Chapter2

Technologies for handing of Big Data

Basanta Joshi, PhD

Asst. Prof., Depart of Electronics and Computer Engineering

Program Coordinator, MSc in Information and Communication Engineering

Member, Laboratory for ICT Research and Development (LICT)

Member, Research Management Cell (RMC)

Institute of Engineering

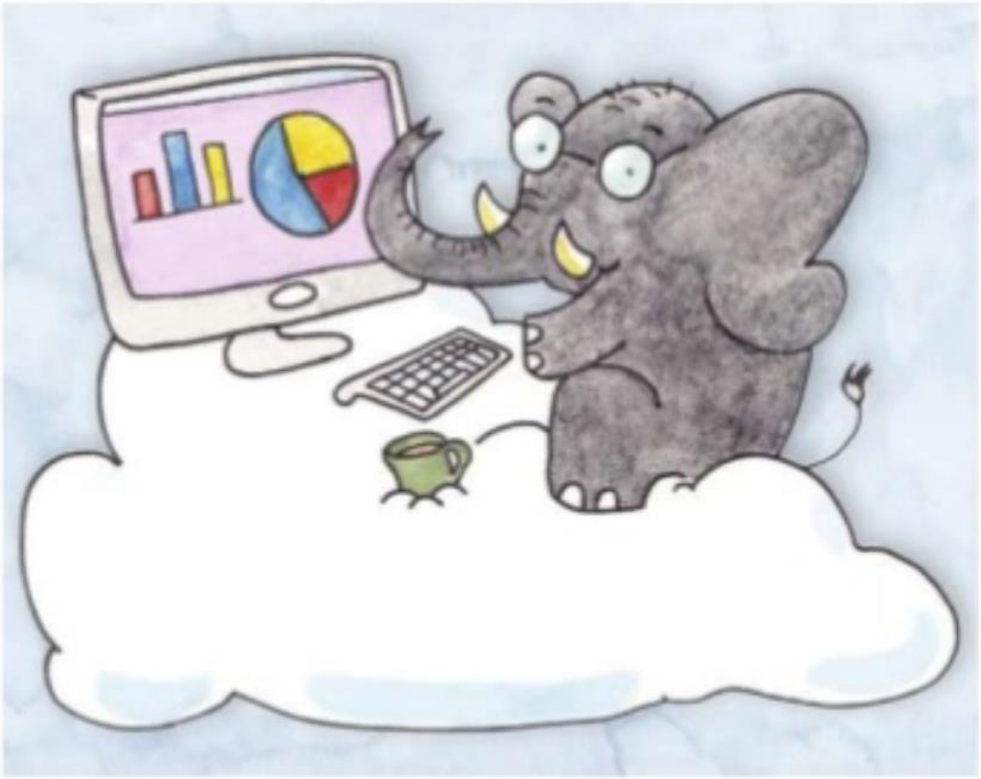
basanta@ioe.edu.np

http://www.basantajoshi.com.np

https://scholar.google.com/citations?user=iocLiGcAAAAJ

https://www.researchgate.net/profile/Basanta_Joshi2

Big Data In the Cloud



Slides from **Matei Zaharia** matei@cs.stanford.edu

Cloud Computing, Big Data

Hardware

Google 1997

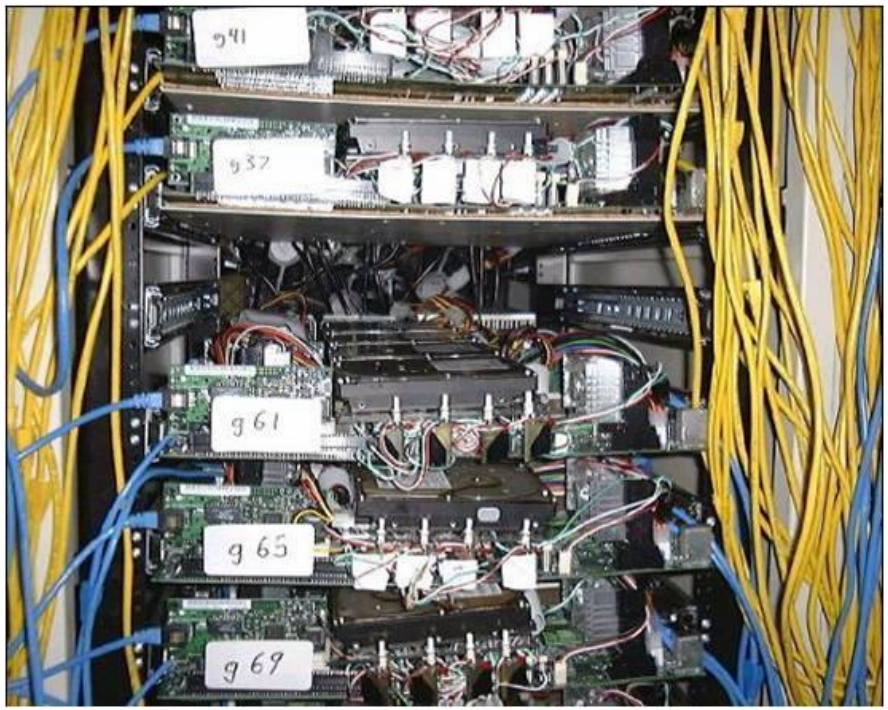
Data, Data, Data

"…Storage space must be used efficiently to store indices and, optionally, the documents themselves. The indexing system must process hundreds of gigabytes of data efficiently…"

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

Google 2001



Commodity CPUs

Lots of disks

Low bandwidth network

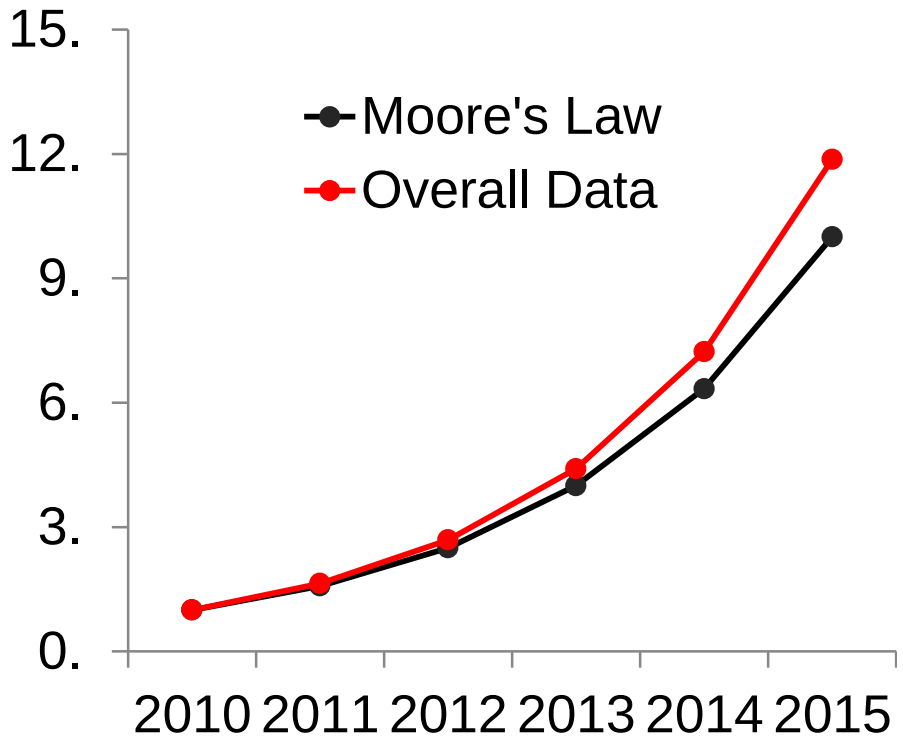
Cheap !

