

“I think there is a world market for maybe five computers”

Thomas Watson, President of IBM, 1943

Lecture A.3: Practical Aspects of Cloud Computing

CS205: Computing Foundations for Computational Science
Dr. David Sondak
Spring Term 2020



HARVARD
School of Engineering
and Applied Sciences



**INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE**
AT HARVARD UNIVERSITY

Lectures developed by Dr. Ignacio M. Illorente

Before We Start

Where We Are

Computing Foundations for Computational and Data Science

How to use modern computing platforms in solving scientific problems

Intro: Large-Scale Computational and Data Science

A. Parallel Processing Fundamentals

A.1. Parallel Processing Architectures

A.2. Large-scale Processing on the Cloud

A.3. Practical Aspects of Cloud Computing

A.4. Application Parallelism

A.5. Designing Parallel Programs

B. Parallel Computing

C. Parallel Data Processing

Wrap-Up: Advanced Topics

CS205: Contents

APPLICATION SOFTWARE

APPLICATION
PARALLELISM

PARALLEL PROGRAM
DESIGN



Optimization

PROGRAMMING MODEL

OpenACC

Spark

OpenMP

Map-Reduce

MPI

B. BIG COMPUTE

PLATFORM

C. BIG DATA



CLOUD COMPUTING



Open
Nebula



FASRC



ODYSSEY
HARVARD FAS
RESEARCH COMPUTING

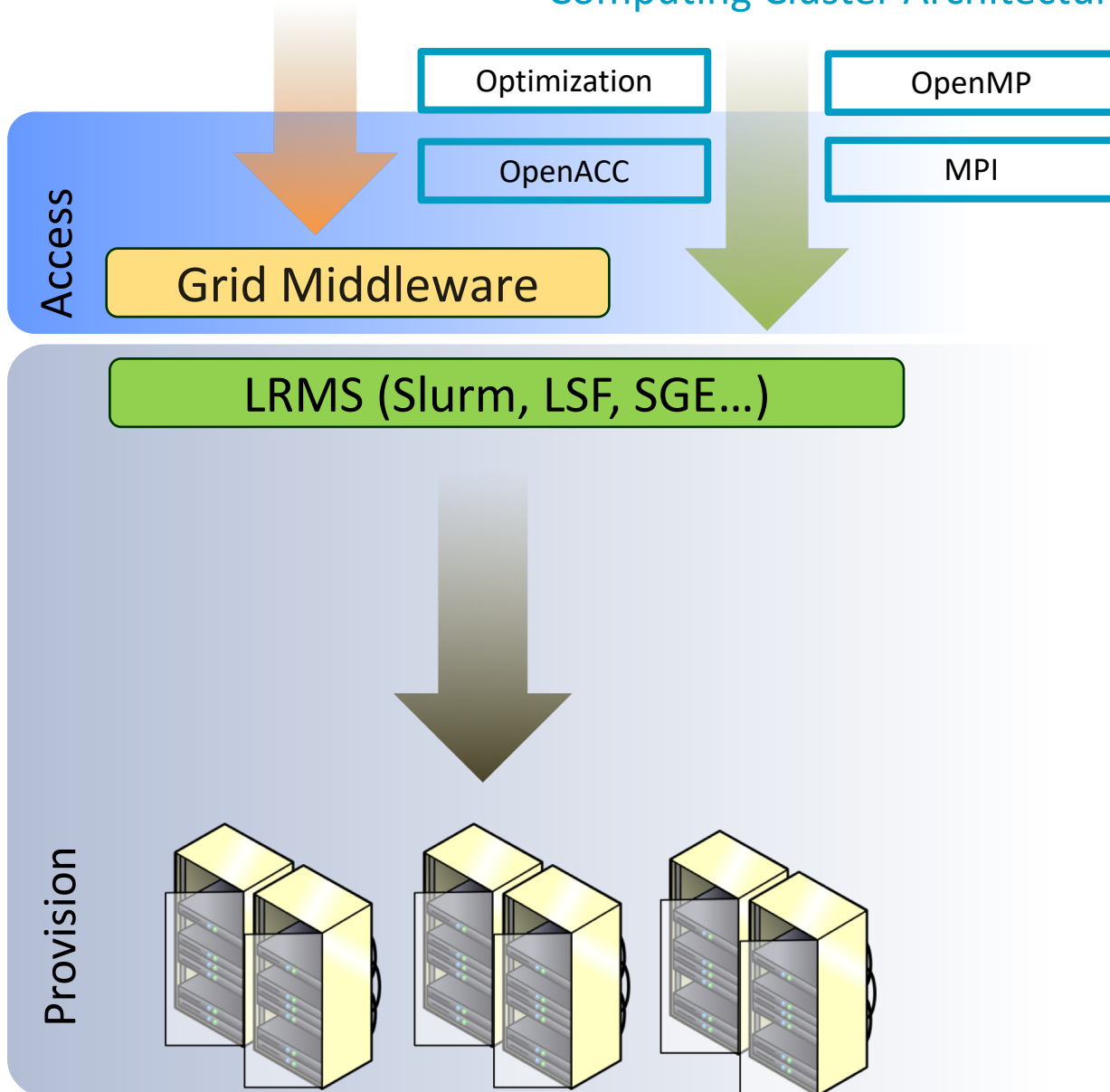


FASRC

PARALLEL ARCHITECTURES

Context

Computing Cluster Architecture



HPC

Programming Model

+

Platform

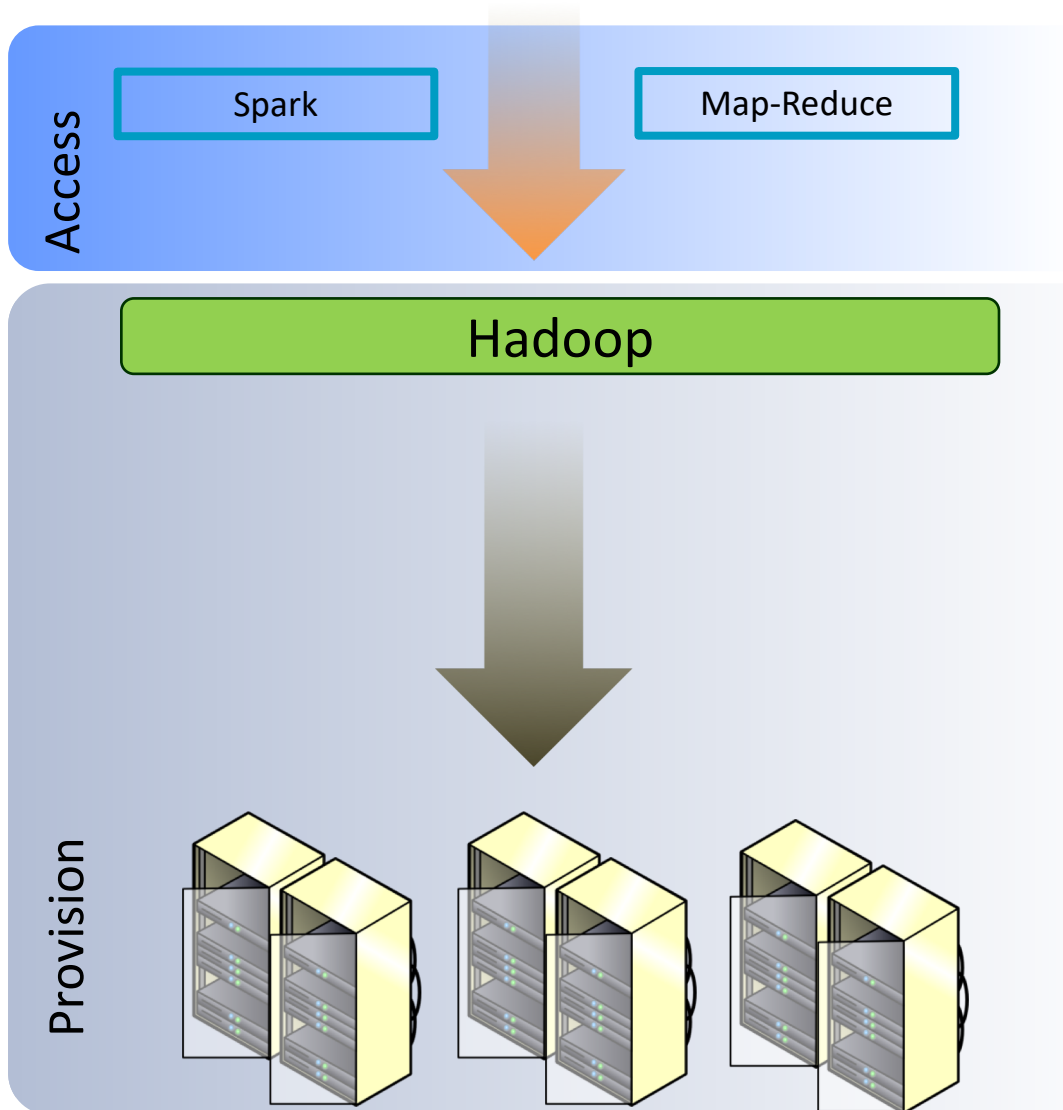
+

Architecture

Context

Computing Cluster Architecture

BIG DATA



Programming Model

+

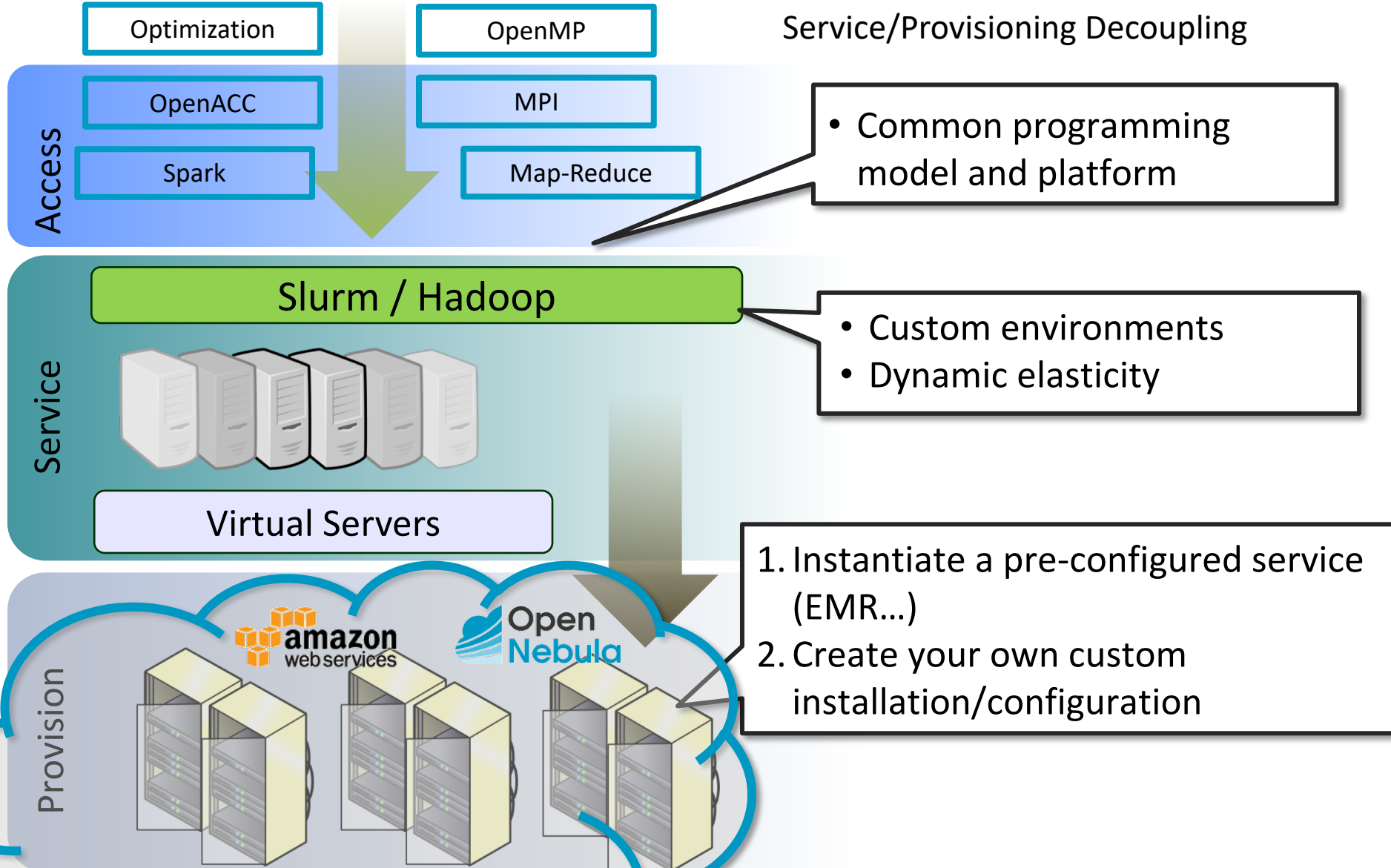
Platform

+

Architecture

Context

Cloud as Infrastructure Tool



Roadmap

Practical Aspects of Cloud Computing

Numerical Reproducibility and Replicability

Economic Aspects

The State of Public Cloud

The Need for Private Clouds

The Anatomy of the Cloud

Numerical Reproducibility and Replicability

The Four Rs

Who has ever written a paper/report including numerical experiments (computational physics, bioinformatics, applied mathematics, statistics....)?

Would I be able to reuse your software/data?

Would I be able to rewrite your software/data?

Would I be able to reproduce your experiment (software/data + execution environment) in a different computing infrastructure and get the same numerical result?

