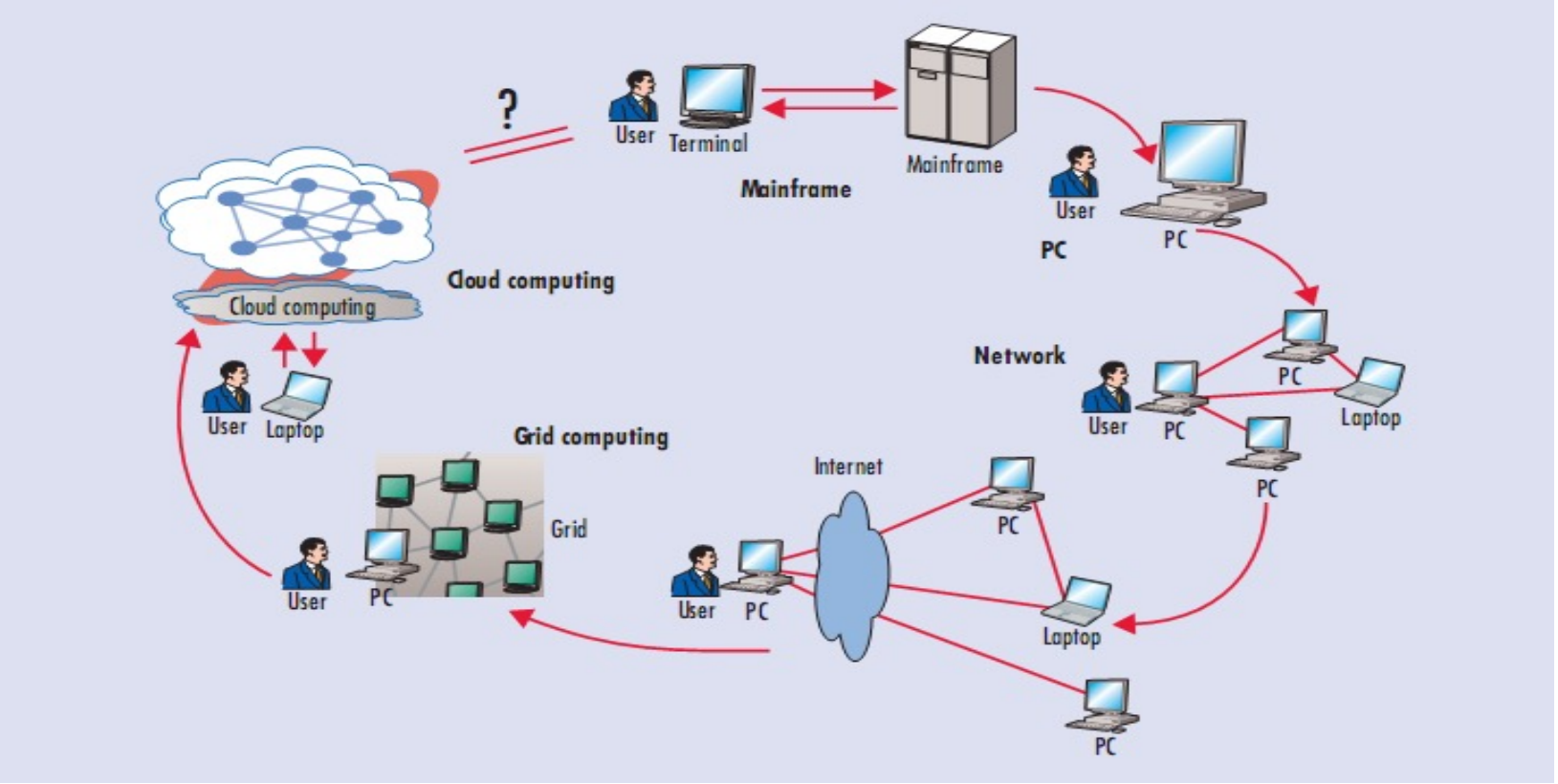
Introduction to Cloud Computing

Dr. Balaji Palanisamy
Associate Professor
School of Computing and Information
University of Pittsburgh
bpalan@pitt.edu

Computing Paradigm Shift

Cloud Challenges

- Cost-effectiveness, Performance and Security



Why? WEB is replacing the Desktop

Scope of this course

- Understand the basic ideas of cloud computing
- Get familiar with
  - Tools
  - Systems
- Expose to some research topics

Two major parts:

- Processing large data with the cloud
- Scaling up/down web applications with the cloud

Note: some programming parts need self-study

Prerequisites

- Some programming skills
  - Java
  - Comfortable with learning new programming exercises
- Sufficient knowledge about
  - Data structure
  - Computer Networks
  - Some knowledge of databases and operating systems is helpful though not required

Assignments and Grading

- Reading Assignment-based Quiz (10%)
- Three mini projects (40%)
- Midterm (20%)
- Final Exam (20%)
- Mini presentation (10%)
- Class participation & Discussion (3% extra credit)

Resources

- updated reference list
- Cloud virtual machines
- Courseweb
  - Submitting reading assignments and projects

Schedule Overview

- Parallel data processing
  - Distributed file systems (GFS, HDFS)
  - MapReduce
  - High-level distributed data management
- Cloud infrastructures
  - Virtualization
  - AWS and Open Stack
  - Mobile cloud computing
- Cloud security and privacy
- Research topics

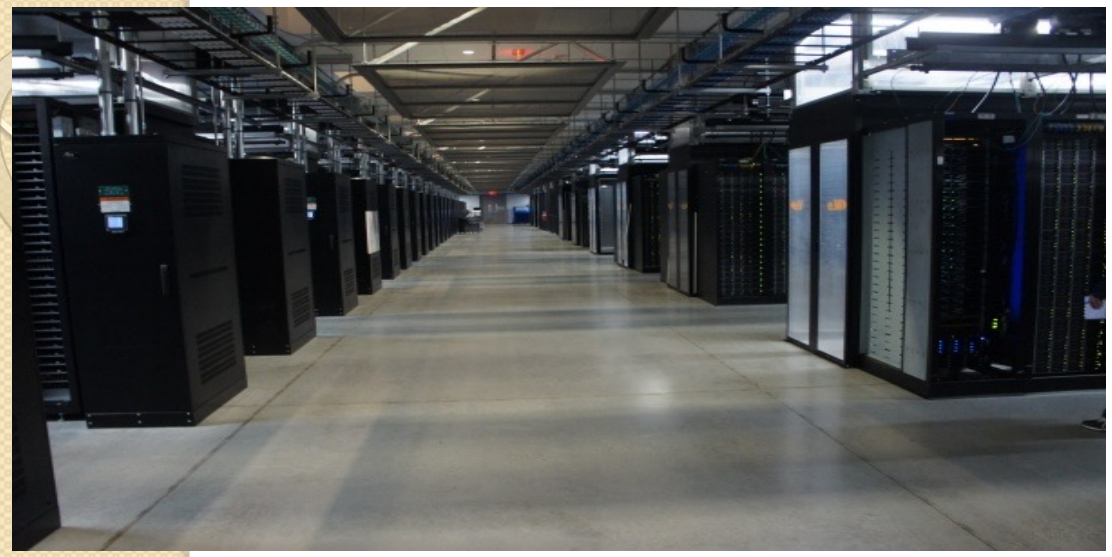
In projects, you will learn

- Hadoop
- Mapreduce, Pig Latin
- Accessing Cloud Virtual Machines and storage
- google app engine

What does a datacenter look like from inside?