Datacenter evolution

Facebook's daily logs: 60 TB

1000 genomes project: 200 TB

Google web index: 10+ PB



(IDC report*)

Slide from Ion Stoica

Datacenter Evolution

Google data centers in The Dalles, Oregon

Datacenter Evolution

Capacity:

~10000 machines

Bandwidth:

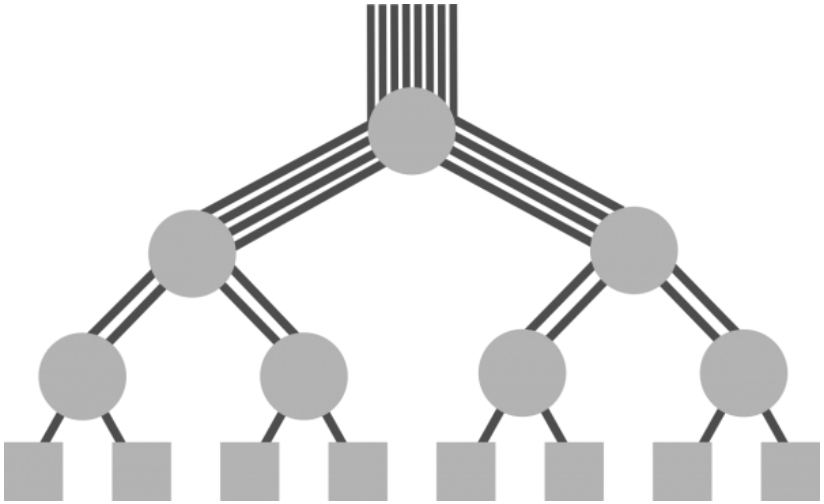
12-24 disks per node

Latency:

256GB RAM cache

Datacenter Networking

Initially tree topology
Over subscribed links



Fat tree, Bcube, VL2 etc.

Lots of research to get full bisection bandwidth

Datacenter Design

Goals

Power usage effectiveness (PUE)

Cost-efficiency

Custom machine design

Open Compute Project (Facebook)

Datacenters → Cloud Computing

"…long-held dream of computing as a utility…"

Above the Clouds: A Berkeley View of Cloud Computing

Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz,
Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia
(Comments should be addressed to abovetheclouds@cs.berkeley.edu)

UC Berkeley Reliable Adaptive Distributed Systems Laboratory *
http://radlab.cs.berkeley.edu/

From Mid 2006

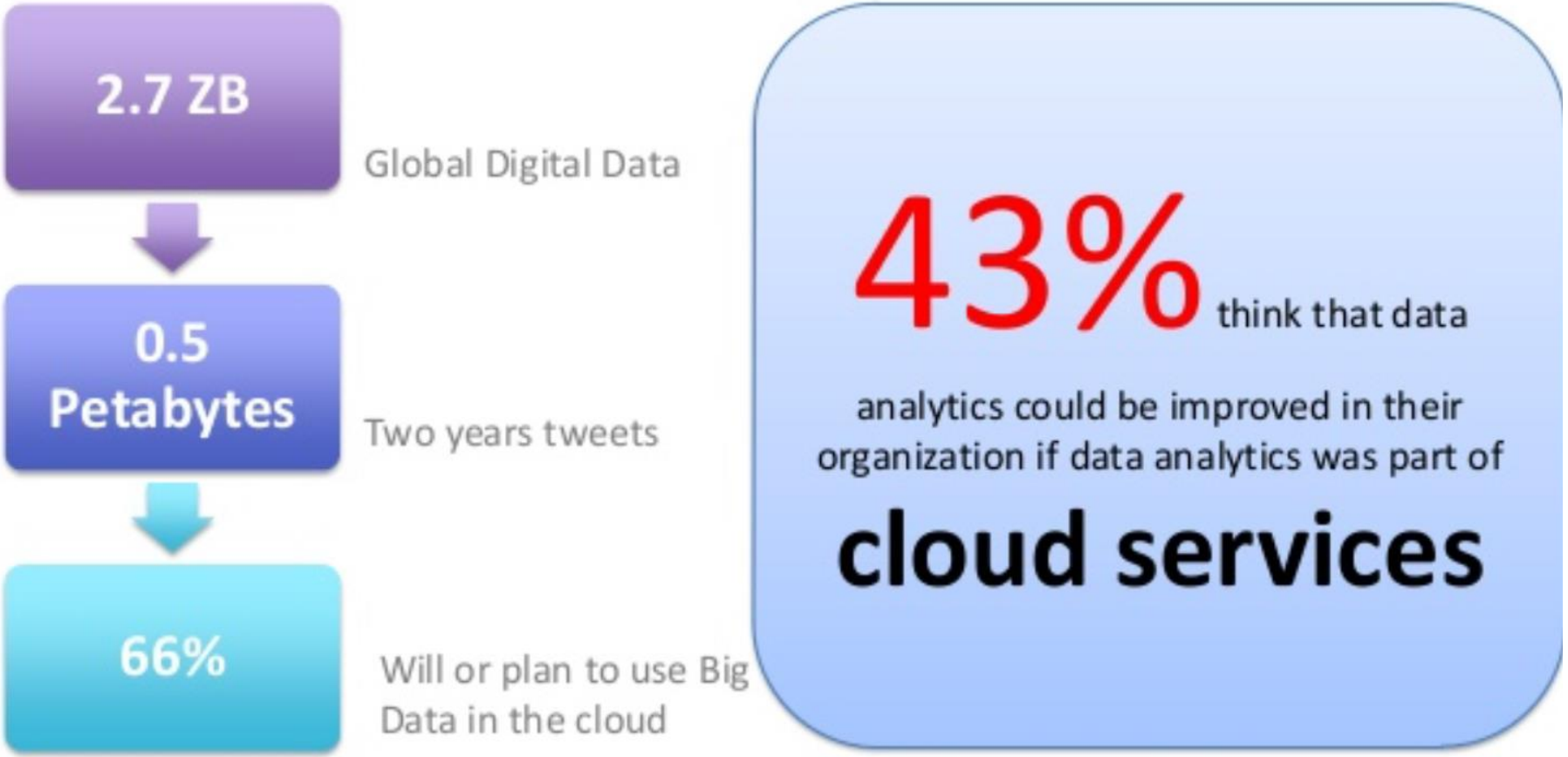
Rent virtual computers in the "Cloud"

On-demand machines, spot pricing

About GigaSpaces

100's of Enterprise Customers

Big Data In The Cloud



The Challenges..