Would I be able to replicate (repeat) your experiment (software/data + execution environment + computing infrastructure) and get the same computing result?

Numerical Reproducibility and Replicability

The Four Rs

Reusability

- Reusability refers to the possibility to reuse the software or parts of it for different purposes, in different environments, and by researchers other than the original authors.

Rewriteability

- Rewriteability refers to the possibility to modify and extend the software or parts of it.

Reproducibility

- Reproducibility of a computational experiment means that it can be repeated by a different researcher in a different computing infrastructure but with the same execution environment and to come to the same numerical results.

Replicability

- The attribute Replicability describes the ability to repeat a computational experiment on the same computing infrastructure and to come to the same numerical results and computing performance.

Numerical Reproducibility and Replicability

The Four Rs

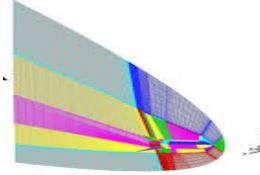
AERODYNAMICS



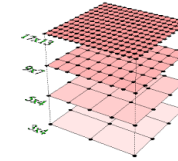
NAVIER-STOKES

$$\frac{\partial u}{\partial t} + \frac{1}{r^2} \frac{\partial(r^2 u)}{\partial r} + \frac{\partial(vu)}{\partial z} = -\frac{\partial p}{\partial r} + \frac{1}{Re} \frac{\partial}{\partial z} \left(\frac{\partial u}{\partial z} - \frac{\partial v}{\partial r} \right) + \frac{1}{Fr^2} g_r,$$
$$\frac{\partial v}{\partial t} + \frac{1}{r^2} \frac{\partial(r^2 uv)}{\partial r} + \frac{\partial(vv)}{\partial z} = -\frac{\partial p}{\partial z} + \frac{1}{Re r^2} \frac{\partial}{\partial r} \left(r^2 \left(\frac{\partial u}{\partial z} - \frac{\partial v}{\partial r} \right) \right) + \frac{1}{Fr^2} g_z,$$
$$\frac{1}{r^2} \frac{\partial(r^2 u)}{\partial r} + \frac{\partial v}{\partial z} = 0,$$

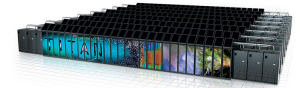
FINITE DIFFERENCE



MULTIGRID



PARALLEL



PHYSICS

ACCURACY

COMPLEXITY

SPEED-UP

EXECUTION ENVIRONMENT

- Algorithm, application version and dependencies
- VIRTUAL MACHINES
- SOFTWARE CONTAINERS

SYSTEM CAPACITY

- Execution time, performance...
- CLOUD PROVIDERS

Economic Aspects

Private vs. Public

ON-PREMISES



CLOUD (VIRTUALIZATION)

VARIABLE COSTS

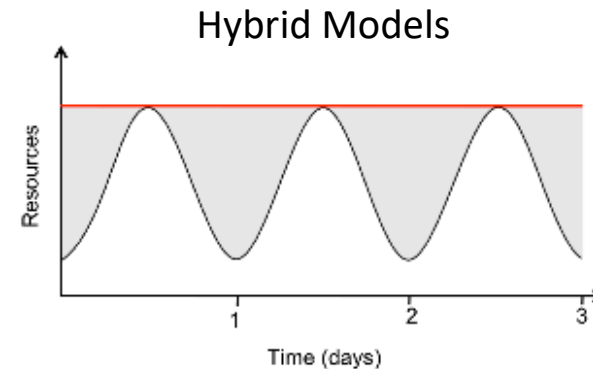
FIXED COSTS
SECURITY ASPECTS
PERFORMANCE
(DATA/ACCESS LATENCY)
INTEGRATE WITH LOCAL
PROCESSES/SERVICES
FLEXIBILITY/ELASTICITY

PERFORMANCE (HPC)