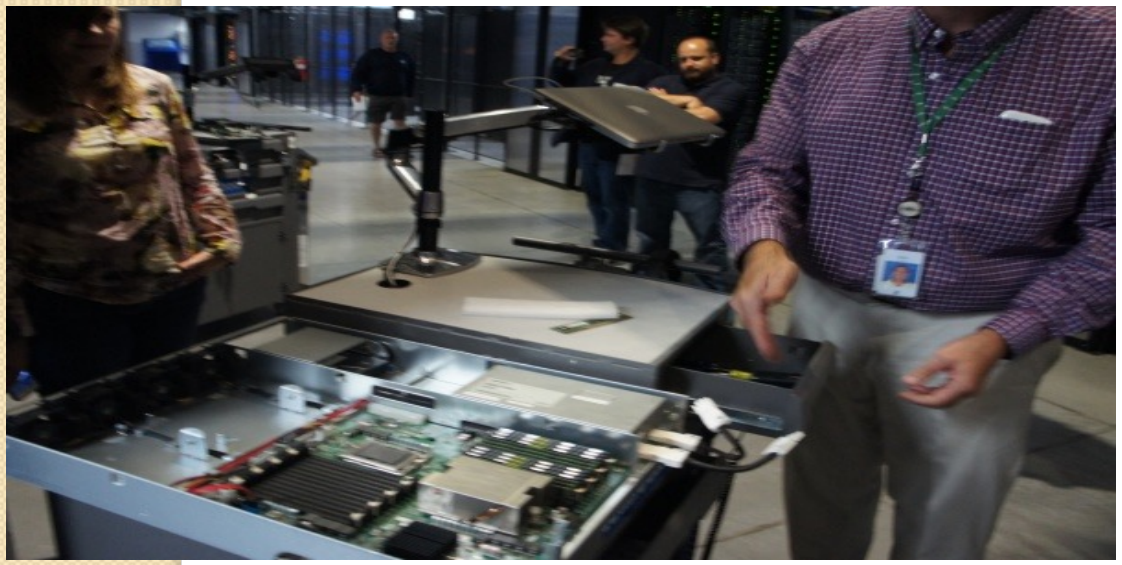
- A virtual walk through a datacenter
- Reference: http://gigaom.com/cleantech/a-rare-look-inside-facebooks-oregon-data-center-photos-video/

Servers



Front

Back



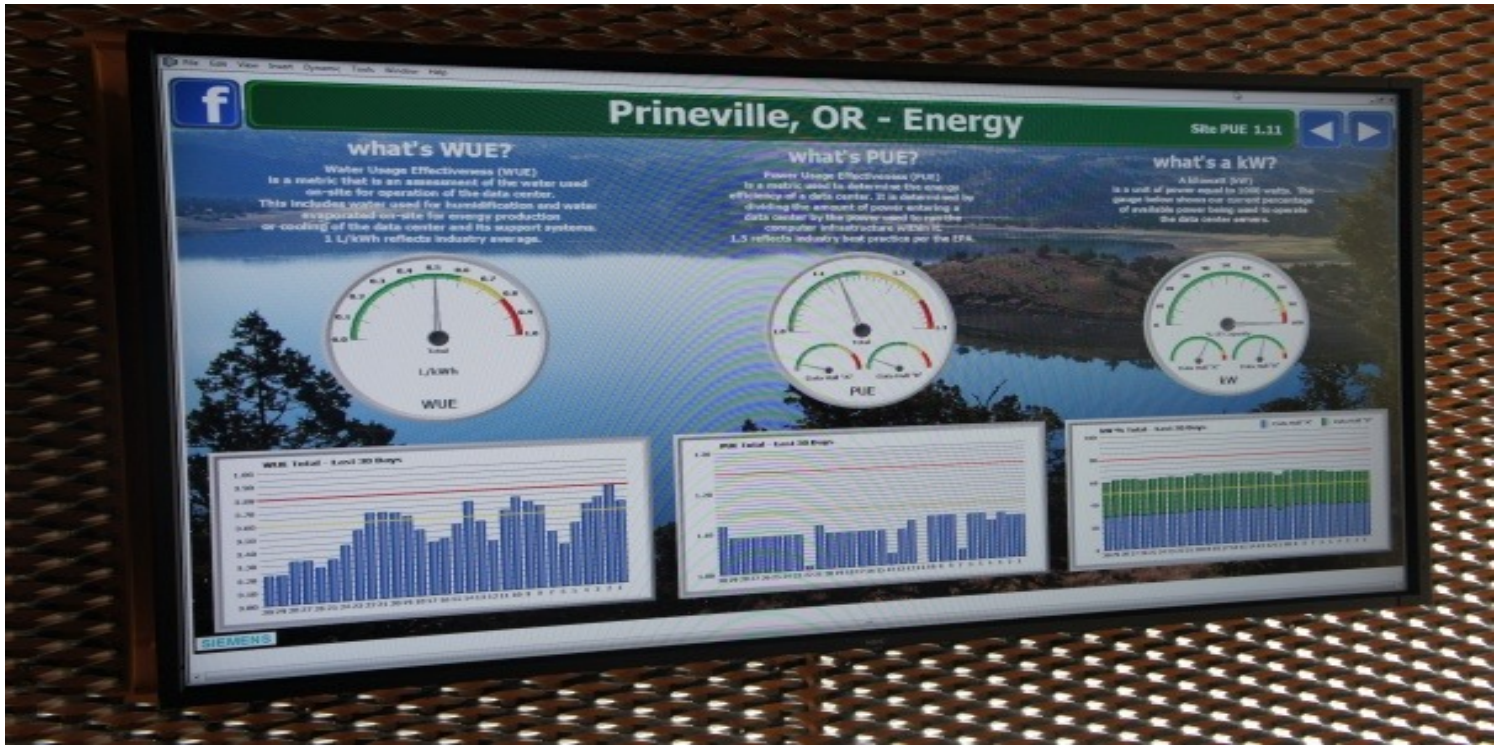
In

Some highly secure (e.g., financial info)

Power

Off-site

On-site

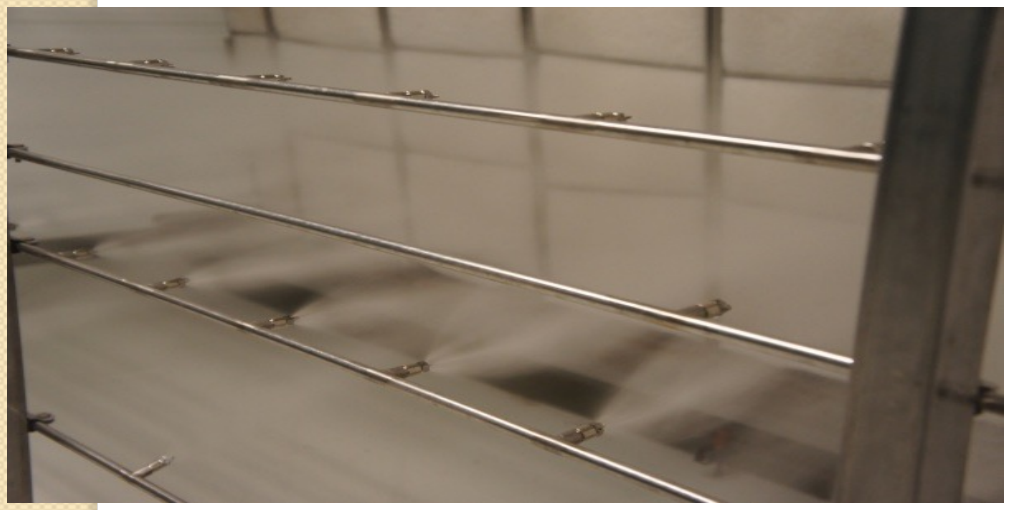


- WUE = Annual Water Usage / IT Equipment Energy (L/kWh) – low is good
- PUE = Total facility Power / IT Equipment Power – low is good (e.g., Google~1.11)

