II

- Single power supply system and single cooling system
- Some components are redundant and backup policies
- Minimum uptime must be 99.749% annually

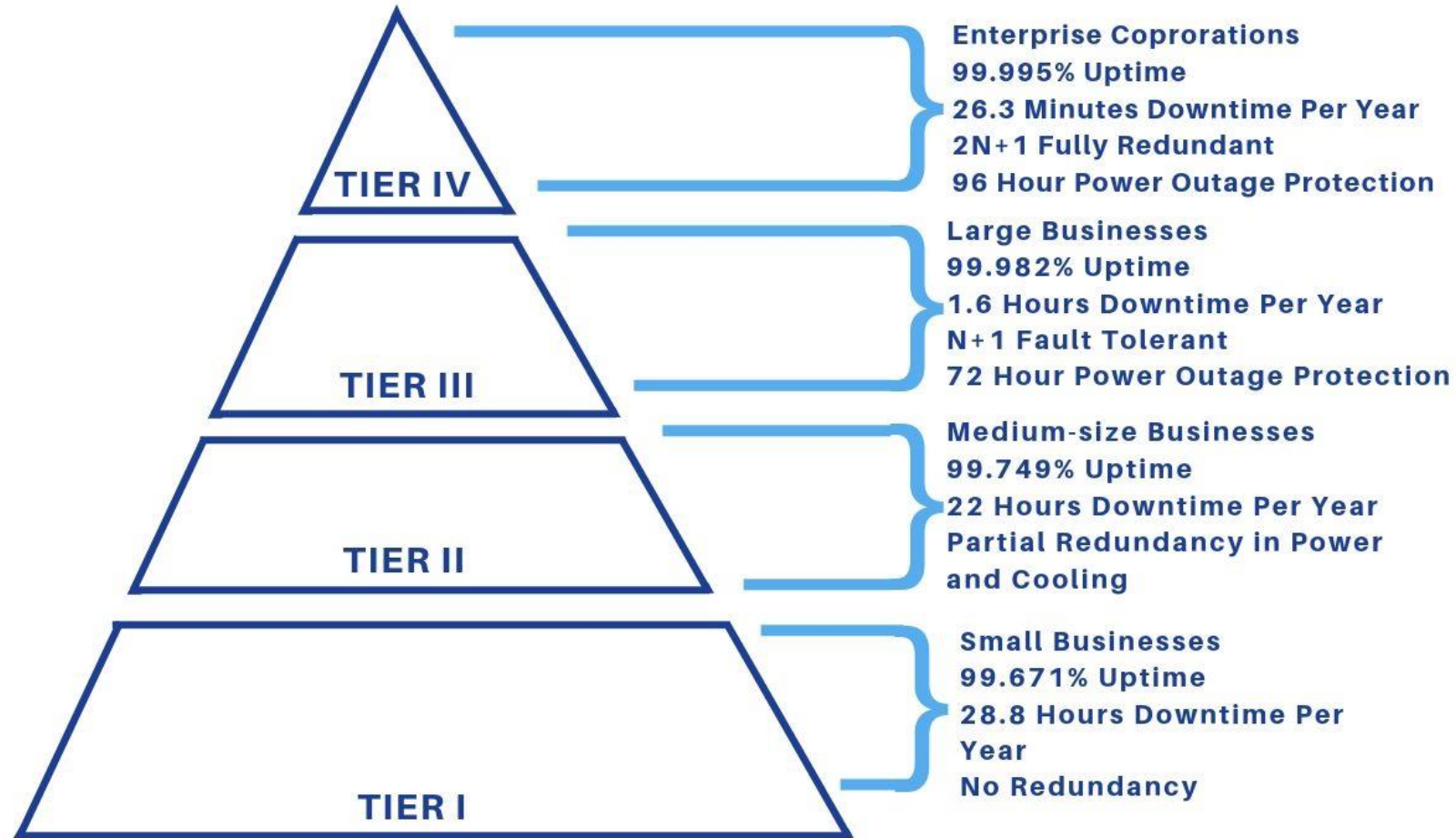
Tier III

- Multiple power supply systems + multiple cooling systems
- All components are redundant
- Most upgrades do not take centre offline
- Minimum uptime must be 99.982% annually

Tier IV

- Designed and created to be totally fault-tolerant
- All components redundant
- Has more power and cooling systems
- Guaranteed uptime is 99.995%

DATA CENTER TIERS





FACEBOOK DATA CENTRE, LULEÅ, SWEDEN

IN SUMMARY

**NIST definition and
characteristics**

**Virtual machines and
containers**

**Industry uptake of Cloud
and economics**

**Deployment and service
models**

What are data centres?

**Data centre types and
tiers**