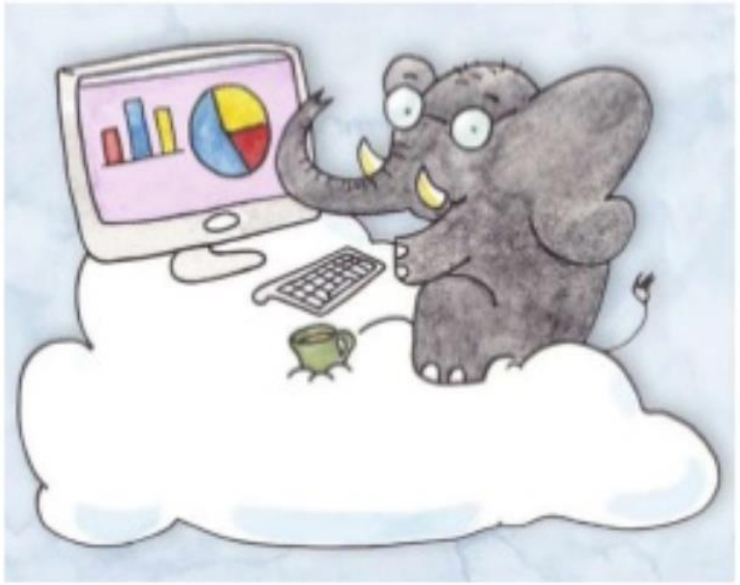
Ever Growing Data

Deeper Correlation

Tight Performance

The Challenge

The Solution



Big Data in the Cloud

Big Data in the Cloud- 3 Reasons

Holger Kisker

Forbes

- **Skills**
  - Do you really need/want this all in-house?
- **Huge amounts of external data.**
  - Does it make sense to move and manage all this data behind your firewall?
- **Focus on the value of your data**
  - Instead of big data management.

Managing Big Data on the Cloud

- Auto start VMs
- Install and configure app components
- Monitor
- Repair
- (Auto) Scale
- Burst…

Big Data in the Cloud..

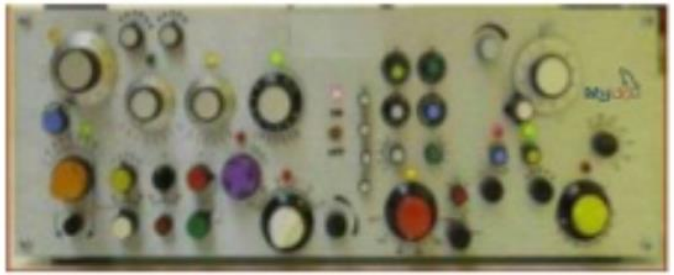
Reduce the Infrastructure Cost

Choose the Right Cloud for the Job

Running Bare-Metal for high I/O workloads, Public cloud for sporadic workloads.

Big Data in the Cloud ..

Reducing The Operational Complexity



- Consistent Management
- Automation Through the Entire Stack

Amazon EC2

Machine	Memory (GB)	Compute Units (ECU)	Local Storage (GB)	Cost / hour
t1.micro	0.615	2	0	$0.02
m1.xlarge	15	8	1680	$0.48
cc2.8xlarge	60.5	88 (Xeon 2670)	3360	$2.40

1 ECU = CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor

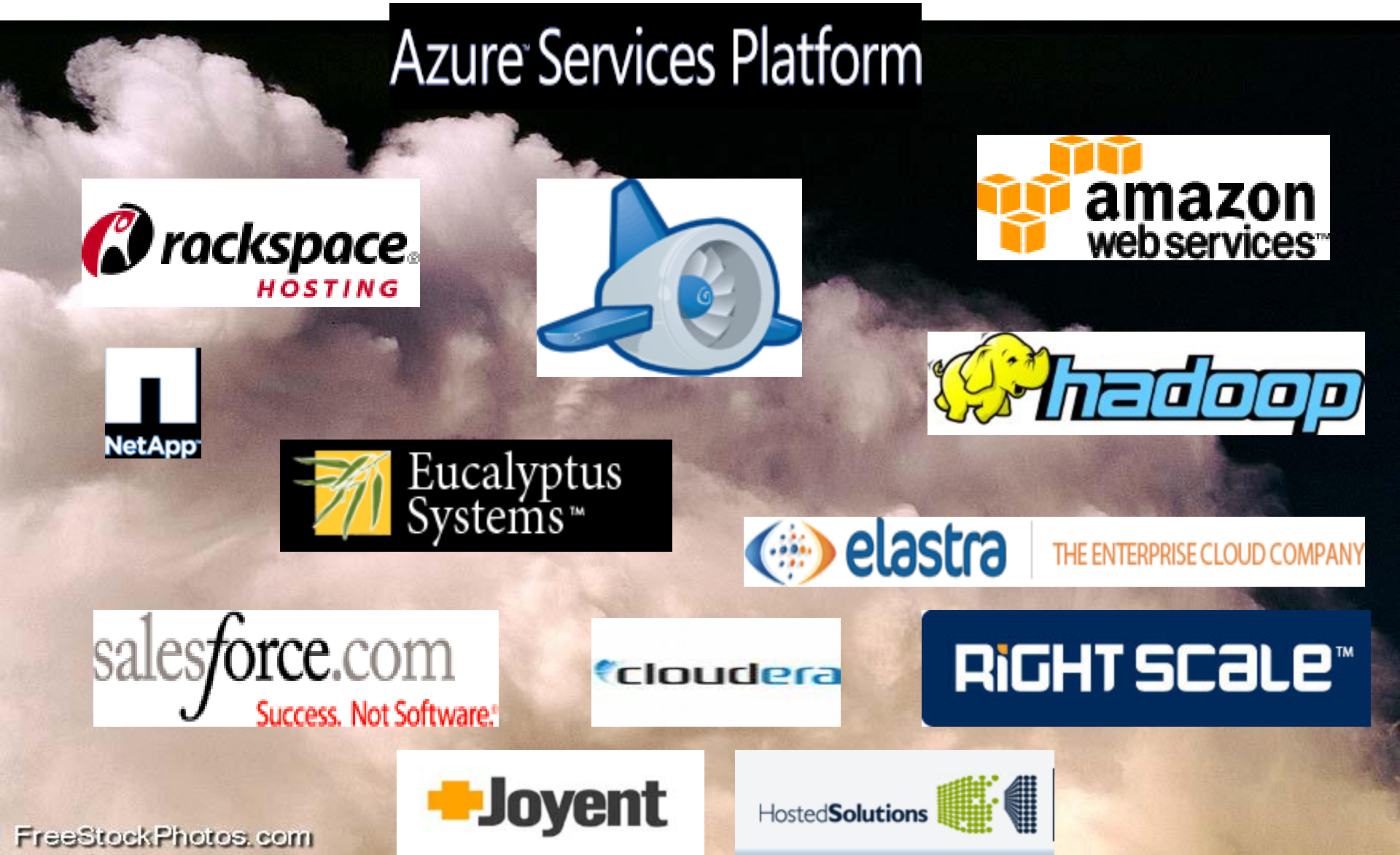
Hardware

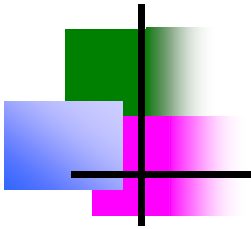
Hopper vs. Datacenter

	Hopper	Datacenter[2]
Nodes	6384	1000s to 10000s
CPUs (per node)	2x12 cores	~2x6 cores
Memory (per node)	32-64GB	~48-128GB
Storage (overall)	~4 PB	120-480 PB
Interconnect	~ 66.4 Gbps	~10Gbps