Cooling

Air sucked in from top (also, Bugzappers)

Water purified



Water sprayed into air

15 motors per server bank

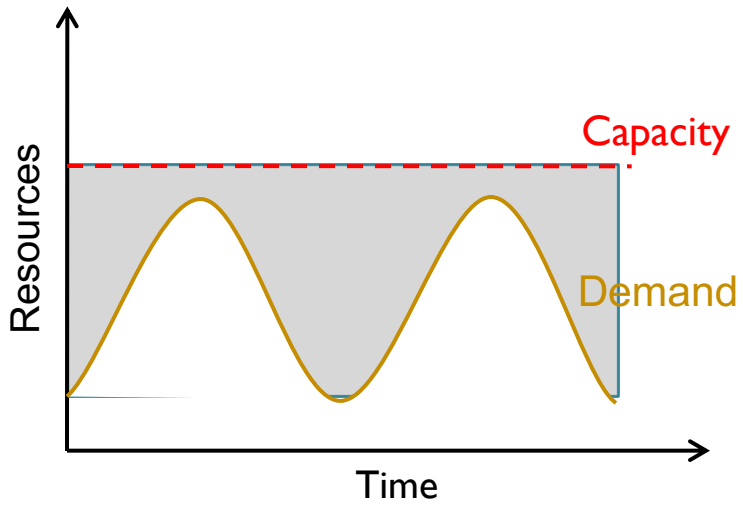
Extra - Fun Videos to Watch

- Microsoft GFS Datacenter Tour (Youtube)
  - http://www.youtube.com/watch?v=hOxA1l1pQIw

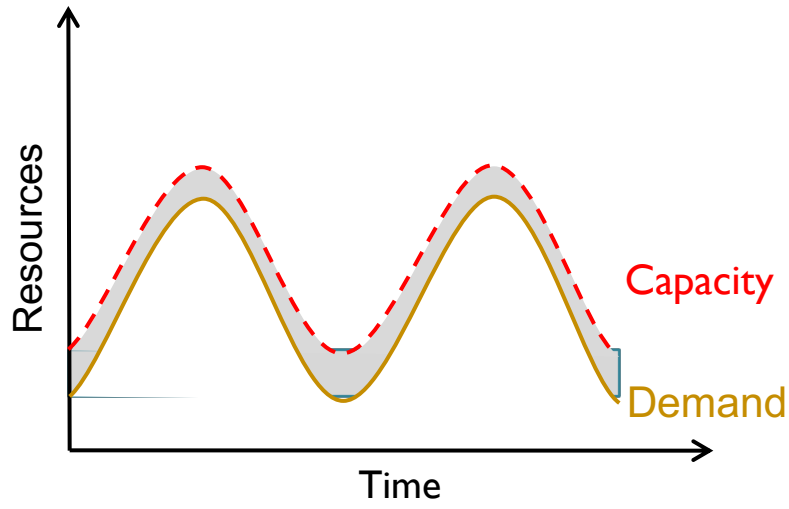
- Timelapse of a Datacenter Construction on the Inside (Fortune 500 company)
  - http://www.youtube.com/watch?v=ujO-xNvXj3g