Economic Aspects

Private vs. Public

5 Servers: 15,000 \$ – 2x Quad-Core Servers
Rack and switch: 1,500 \$
Total Hardware: 16,500 \$
Cost Equipment Annual: 5,500 \$ (depreciation 3 years)
Power: 3,500 \$ (0.4kW per server)
Admin: 2,000 \$ (420 server)
Cooling/install: 5,500 \$
Cost Maintenance Annual: 11,000 \$
Total Annual: 16,500 \$
Hours Year: 8,760
Cost per Hour and Server: 0.37 \$
Cost Instance Extra Large: 0.68 \$
Instance Usage: 54%



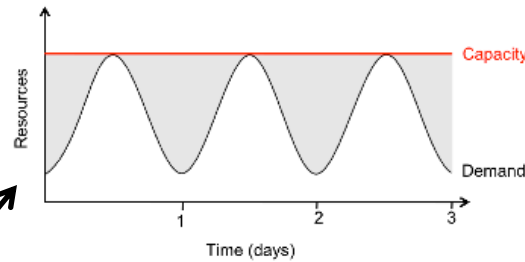
Economic Aspects

Sources of Variability

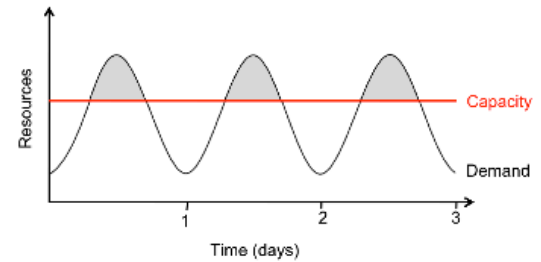
Variability of the Demand

- Data centers work on average at 5-20% of capacity
- Systems are sized to meet peak of demand
- Example with daily patterns

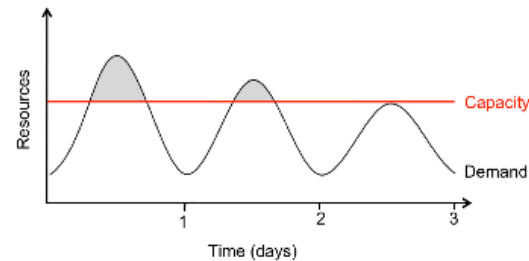
Max: 500 servers
Min: 100 servers
Average: 300 servers
Utilization: 7,200 s-h
Provision: 12,000 s-h
Ratio: 1.7



(a) Provisioning for peak load



(b) Underprovisioning 1



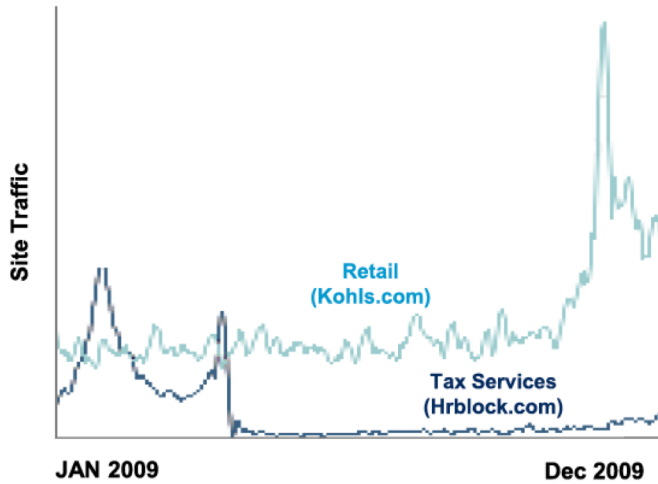
(c) Underprovisioning 2

Source: *Above the Clouds: A Berkeley View of Cloud Computing*, Berkeley

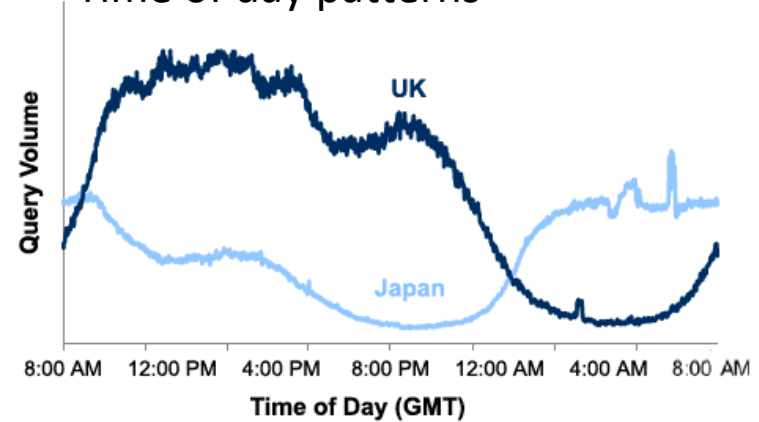
Economic Aspects

Sources of Variability

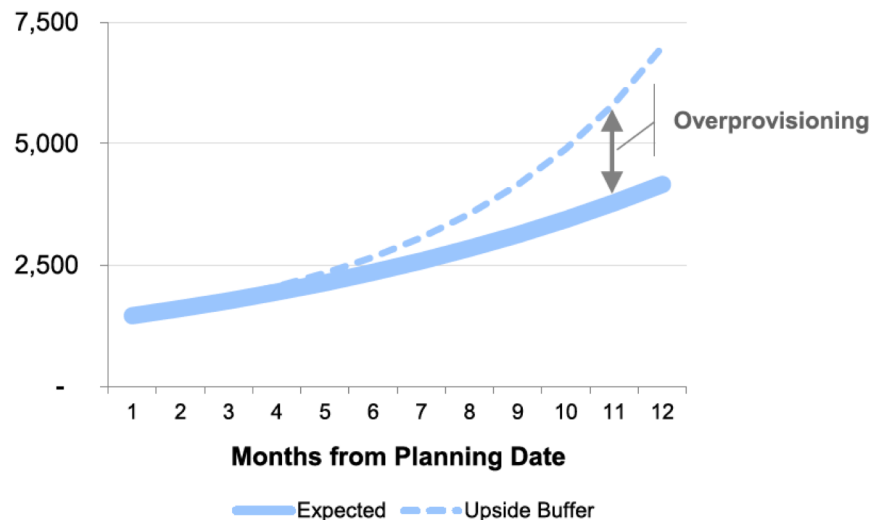
- Industry-specific variability



- Time of day patterns



- Uncertain growth patterns



Source: *The Economics of the Cloud*, Microsoft

Economic Aspects

Hidden Costs and Overheads

Consider All Hidden Costs and Overheads

- Cost and overhead to upload data

Data Set: 500 GB

Local Infrastructure: 10 servers, each needs 1 hour to process 1 GB Local Time: 50 hours

Amazon EC2: If we now go to AWS, how much time is needed?
(assume virtual servers with same capacity and not upload overhead)



Source: Above the Clouds: A Berkeley View of Cloud Computing, Berkeley

Execution time on AWS?

1 hour

25
hours

50
hours

Economic Aspects

Hidden Costs and Overheads

Consider All Hidden Costs and Overheads

- Cost and overhead to upload data

Considerations about Elasticity

- 100 servers 1 hours = 1 server 100 hours
- Massive parallelism is usually cheaper

Upload Data to S3: 20 Mbits/sec => 55 hours
Best Remote Time: 56 hours
Additional Cost for Data Upload?



Source: Above the Clouds: A Berkeley View of Cloud Computing, Berkeley

Additional cost for data upload?



The State of Public Cloud

August 2006 – Fourteen Years Ago



Products

Solutions

Pricing

More ▾

English ▾

My Account ▾

Sign In to the Console

ABOUT AWS

About AWS >

Global Infrastructure >

What's New >

AWS in the News >

Events & Webinars >

RELATED LINKS

What is Cloud Computing?