[2]http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/

Paradigm Shift in Computing

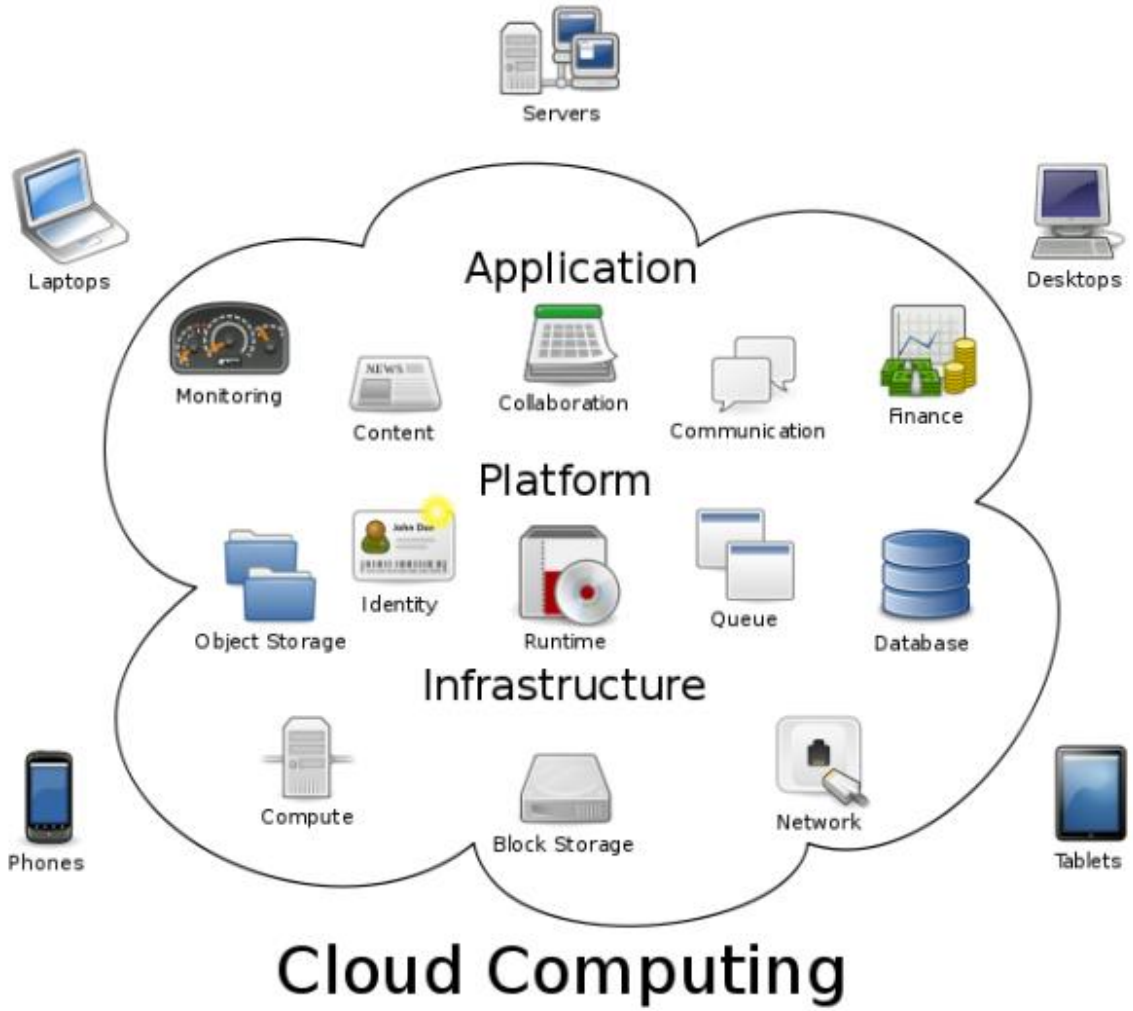




What is Cloud Computing?

- IT resources provided as a service
  - Compute, storage, databases, queues
- Clouds leverage economies of scale of commodity hardware
  - Cheap storage, high bandwidth networks & multicore processors
  - Geographically distributed data centers
- Offerings from Microsoft, Amazon, Google, …

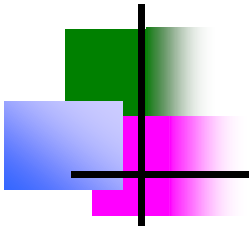
What is Cloud Computing?



Cloud Computing: History

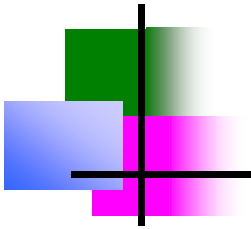
> If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility... The computer utility could become the basis of a new and important industry.

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



Cloud Computing: Why Now?

- Experience with very large datacenters
  - Unprecedented economies of scale
  - Transfer of risk
- Technology factors
  - Pervasive broadband Internet
  - Maturity in Virtualization Technology
- Business factors
  - Minimal capital expenditure
  - Pay-as-you-go billing model

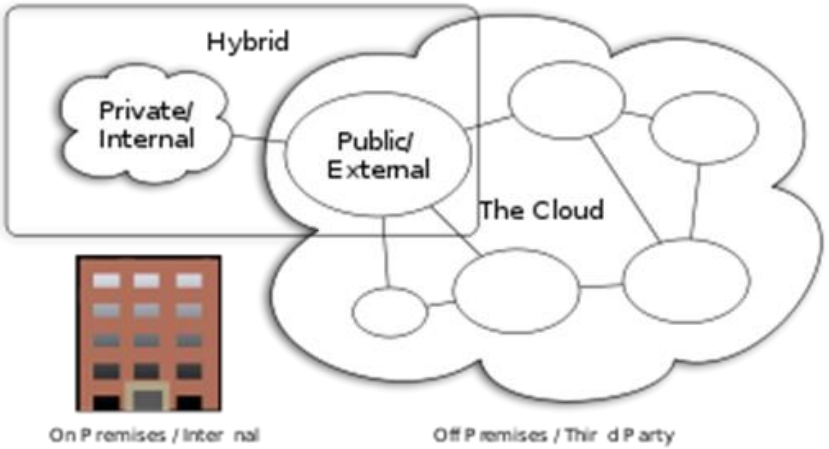


Benefits

- Cost & management
  - Economies of scale, “out-sourced” resource management
- Reduced Time to deployment
  - Ease of assembly, works “out of the box”
- Scaling
  - On demand provisioning, co-locate data and compute
- Reliability
  - Massive, redundant, shared resources
- Sustainability
  - Hardware not owned

Types of Cloud Computing

- **Public Cloud**: Computing infrastructure is hosted at the vendor's premises.
- **Private Cloud**: Computing architecture is dedicated to the customer and is not shared with other organisations.
- **Hybrid Cloud**: Organisations host some critical, secure applications in private clouds. The not so critical applications are hosted in the public cloud
  - **Cloud bursting**: the organisation uses its own infrastructure for normal usage, but cloud is used for peak loads.
- **Community Cloud**