Economics of Cloud Users

- Pay by use instead of provisioning for peak



Static data center



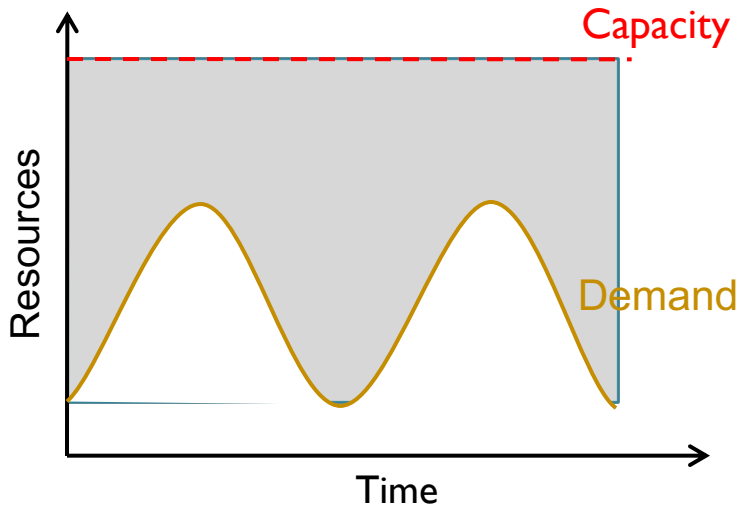
Data center in the cloud

Unused resources

Slide Credits: Berkeley RAD Lab

Economics of Cloud Users

- Risk of over-provisioning: underutilization



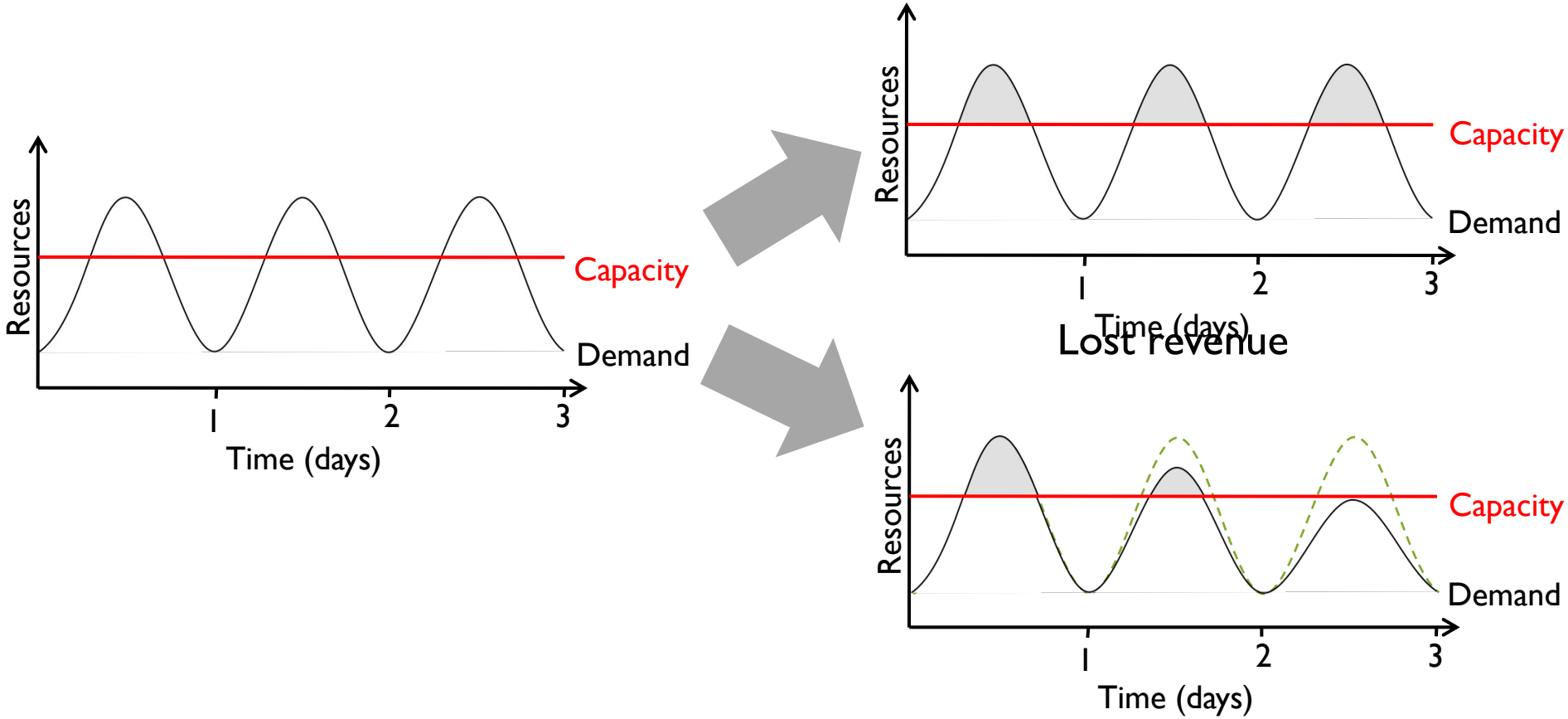
Static data center

Unused resources

Slide Credits: Berkeley RAD Lab

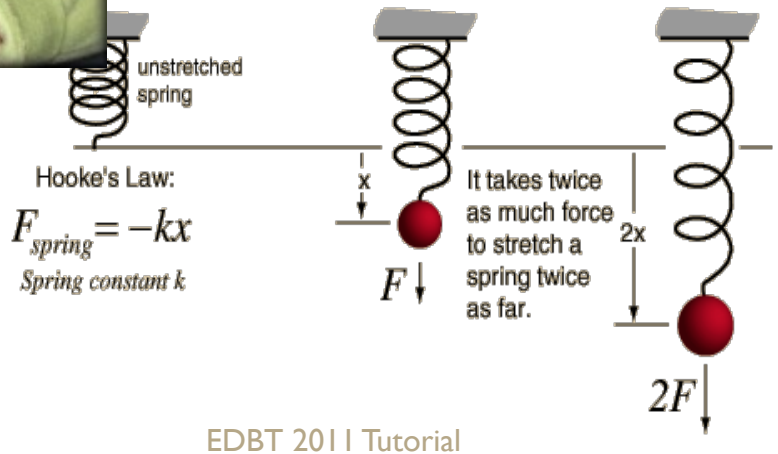
Economics of Cloud Users

- Heavy penalty for under-provisioning



Slide Credits: Berkeley RAD Lab

Cloud Challenges: Elasticity



So, What really is Cloud Computing?

Cloud computing is a new computing paradigm, involving data and/or computation outsourcing, with

- Infinite and elastic resource scalability
- On demand "just-in-time" provisioning
- No upfront cost … pay-as-you-go

That is, use as much or as less you need, use only when you want, and pay only what you use,

Cloud computing provides numerous economic advantages

For clients:

- No upfront commitment in buying/leasing hardware
- Can scale usage according to demand
- Barriers to entry lowered for startups

For providers:

- Increased utilization of datacenter resources

Cloud computing means **selling "X as a service"**

IaaS: Infrastructure as a Service

- Selling virtualized hardware

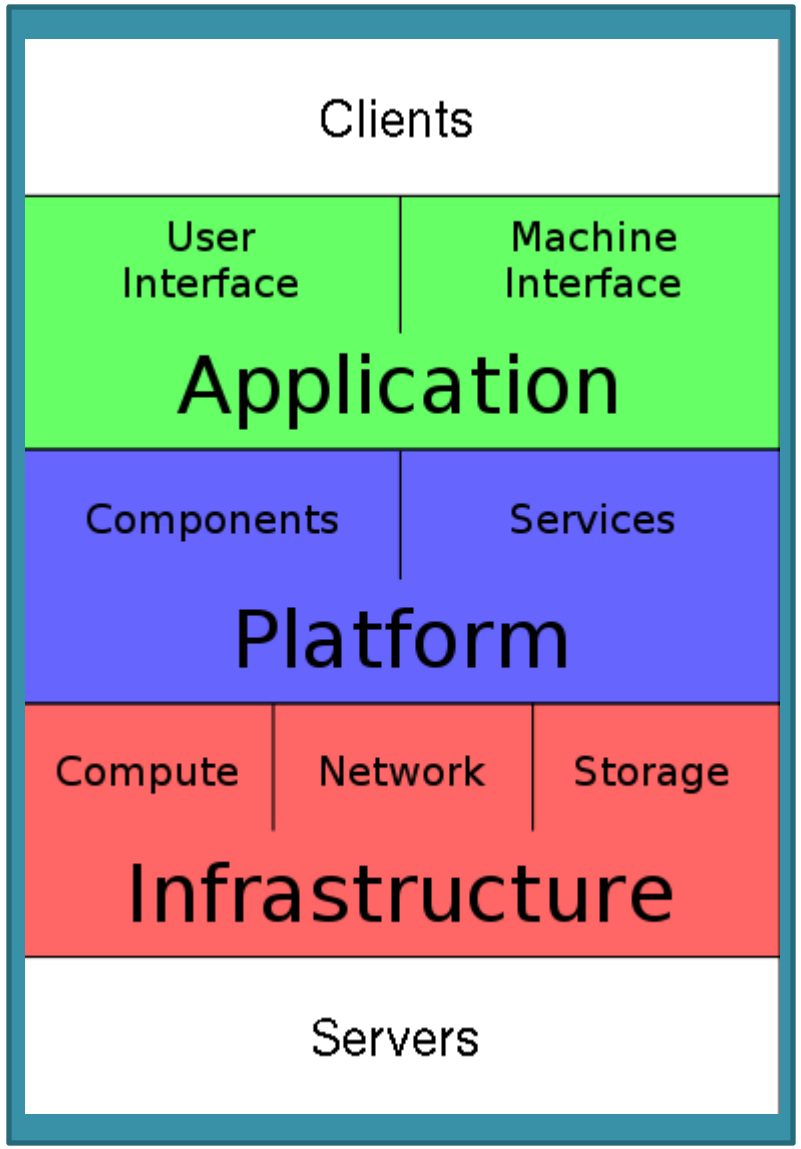
PaaS: Platform as a service

- Access to a configurable platform/API

SaaS: Software as a service

- Software that runs on top of a cloud

Cloud computing architecture



e.g., Web browser

SaaS , e.g., Google Docs

PaaS, e.g., Google AppEngine

IaaS, e.g., Amazon EC2

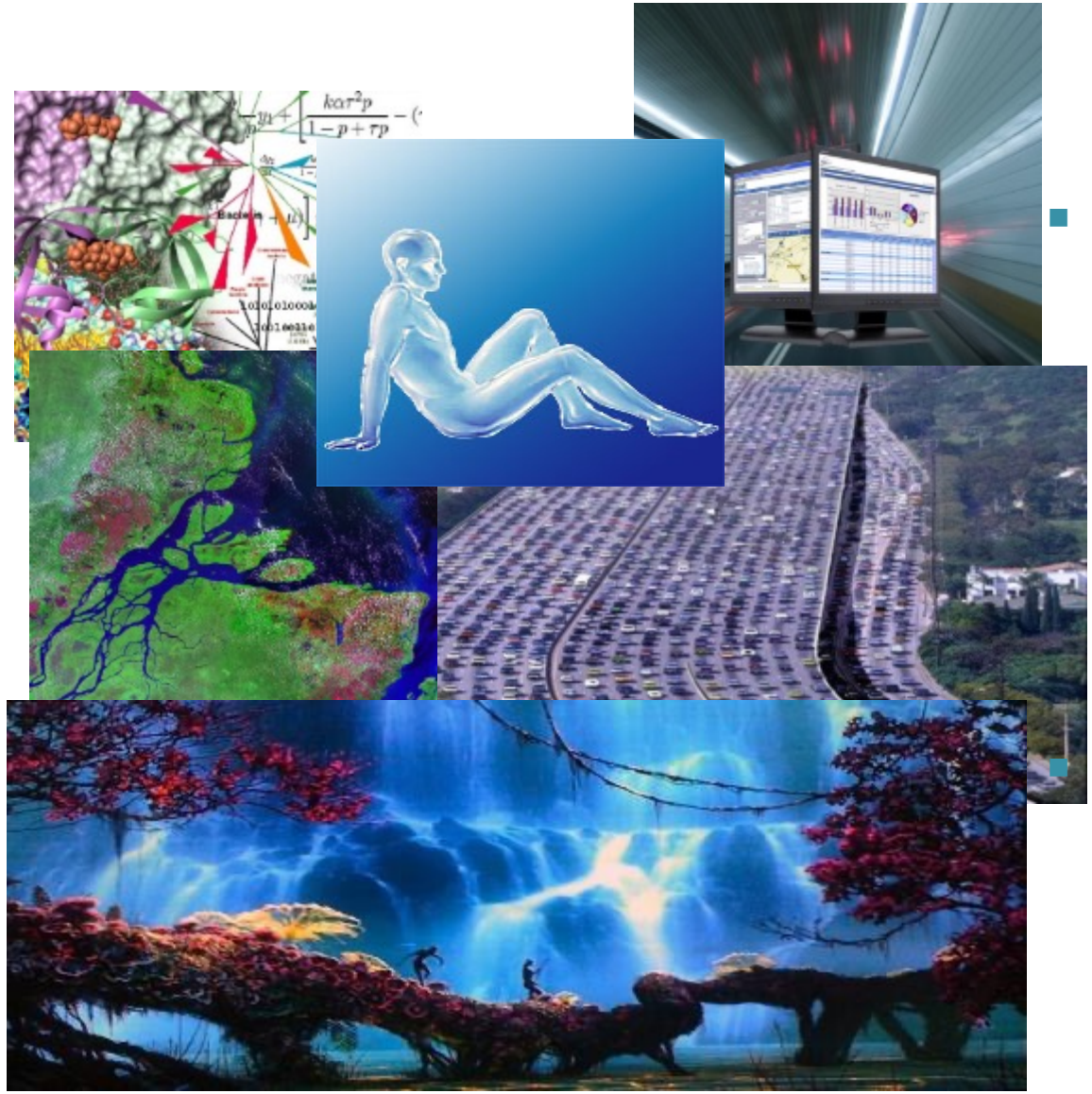
Our Data-driven World

- Science
  - Data bases from astronomy, genomics, environmental data, transportation data, …
- Humanities and Social Sciences
  - Scanned books, historical documents, social interactions data, …
- Business & Commerce
  - Corporate sales, stock market transactions, census, airline traffic, …
- Entertainment
  - Internet images, Hollywood movies, MP3 files, …
- Medicine
  - MRI & CT scans, patient records, …

Data-rich World

- Data capture and collection:
  - Highly instrumented environment
  - Sensors and Smart Devices
  - Network
- Data storage:
  - Seagate 1 TB Barracuda @ $72.95 from Amazon.com (73¢/GB)

What can we do with this wealth?



- What can we do?
  - Scientific breakthroughs
  - Business process efficiencies
  - Realistic special effects
  - Improve quality-of-life: healthcare, transportation, environmental disasters, daily life, …
- Could We Do More?
  - YES: but need major advances in our capability to analyze this data

Cloud Computing Modalities

"Can we outsource our IT software and hardware infrastructure?"

- Hosted Applications and services
- Pay-as-you-go model
- Scalability, fault-tolerance, elasticity, and self-manageability

"We have terabytes of click-stream data – what can we do with it?"

- Very large data repositories
- Complex analysis
- Distributed and parallel data processing

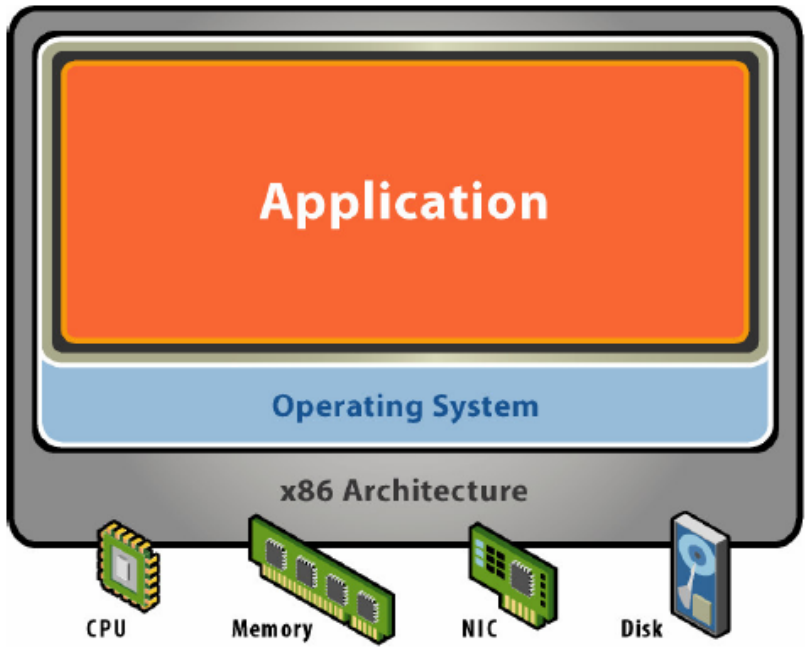
Cloud is often a collection of a large number of virtual devices such as VMs and virtualized disks

Virtualization

- Virtualization deals with "extending or replacing an existing interface so as to mimic the behavior of another system"

- Virtual system examples: virtual private network, virtual memory, virtual machine

Starting Point: A Physical Machine



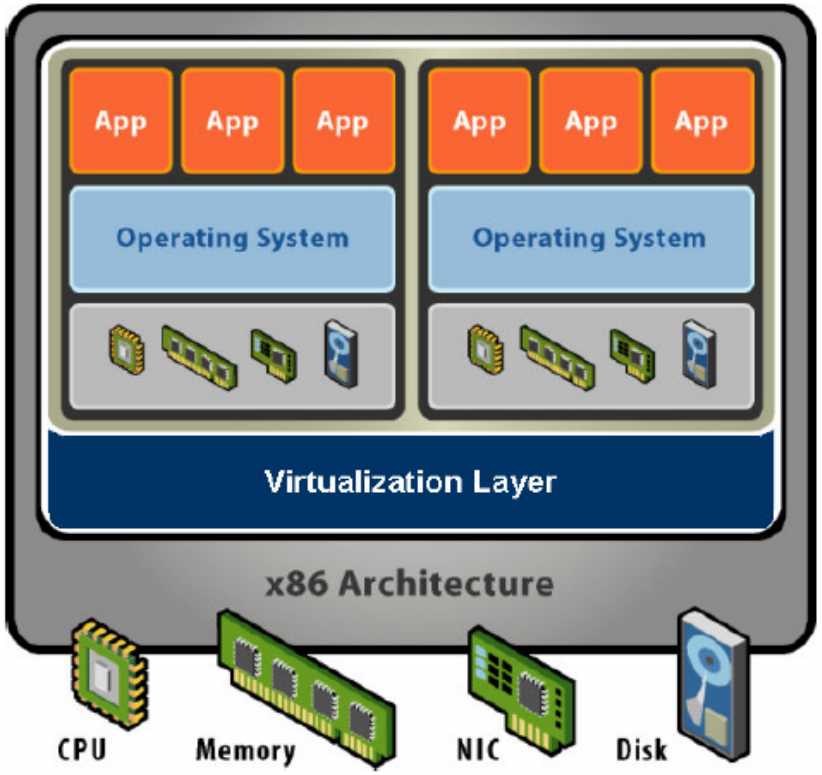
Physical Hardware

- Processors, memory, chipset, I/O bus and devices, etc.
- Physical resources often underutilized