AWS Free Usage Tier

AWS Blog

AWS Careers

AWS Training

Manage Your Resources

Sign In to the Console

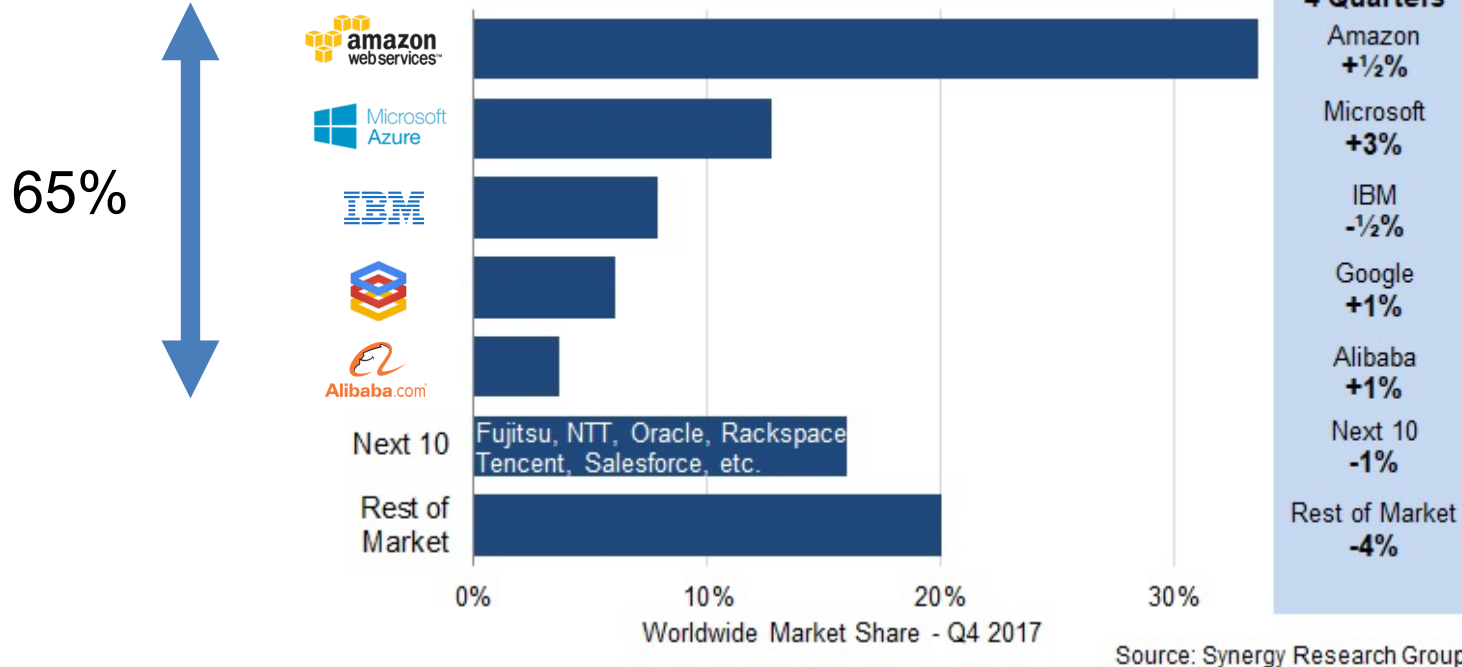
Announcing Amazon Elastic Compute Cloud (Amazon EC2) - beta

Posted On: Aug 24, 2006

Amazon Elastic Compute Cloud ([Amazon EC2](#)) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers. Just as Amazon Simple Storage Service (Amazon S3) enables storage in the cloud, Amazon EC2 enables "compute" in the cloud. Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use.

The State of Public Cloud

Public Cloud Market Becoming an Oligopoly



Growing market
Cloud services grew by 50% annually in Q4

Highly concentrated market
The 5 big players aggregate the 65% and the next 10 players the 15%


Increasing concentration
Big players gaining market share


The State of Public Cloud

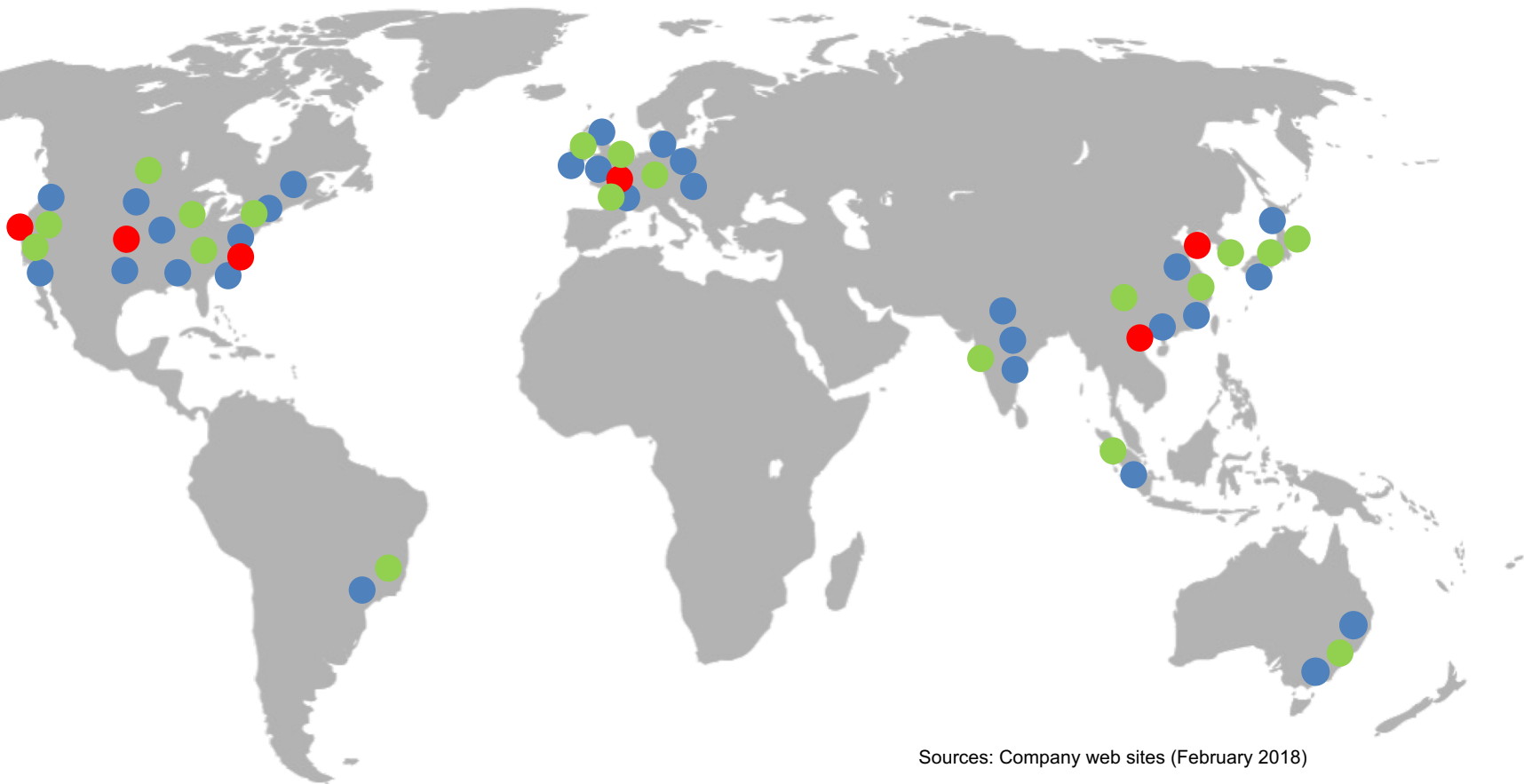
A Global Infrastructure

Regions 2016=>2017

 14=>19
22 in 2020

 30=> 36
56 in 2020





 6=>15
21 in 2020



Sources: Company web sites (February 2018)

The State of Public Cloud

Focus on Deploying Massive, Centralized Datacenters

	Zones (2016=>2017)	Servers (Millions)	New Zones in 2018
	38=>50	2-4	12
	30=>36	1-3	8
	31=>44	1-3	12
	39	1-3	8

↑
25%
↓

Each zone can have 1 or more datacenters
Most datacenters house between 50K and 80K servers

The 4 big players house approx. 10 million servers

This number is estimated to grow 25% annually (probably much higher if we consider their 50% annual growth in revenues)

Sources:

- Company web sites
- <http://datacenterfrontier.com/inside-amazon-cloud-computing-infrastructure>
- <http://www.datacenterknowledge.com/archives/2013/07/15/ballmer-microsoft-has-1-million-servers/>

The State of Public Cloud

Centralization...After All, Was Thomas Watson Right?

"There is only a market need in the world for five clouds"
Thomas Watson, IBM, 1854-1956.

The Need for Private Clouds

Centralized Public Cloud Does Not Fit All

Not so fast?

Main barriers to adoption

1. Cost
2. Performance
3. Security

The Need for Private Clouds

You Can Do It Cheaper

Cost

1. You may have sufficient scale to be able to do it cheaper
2. Public cloud is becoming more expensive
3. Cost benefit strongly depends on variability

The Need for Private Clouds

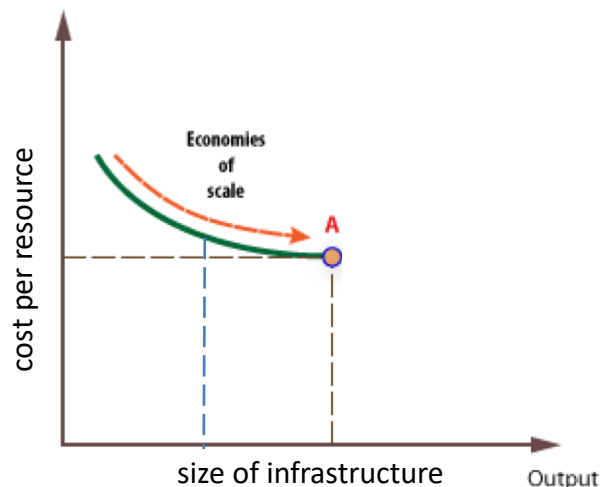
Many Datacenters Do Cheaper Themselves

Public cloud is a high-volume business

Larger data centers can in principle deploy computational resources at significantly lower cost than smaller ones

Economies of Scale

- Infrastructure innovation
- Technical economies
- Purchasing economies
- Savings
- ...