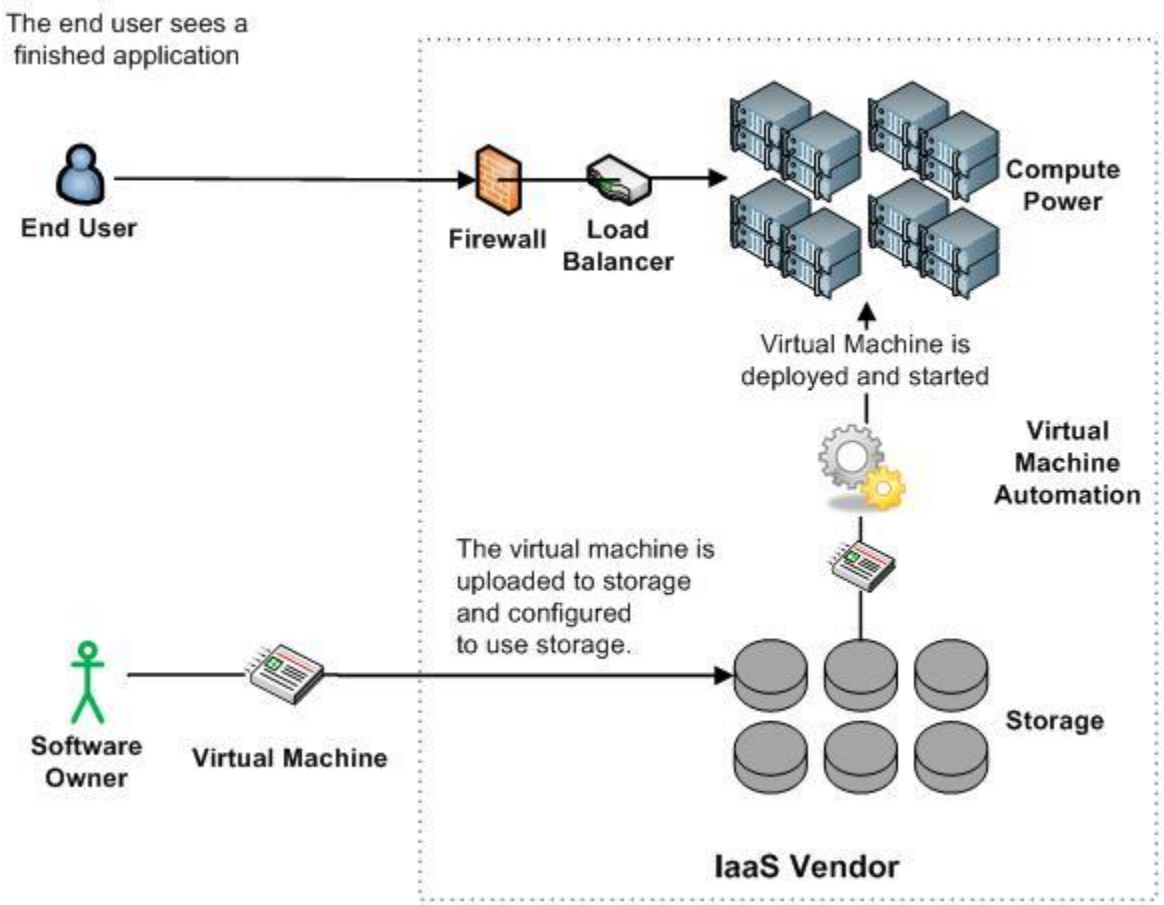


Cloud Computing Types

Classification of Cloud Computing based on Service Provided

- Infrastructure as a service (IaaS)
  - Offering hardware related services using the principles of cloud computing. These could include storage services (database or disk storage) or virtual servers.
  - Amazon EC2, Amazon S3, Rackspace Cloud Servers and Flexiscale.
- Platform as a Service (PaaS)
  - Offering a development platform on the cloud.
  - Google's Application Engine, Microsofts Azure, Salesforce.com's force.com .
- Software as a service (SaaS)
  - Including a complete software offering on the cloud. Users can access a software application hosted by the cloud vendor on pay-per-use basis. This is a well-established sector.
  - Salesforce.coms' offering in the online Customer Relationship Management (CRM) space, Googles gmail and Microsofts hotmail, Google docs.

Infrastructure as a Service (IaaS)



More Refined Categorization

- Storage-as-a-service
- Database-as-a-service
- Information-as-a-service
- Process-as-a-service
- Application-as-a-service
- Platform-as-a-service
- Integration-as-a-service
- Security-as-a-service
- Management/ Governance-as-a-service
- Testing-as-a-service
- Infrastructure-as-a-service

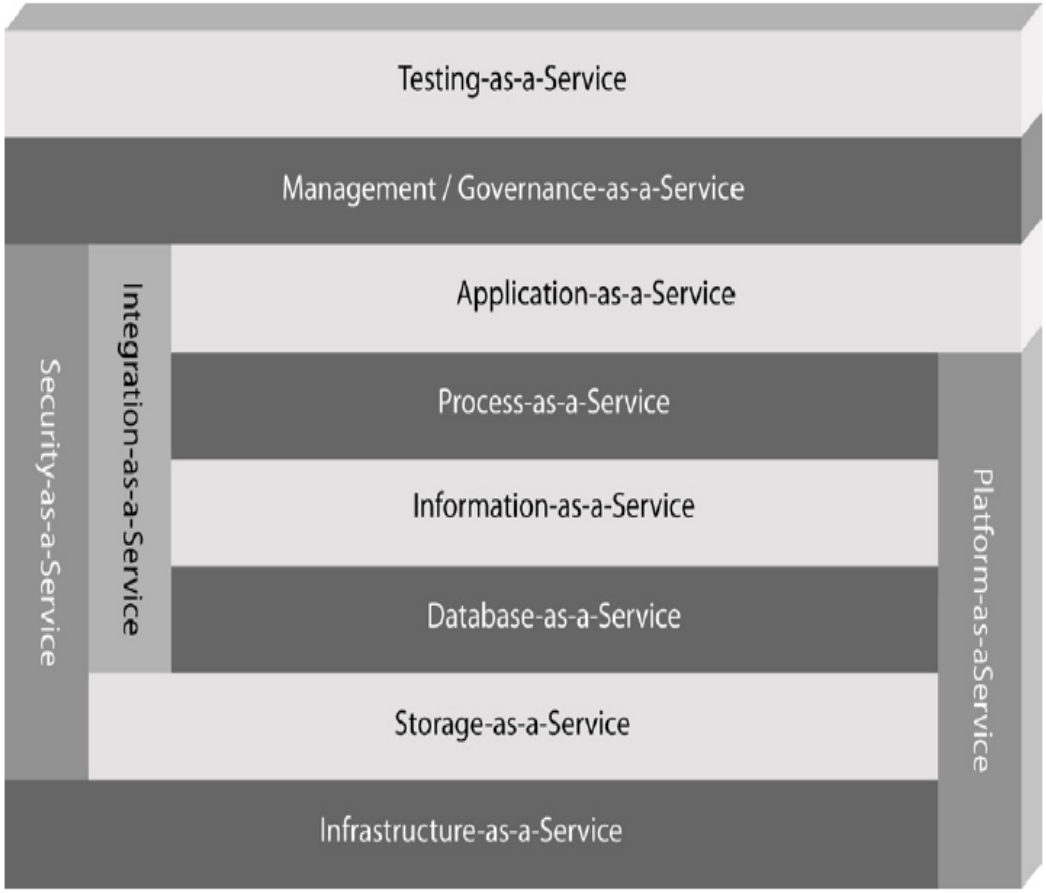


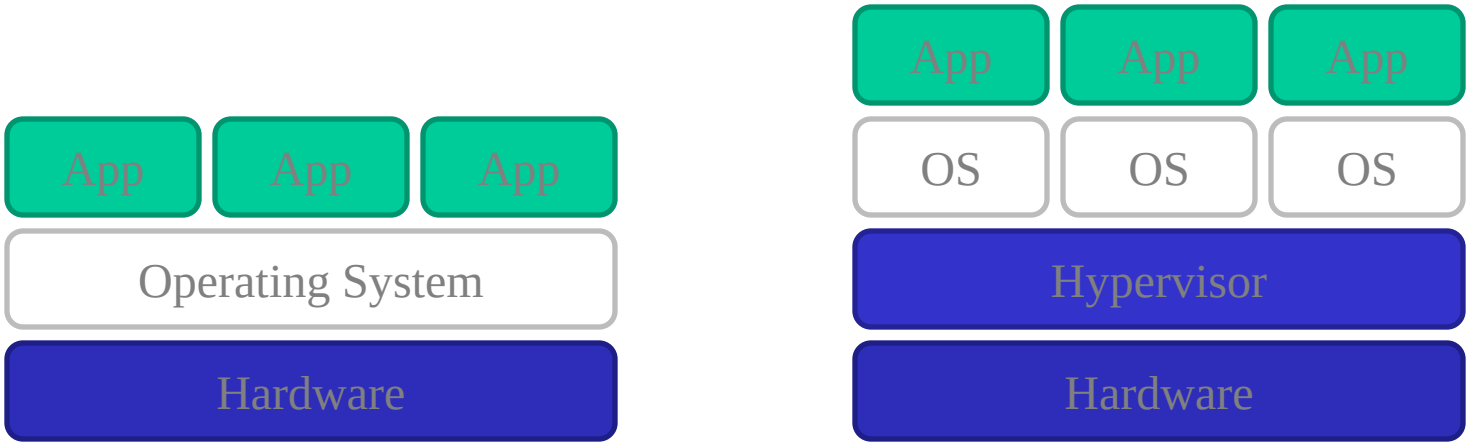
Figure 1: The patterns or categories of cloud computing providers allow you to use a discrete set of services within your architecture.