Software

- Tightly coupled to hardware
- Single active OS image
- OS controls hardware

What is a Virtual Machine?



Hardware-Level Abstraction

- Virtual hardware: processors, memory, chipset, I/O devices, etc.
- Encapsulates all OS and application state

Virtualization Software

- Extra level of indirection decouples hardware and OS
- Multiplexes physical hardware across multiple “guest” VMs
- Strong isolation between VMs
- Manages physical resources, improves utilization

VM Isolation

Secure Multiplexing

- Run multiple VMs on single physical host
- Processor hardware isolates VMs, *e.g.* MMU

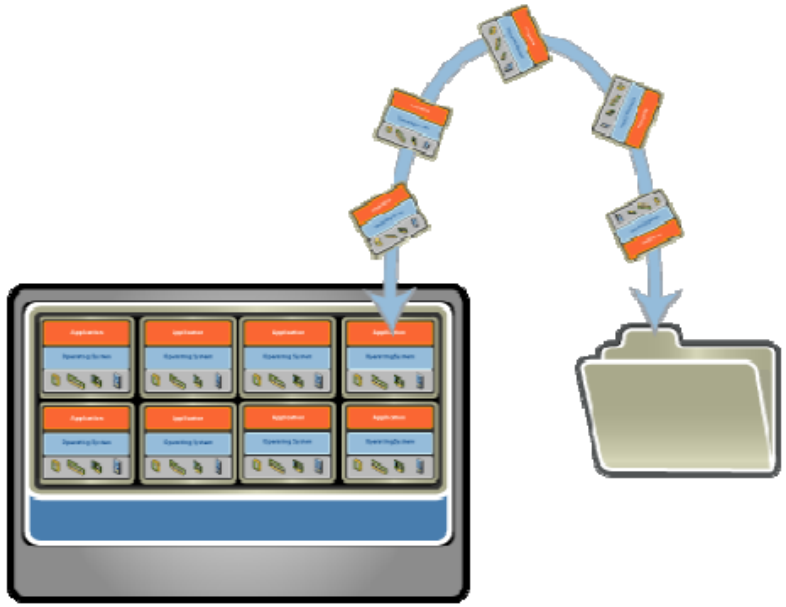
Strong Guarantees

- Software bugs, crashes, viruses within one VM cannot affect other VMs

Performance Isolation

- Partition system resources
- Example: VMware controls for reservation, limit, shares

VM Encapsulation



Entire VM is a File

- OS, applications, data
- Memory and device state

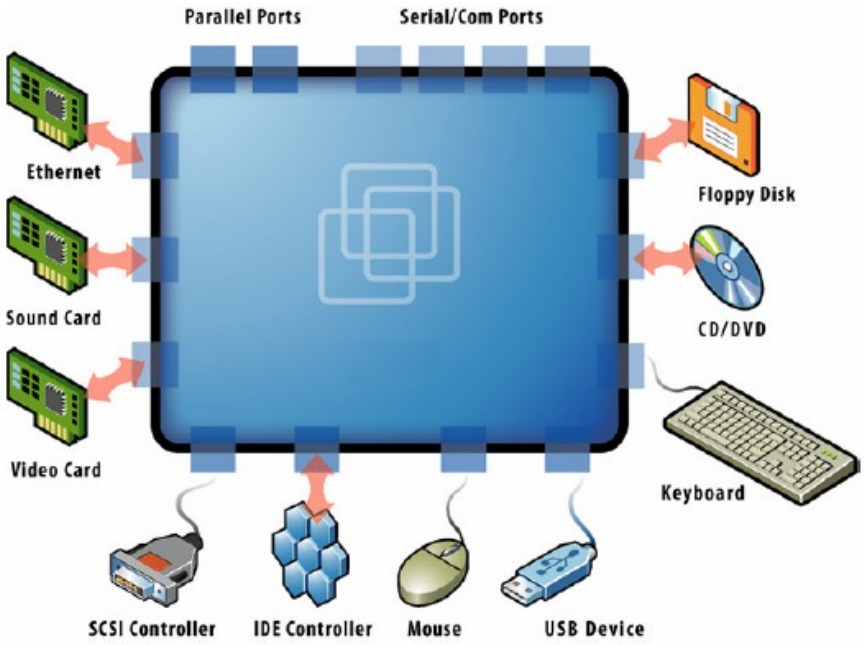
Snapshots and Clones

- Capture VM state on the fly and restore to point-in-time
- Rapid system provisioning, backup, remote mirroring

Easy Content Distribution

- Pre-configured apps, demos
- Virtual appliances

VM Compatibility



Hardware-Independent

- Physical hardware hidden by virtualization layer
- Standard virtual hardware exposed to VM

Create Once, Run Anywhere

- No configuration issues
- Migrate VMs between hosts

Legacy VMs

- Run ancient OS on new platform
- *E.g.* DOS VM drives virtual IDE and vLance devices, mapped to modern SAN and GigE hardware

Common Virtualization Uses Today

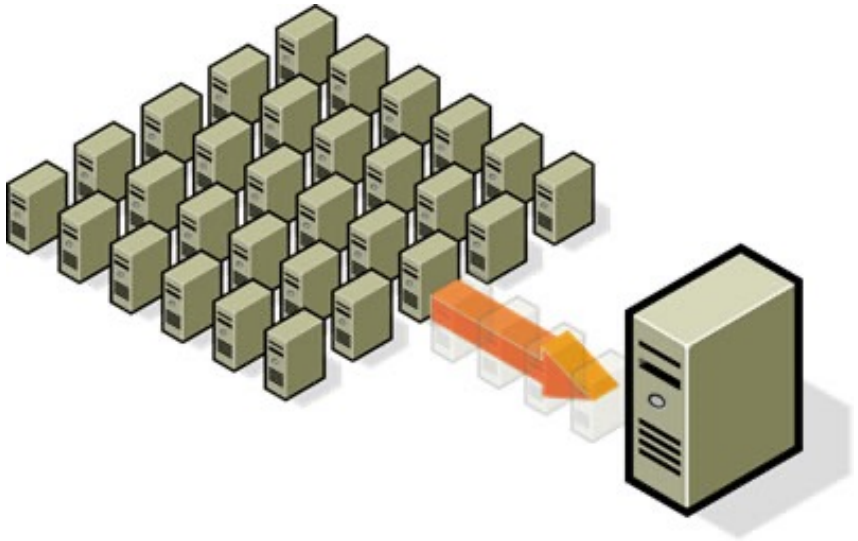
Test and Development – Rapidly provision test and development servers; store libraries of pre-configured test machines