Diseconomies of Scale

- Communication overheads
- Coordination
- Complex management
- Complex operation
- ...

Huge data centers are only marginally less expensive than medium sized data centers

Point where economies of scale advantages go down (50K servers)

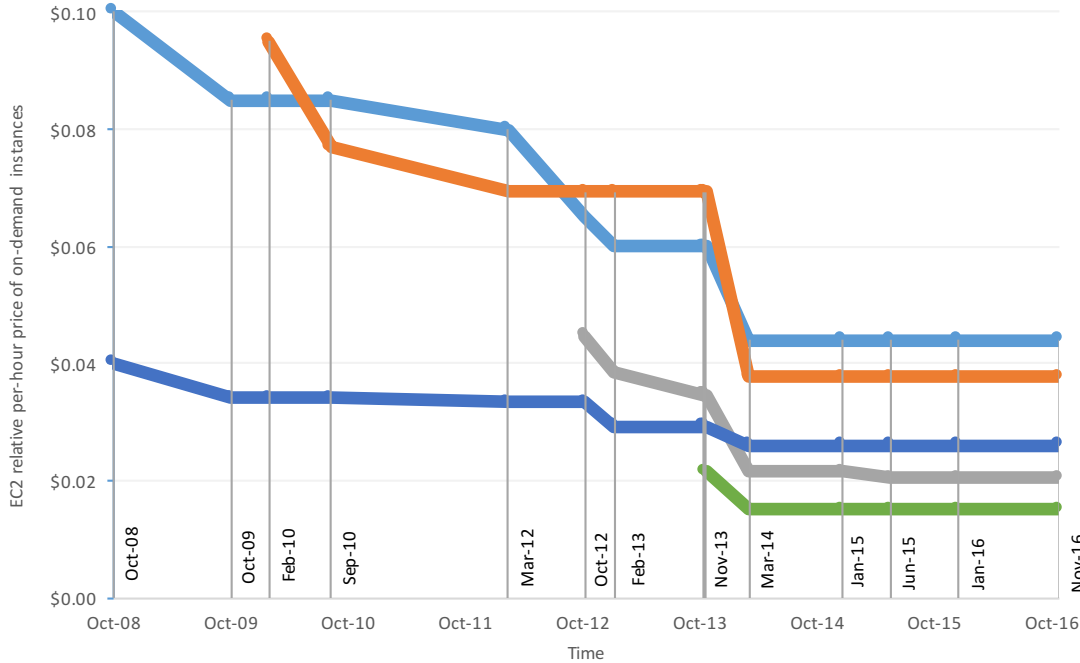
Many datacenters have sufficient scale to be able to do it cheaper themselves

Sources:

- <http://datacenterfrontier.com/inside-amazon-cloud-computing-infrastructure/>
- http://economicsonline.co.uk/Business_economics/Economies_of_scale.html

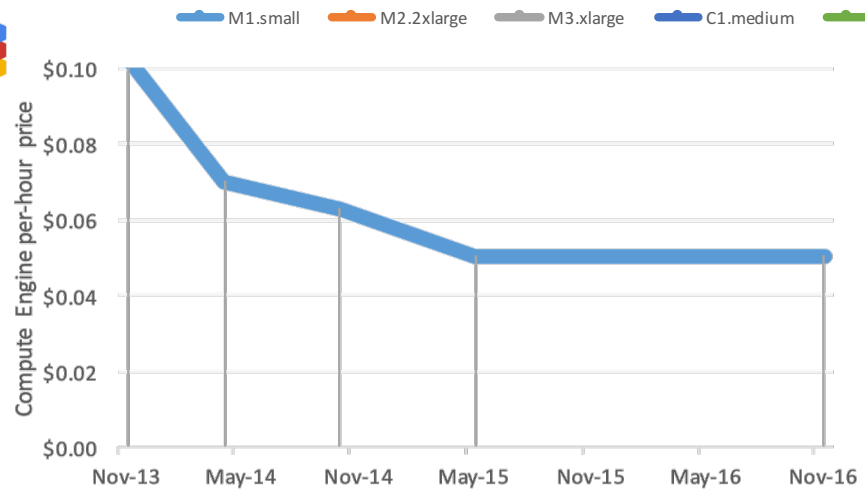
The Need for Private Clouds

Public Cloud is Becoming more Expensive



10% annual drop

Public Cloud pricing is not tracking Moore's Law with hardware cost dropping 30% annually



21% annual drop

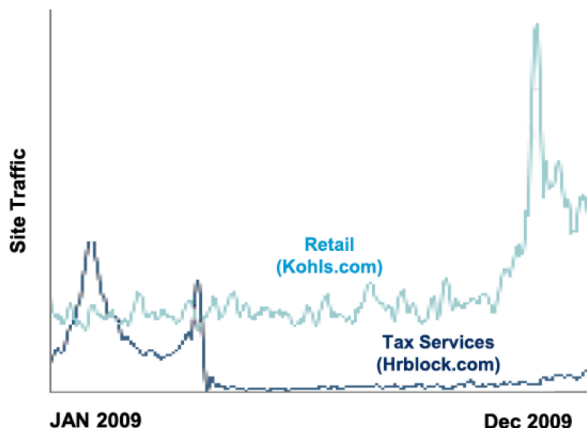
DOE centers have historically delivered average improvements in computing capability of 40%-80% per year with relatively flat budgets

Sources:
 • Companies web sites
 • http://science.energy.gov/~media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf
 Dr. David Sondak