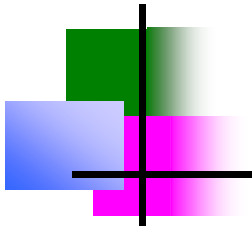
InfoWorld Cloud Computing Deep Dive

Key Ingredients in Cloud Computing

- Service-Oriented Architecture (SOA)
- Utility Computing (on demand)
- Virtualization (P2P Network)
- SAAS (Software As A Service)
- PAAS (Platform AS A Service)
- IAAS (Infrastructure AS A Service)
- Web Services in Cloud

Enabling Technology: Virtualization



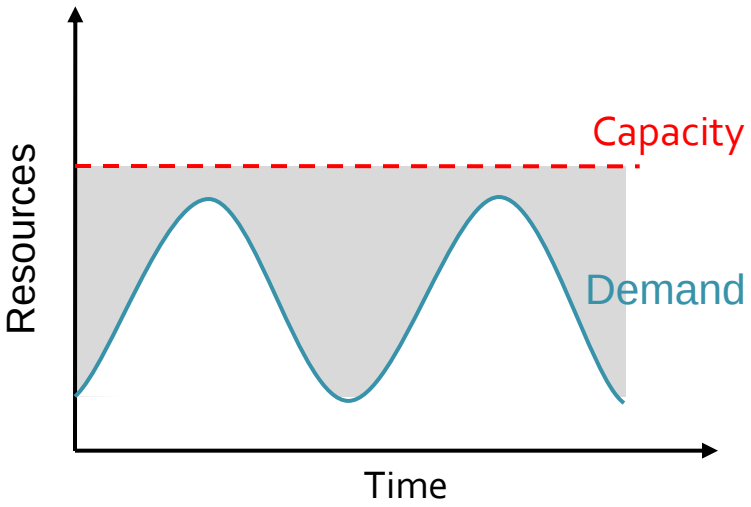


Everything as a Service

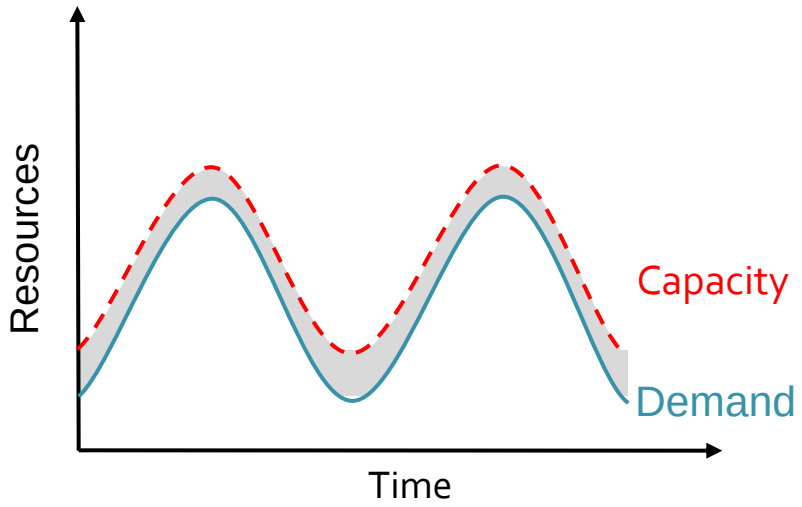
- Utility computing = Infrastructure as a Service (IaaS)
  - Why buy machines when you can rent cycles?
  - Examples: Amazon's EC2, Rackspace
- Platform as a Service (PaaS)
  - Give me nice API and take care of the maintenance, upgrades, …
  - Example: Google App Engine
- Software as a Service (SaaS)
  - Just run it for me!
  - Example: Gmail, Salesforce