Business Continuity – Reduce cost and complexity by encapsulating entire systems into single files that can be replicated and restored onto any target server

Enterprise Desktop – Secure unmanaged PCs without compromising end-user autonomy by layering a security policy in software around desktop virtual machines

Common Virtualization Uses Today

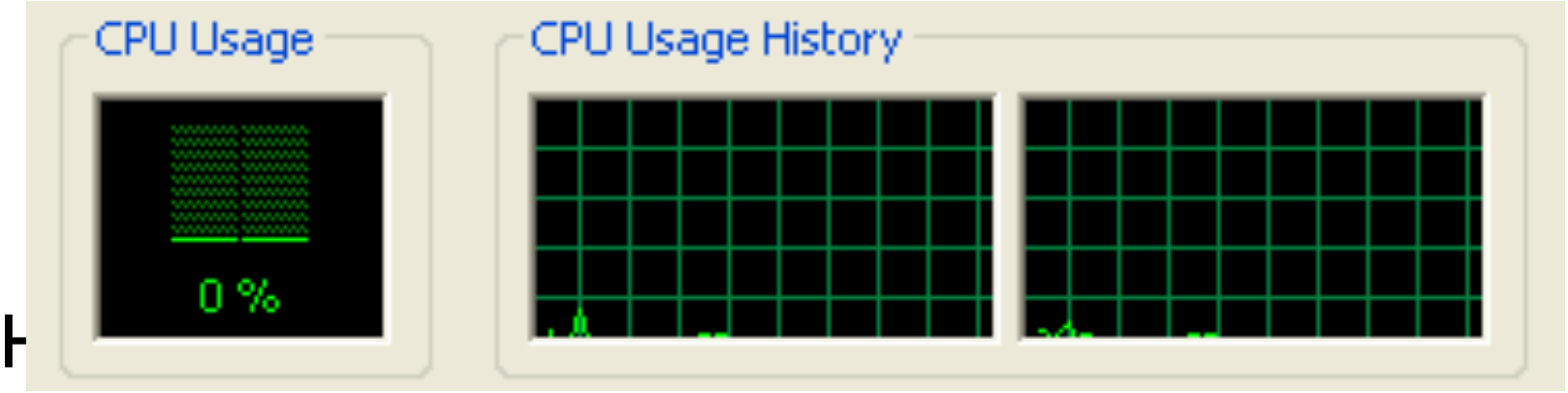
- Reduce costs by consolidating services onto the fewest number of physical machines



http://www.vmware.com/img/serverconsolidation.jpg

Non-virtualized Data Centers

- Too many servers for too little work

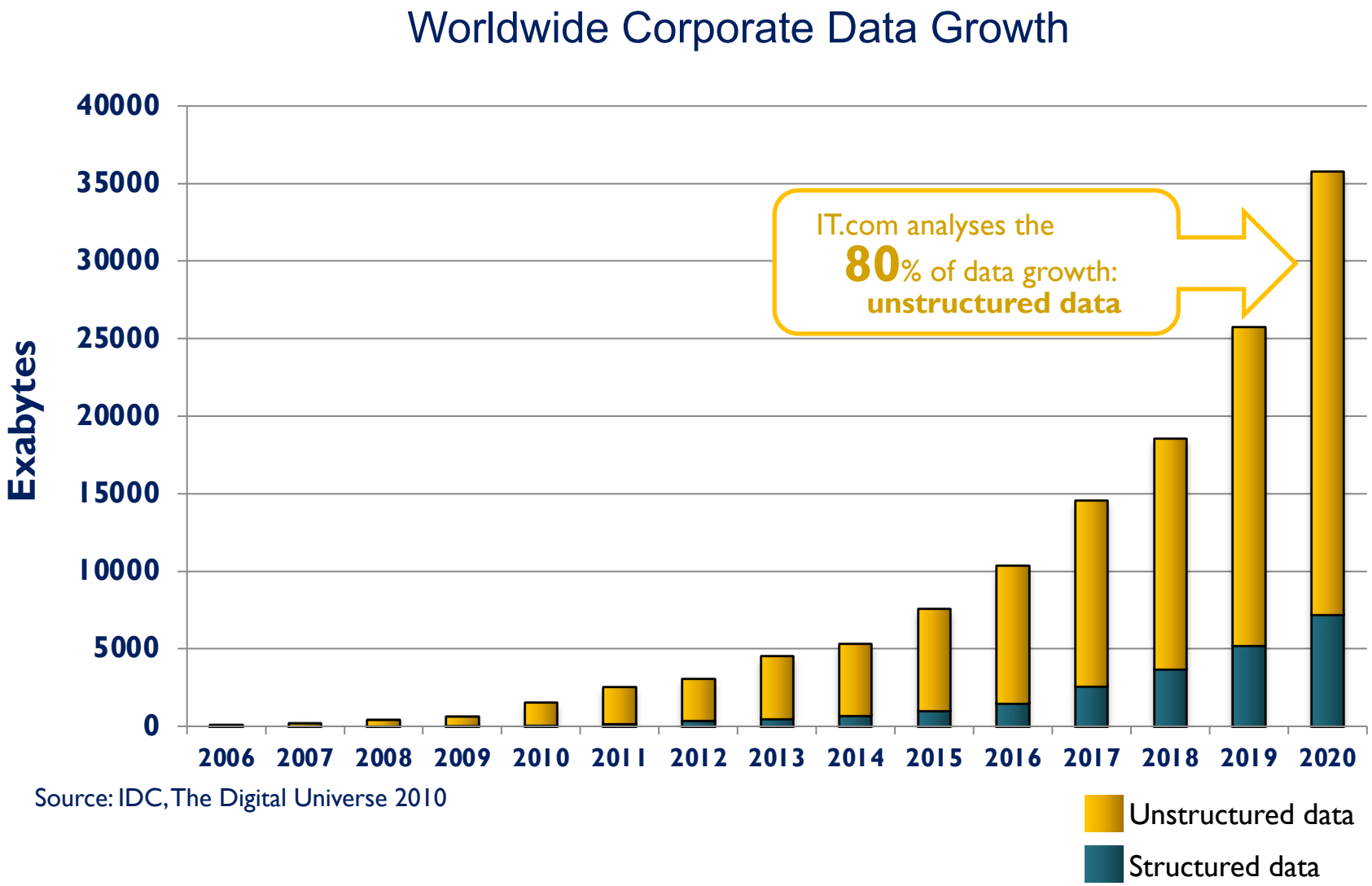


- H
  - Maintenance
  - Networking
  - Floor space
  - Cooling
  - Power
  - Disaster Recovery