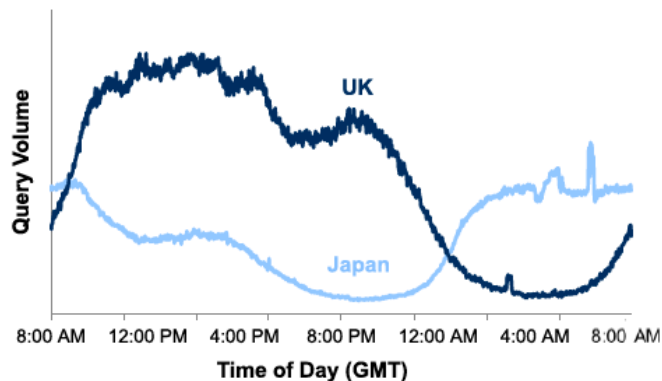
The Need for Private Clouds

Cost Benefit Strongly Depends on Variability

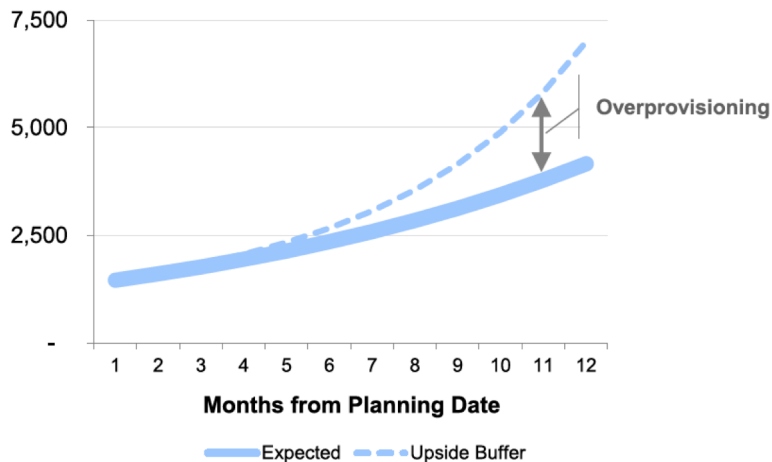
- Industry-specific patterns



- Time of day patterns



- Uncertain grown patterns



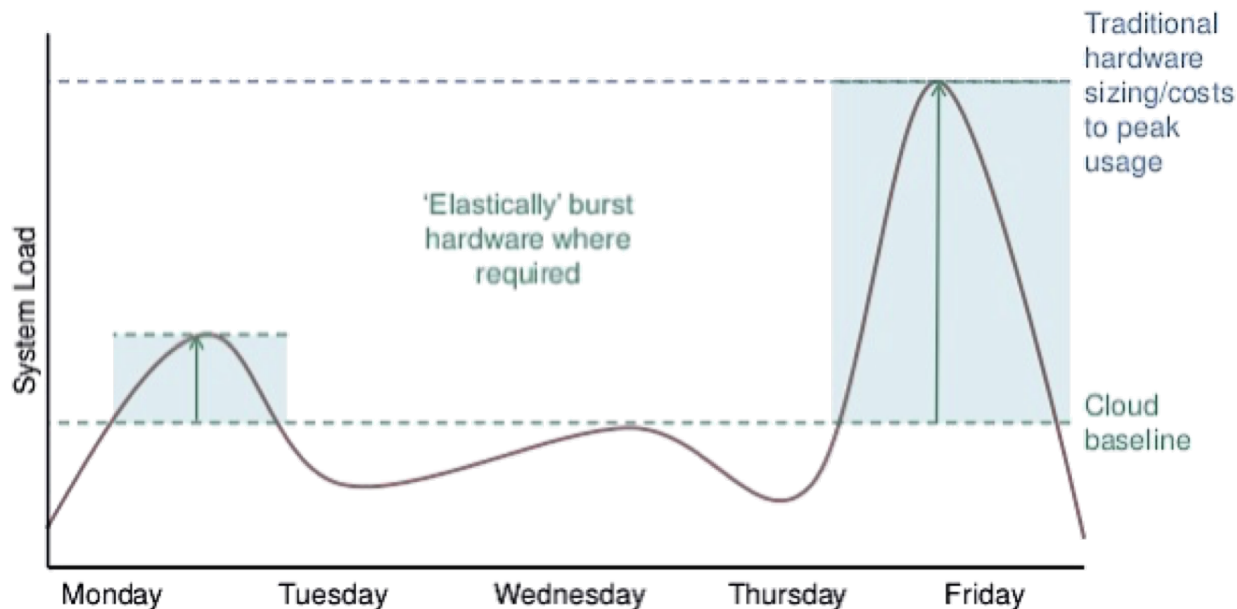
Source: *The Economics of the Cloud*, Microsoft

The Need for Private Clouds

Cost Benefit Strongly Depends on Variability

Hybrid cloud offers the optimal cost solution

- Workloads that are highly variable over time or with uncertain growth pattern
- Public cloud can be as much as four times more expensive than traditional data center, mostly for traditional legacy apps running 24/7 at peak on redundant architecture
- There is always a level of utilization at which private cloud becomes more cost-effective than public cloud



Sources: <http://www.rightscale.com/blog/cloud-industry-insights/cloud-computing-trends-2015-state-cloud-survey>

The Need for Private Clouds

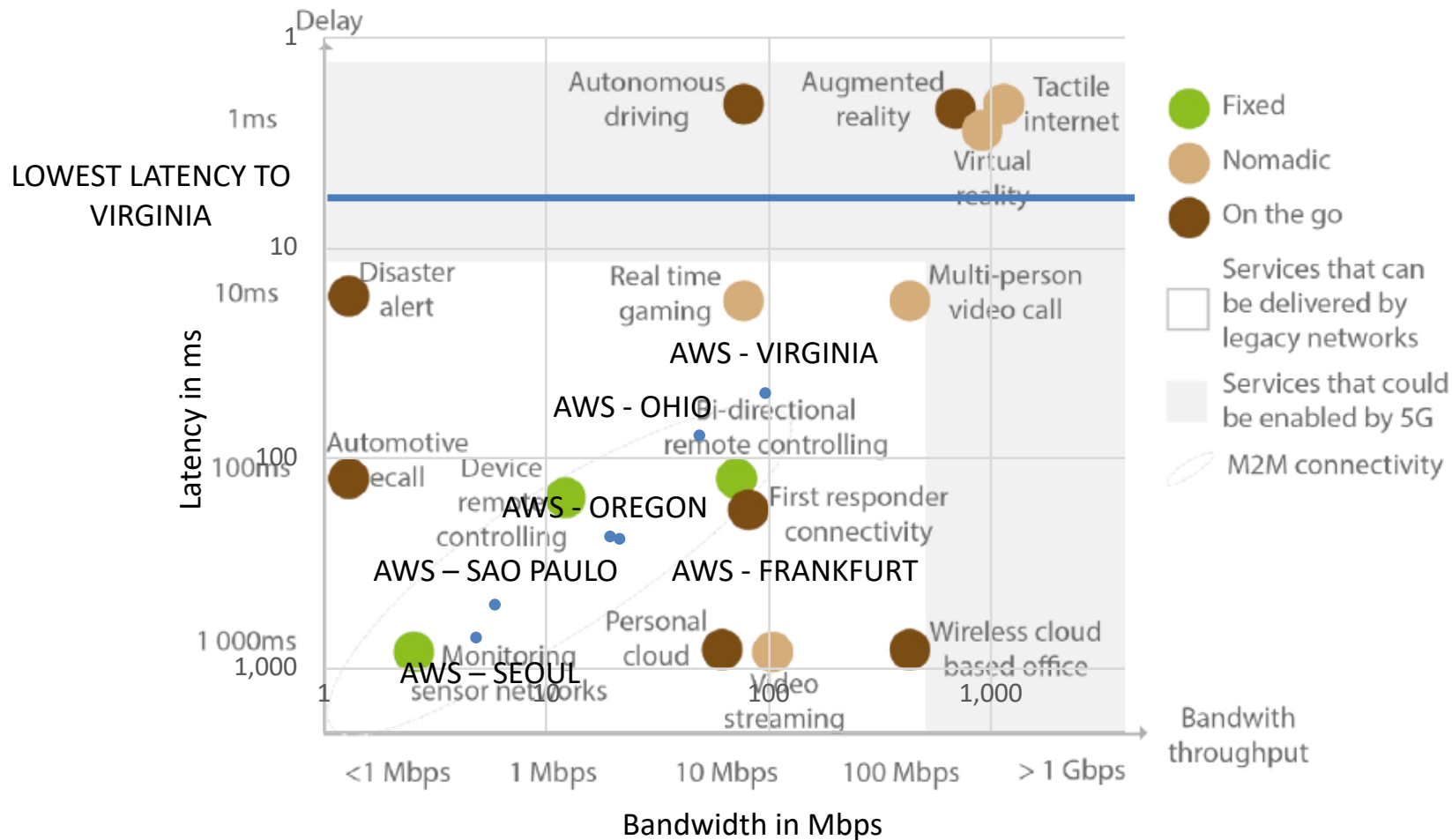
Not All Applications are Public Cloud-friendly

Performance needs?

1. Proximity to users or on-premises DC
2. Special high performance infrastructure

The Need for Private Clouds

When Proximity Matters



The Need for Private Clouds

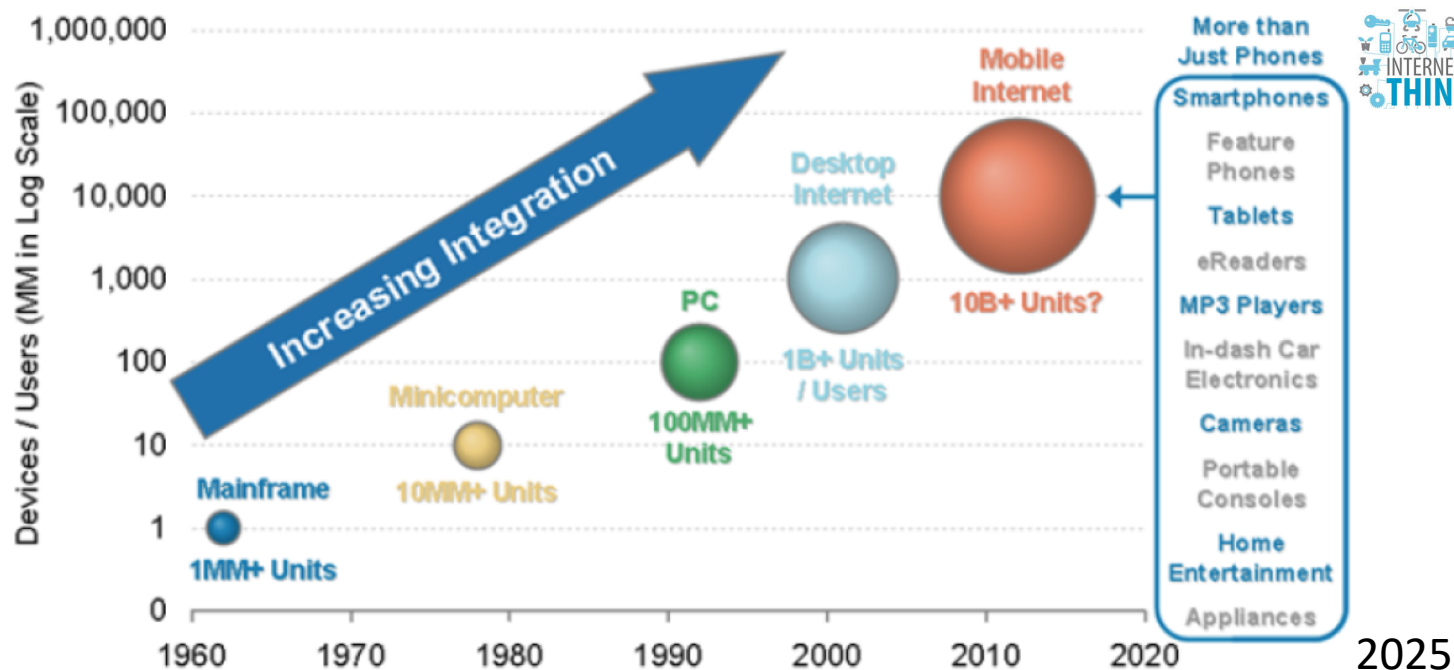
When Proximity Matters

Example: Distributed Big Data Processing

Each new computing cycle typically generates around 10x the installed base of the previous cycle