Economics of Cloud Users

- Pay by use instead of provisioning for peak



Static data center



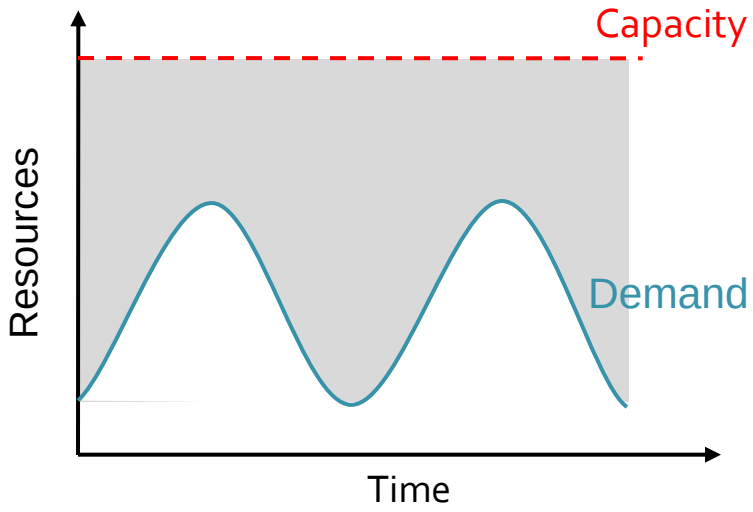
Data center in the cloud

Unused resources

Slide Credits: Berkeley RAD Lab

Economics of Cloud Users

- Risk of over-provisioning: underutilization



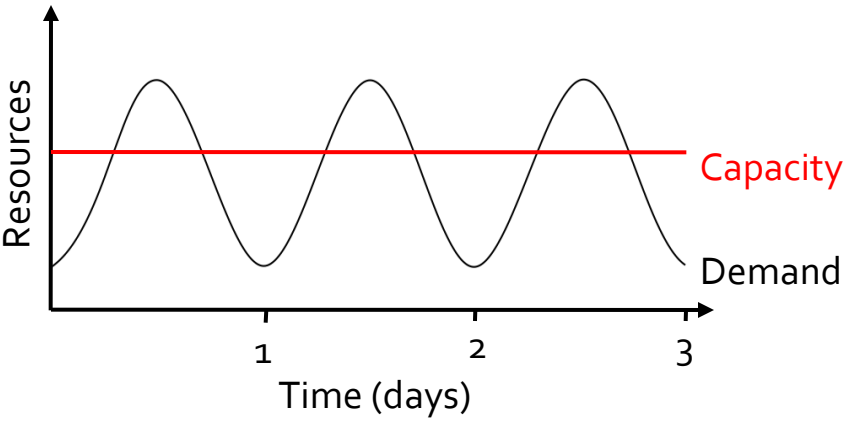
Static data center

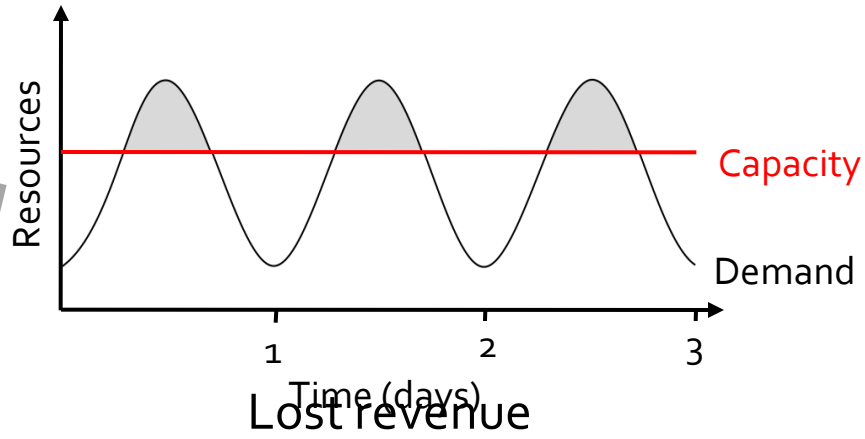
Unused resources

Slide Credits: Berkeley RAD Lab

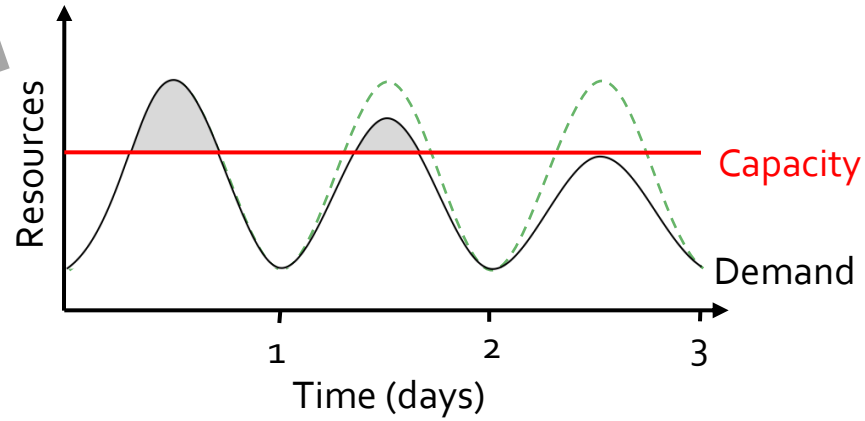
Economics of Cloud Users

- Heavy penalty for under-provisioning



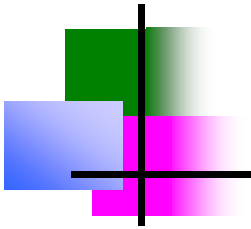


Lost revenue



Lost users

Slide Credits: Berkeley RAD Lab



The Big Picture

- Unlike the earlier attempts:
  - Distributed Computing
  - Distributed Databases
  - Grid Computing

- Cloud Computing is likely to persist:
  - Organic growth: Google, Yahoo, Microsoft, and Amazon
  - Poised to be an integral aspect of National Infrastructure in US and other countries

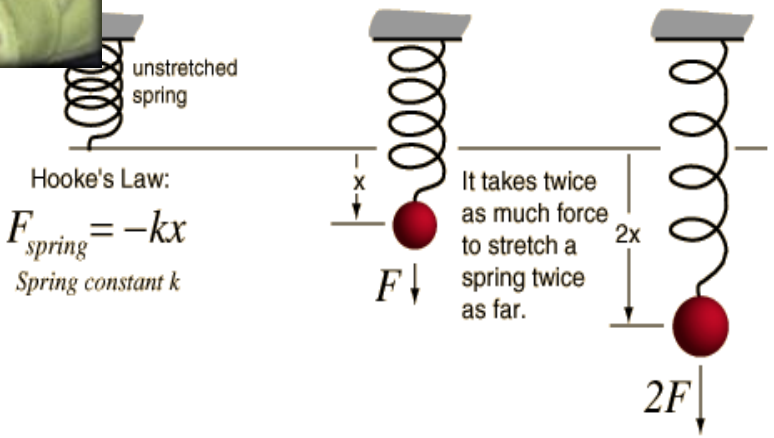
Cloud Reality

- Facebook Generation of Application Developers

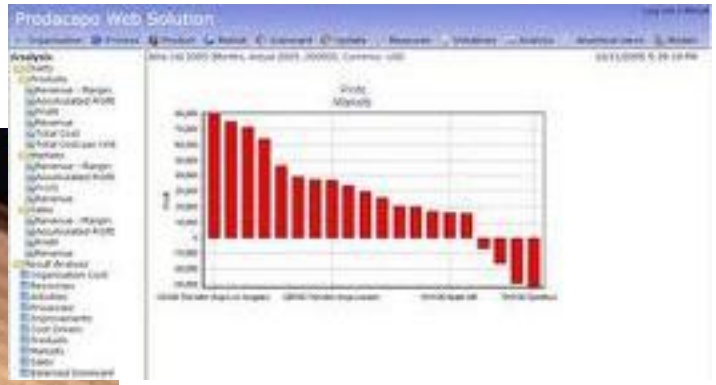
- Animoto.com:
  - Started with 50 servers on Amazon EC2
  - Growth of 25,000 users/hour
  - Needed to scale to 3,500 servers in 2 days (RightScale@SantaBarbara)

- Many similar stories:
  - RightScale
  - Joyent
  - …

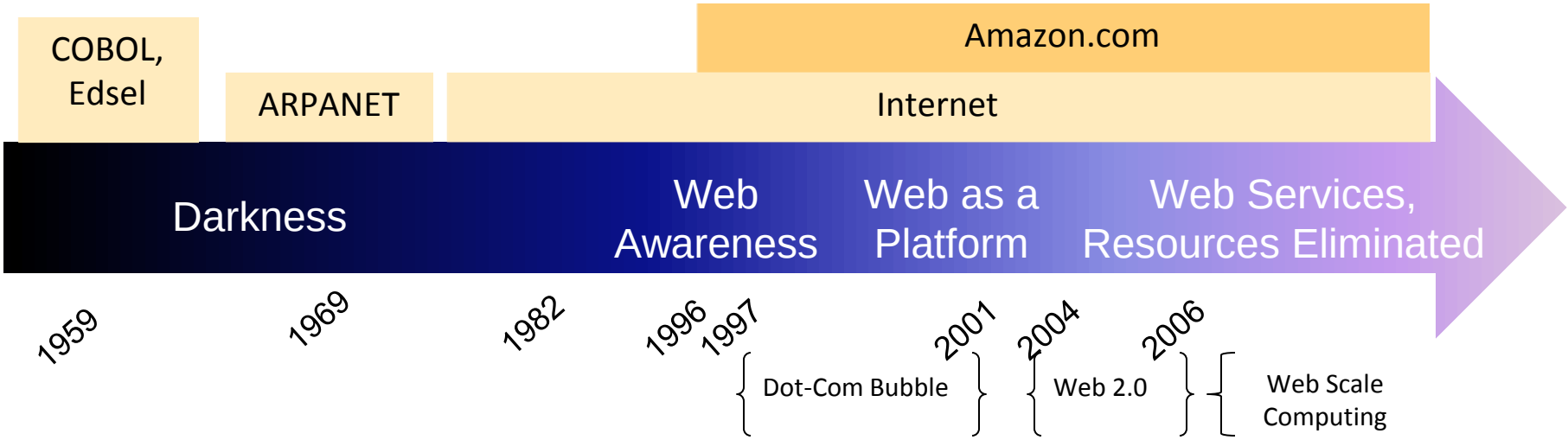
Cloud Challenges: Elasticity

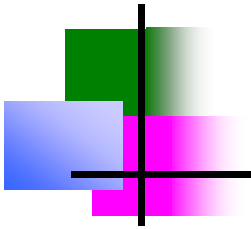


Cloud Challenges: Differential Pricing Models



The Obligatory Timeline Slide (Mike Culver @ AWS)