Devices or users in millions; logarithmic scale

100 BILLION DEVICES



@KPCB

Source: Morgan Stanley Mobile Internet Report (12/09)

The Need for Private Clouds

Performance

Example: High Performance Computing



TOP 10 Sites for November 2018

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,397,824	143,500.0	200,794.9	9,783

143 PFlops with 2M cores and 4.5K servers

Smaller than one AWS datacenter (40-80K servers)

But with different networking, CPU and I/O trade-offs more oriented to the efficient execution of tightly coupled application

The Need for Private Clouds

Public Cloud Is a Fraction of IT Infrastructure Market

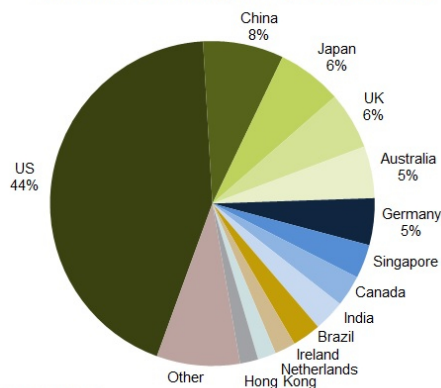
Public Cloud represents 20% of Internet computing power
10 million blade servers shipped annually
50 million servers in operation in 2017 (not including public cloud)

Public Cloud represents a tiny fraction of existing data centers
3 million data centers only in the U.S. in 2013
Number datacenters growing, peaking at 8.6 M worldwide by 2017

Public Cloud represents 50% of existing hyperscale DCs
There are now close to 400 Hyper-Scale Data Centers in the world

Hyperscale Data Center Operators

Data Center Locations by Country - December 2017

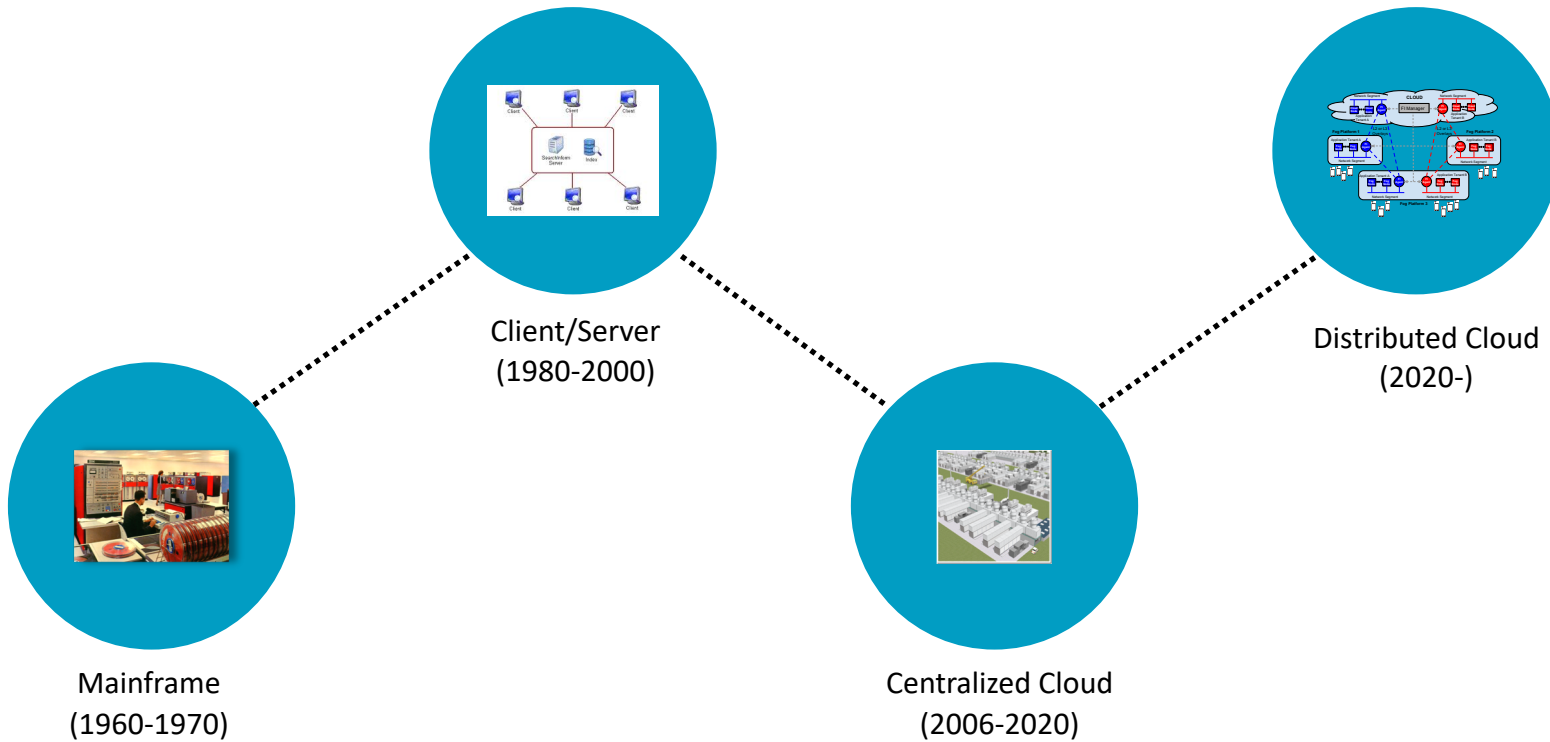


Sources:

- <http://www.gartner.com/newsroom/id/3530117>
- <http://energy.gov/eere/articles/10-facts-know-about->
- <http://worldstopdatacenters.com/idc-number-of-data-centers-will-grow-to-8-million-in-2017-when-begin-to-decline>
- <http://www.datacenterknowledge.com/cloud/research-there-are-now-close-400-hyper-scale-data-centers-world>

Evolution Toward a Distributed Cloud

Back to the Future

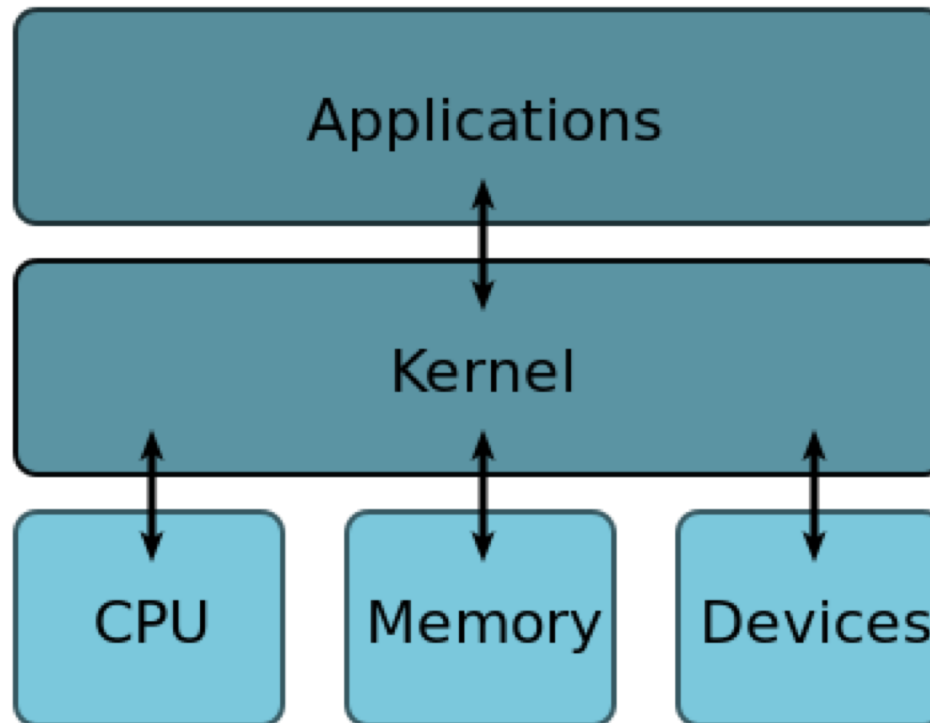


The Anatomy of the Cloud

What is an Operating System?

“An operating system (OS) is system software that manages computer hardware and software resources and provides common services for computer programs”

(source: Wikipedia)



The Anatomy of the Cloud

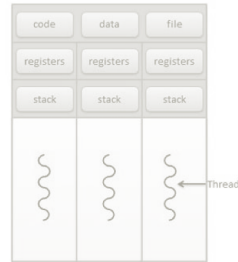
What is an Operating System?



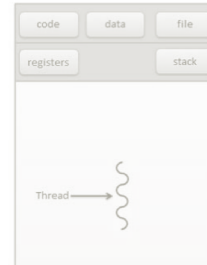
PROCESSES



Single threaded Process



Multi-threaded Process



Single threaded Process

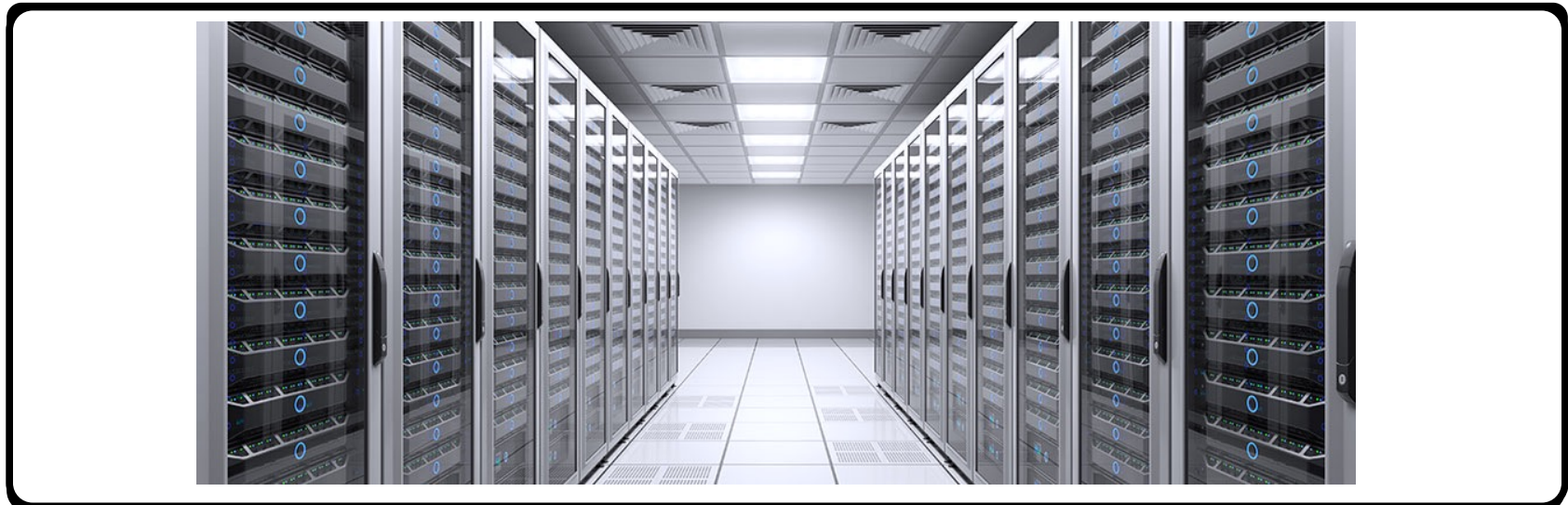
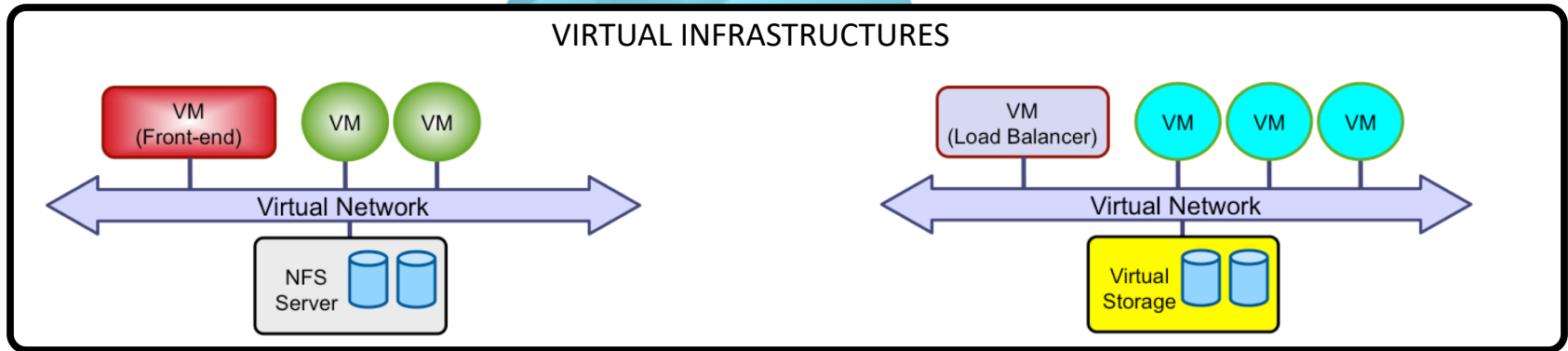


Multi-threaded Process



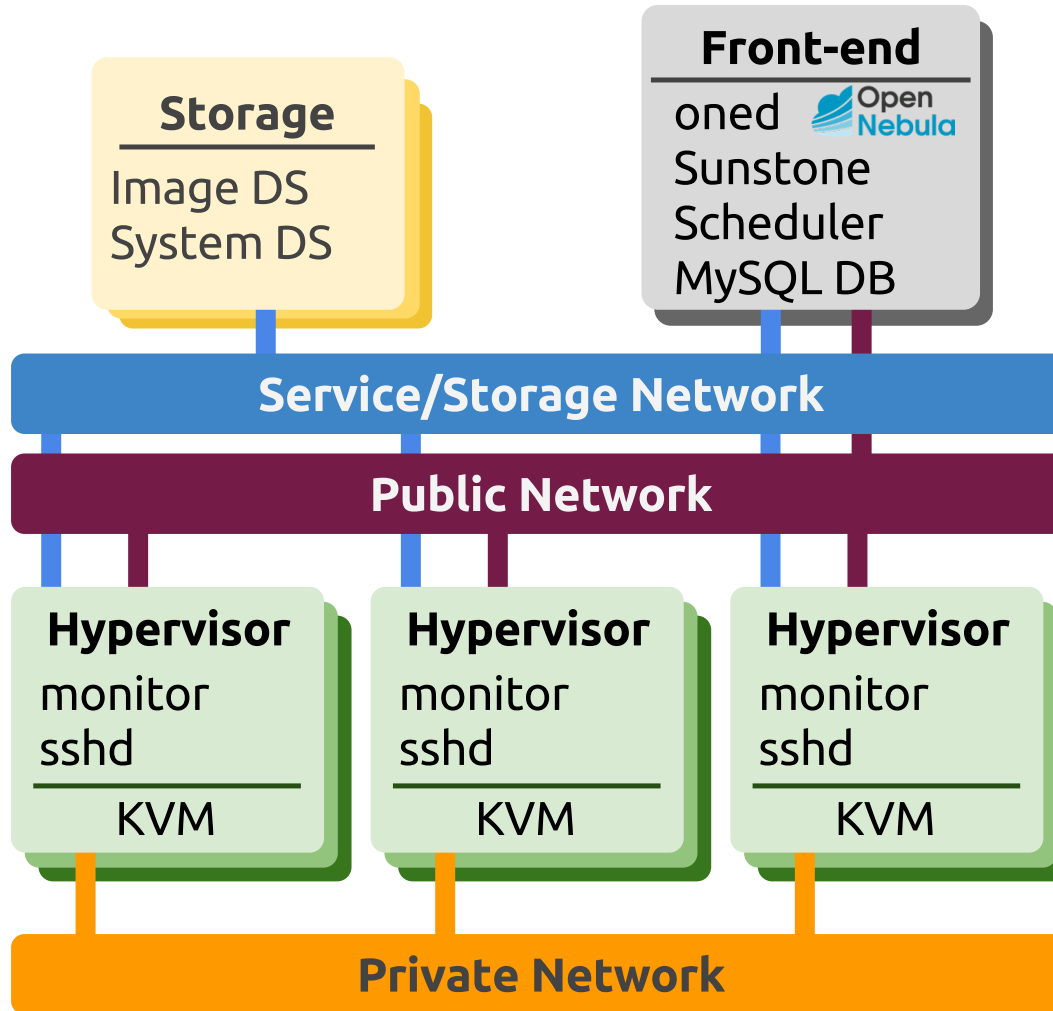
The Anatomy of the Cloud

What is a Cloud Management Platform?



The Anatomy of the Cloud

The Internals of a Cloud Instance



Reading Assignments / Open Discussion

Potential of Cloud for Scientific Applications

J. Riley, J. Noss, W. Dillingham, J. Cuff, I. M. Llorente, “*A High-Availability Cloud for Research Computing*”,
IEEE Computer, Volume: 50, Issue: 6, 2017.

What are the basic components of the architecture?

Why a private cloud?

What is data centric cloud computing?

Next Steps

- Get ready for **next lecture**:
A.4. Application parallelism

Questions

Practical Aspects of Cloud Computing