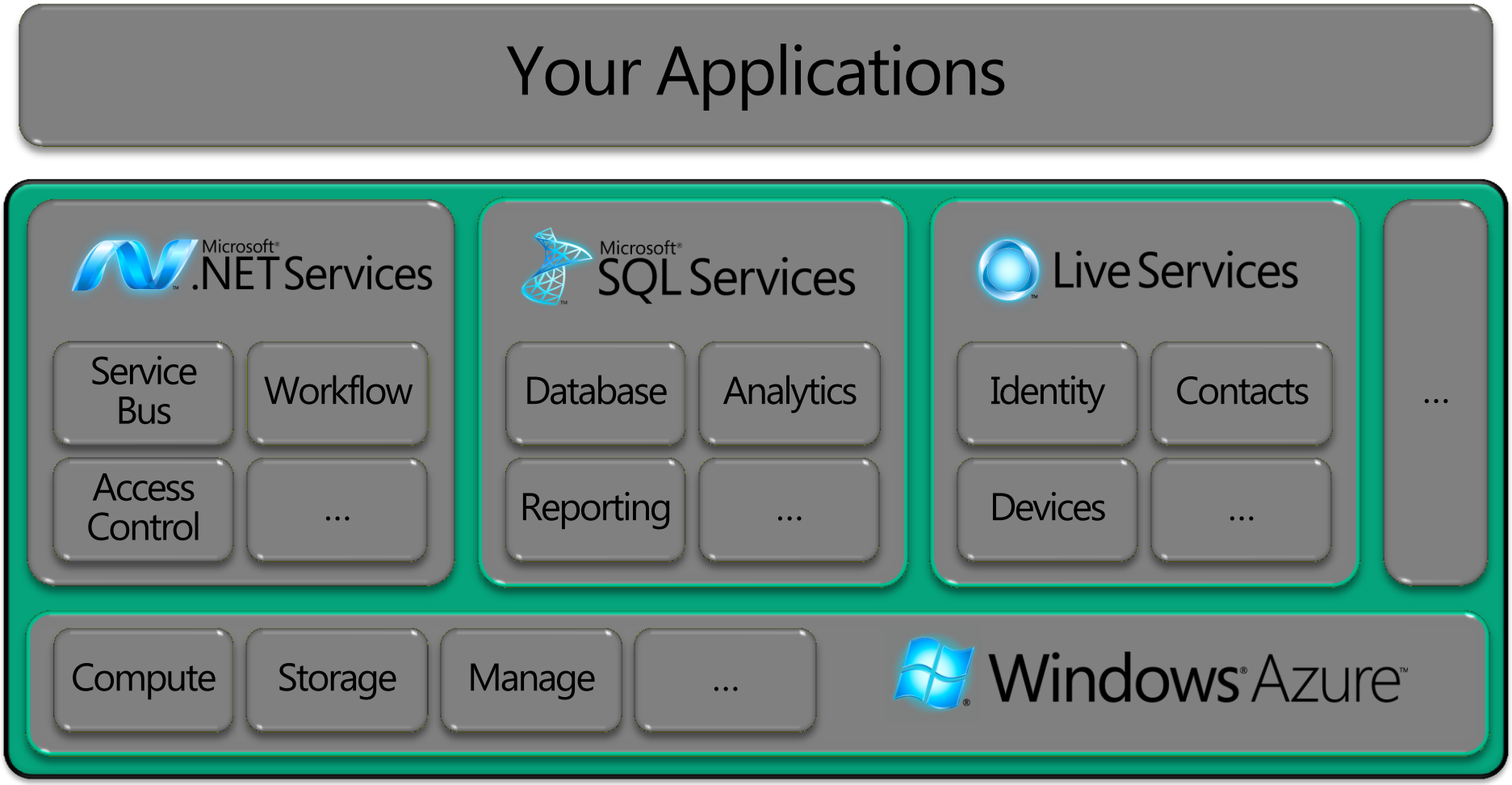




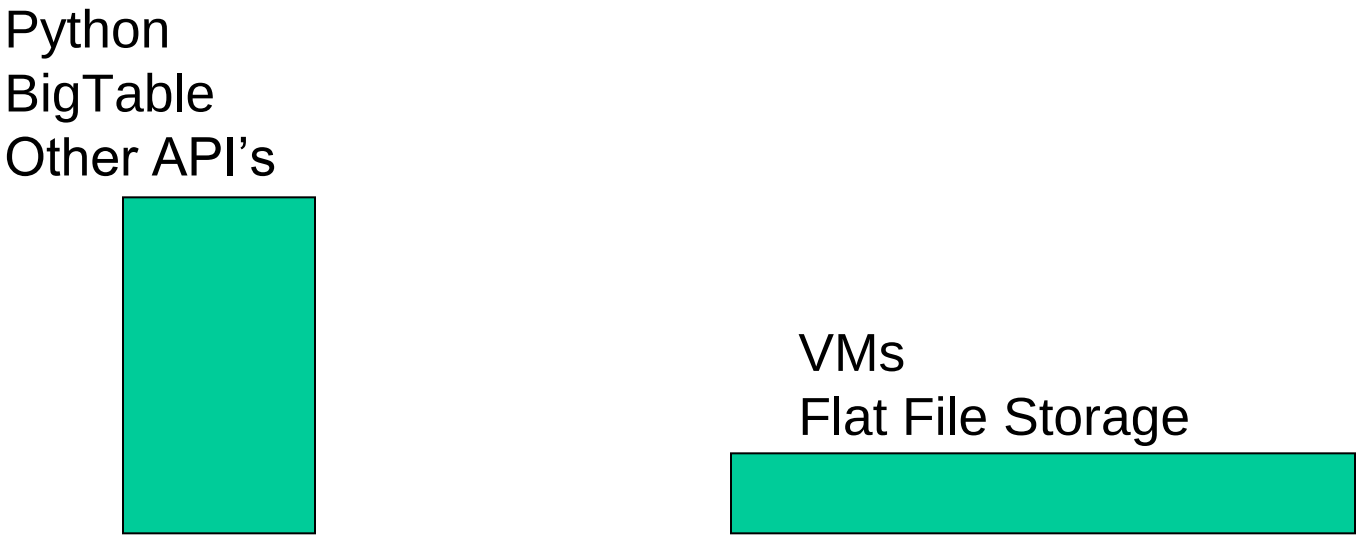
AWS

- Elastic Compute Cloud – EC2 (IaaS)
- Simple Storage Service – S3 (IaaS)
- Elastic Block Storage – EBS (IaaS)
- SimpleDB (SDB) (PaaS)
- Simple Queue Service – SQS (PaaS)
- CloudFront (S3 based Content Delivery Network – PaaS)
- Consistent AWS Web Services API

What does Azure platform offer to developers?



Google's AppEngine vs Amazon's EC2



AppEngine:

- Higher-level functionality (e.g., automatic scaling)
- More restrictive (e.g., respond to URL only)
- Proprietary lock-in

EC2/S3:

- Lower-level functionality
- More flexible
- Coarser billing model

Thank you !!!