Dynamic Data Center

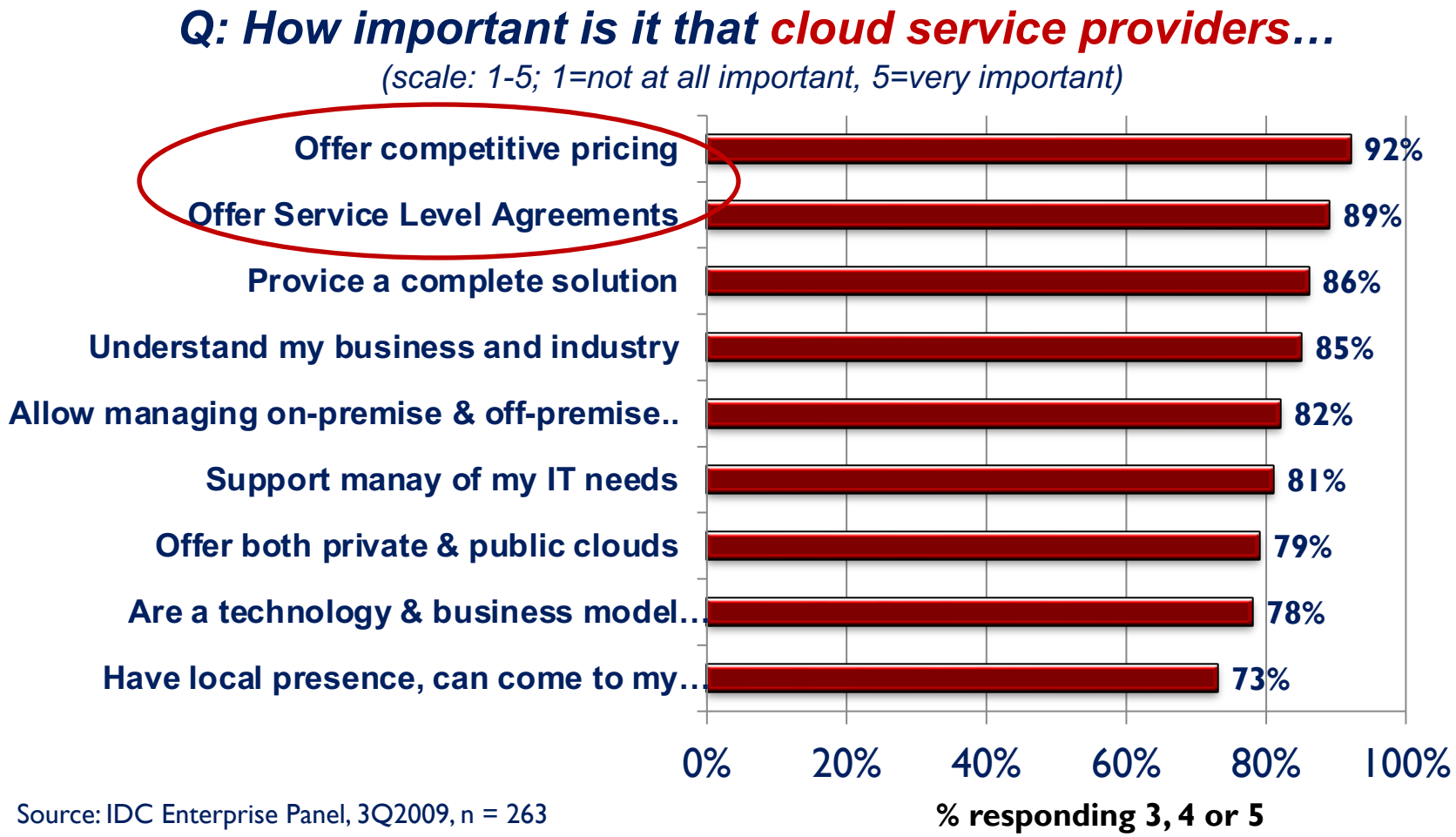
- Virtualization helps us break the "one service per server" model
- Consolidate many services into a fewer number of machines when workload is low, reducing costs
- Conversely, as demand for a particular service increases, we can shift more virtual machines to run that service
- We can build a data center with fewer total resources, since resources are used as needed instead of being dedicated to single services

Data Growth



Source: IDC, The Digital Universe 2010

Cloud Computing Challenges

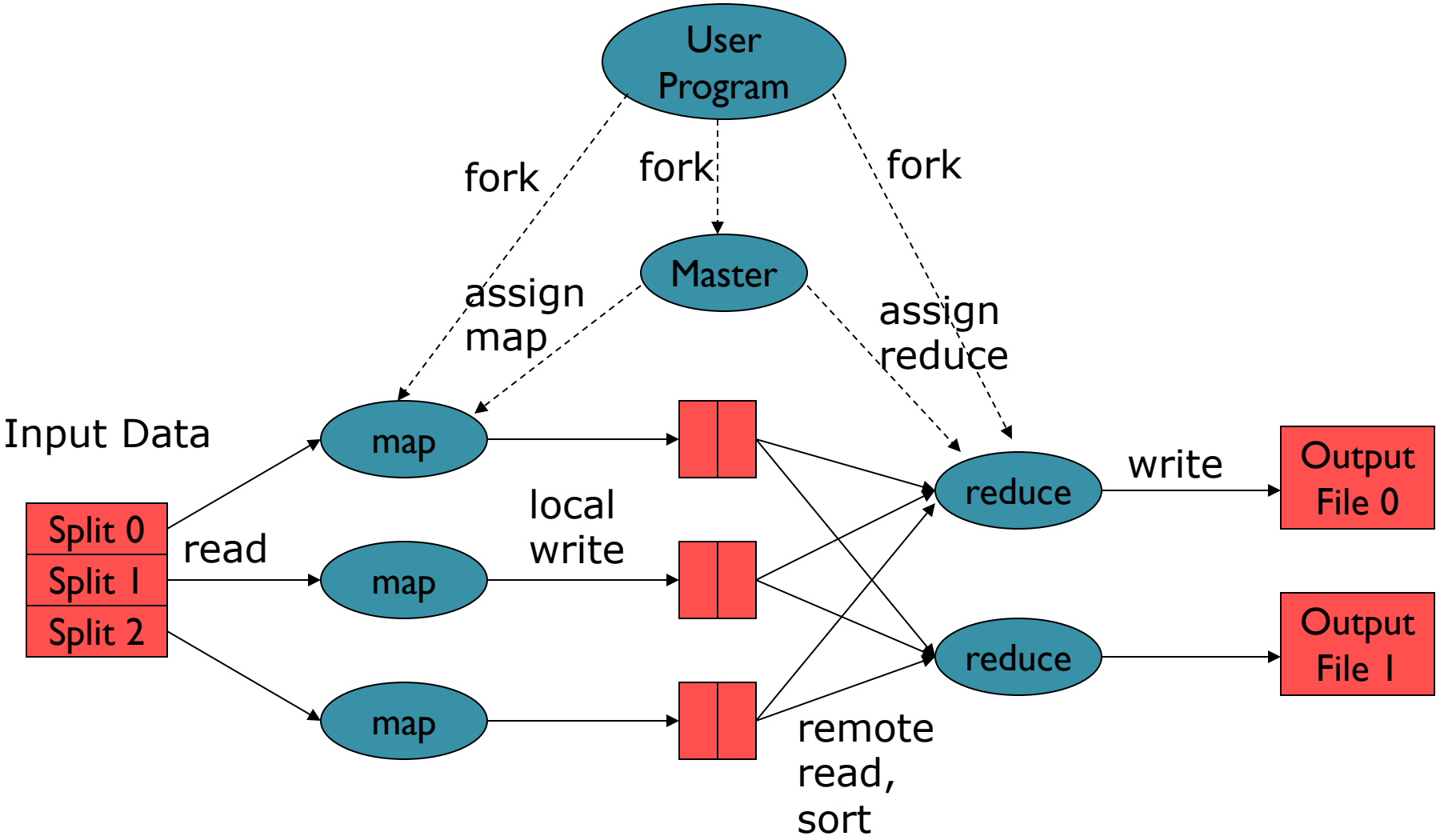


Source: IDC Enterprise Panel, 3Q2009, n = 263

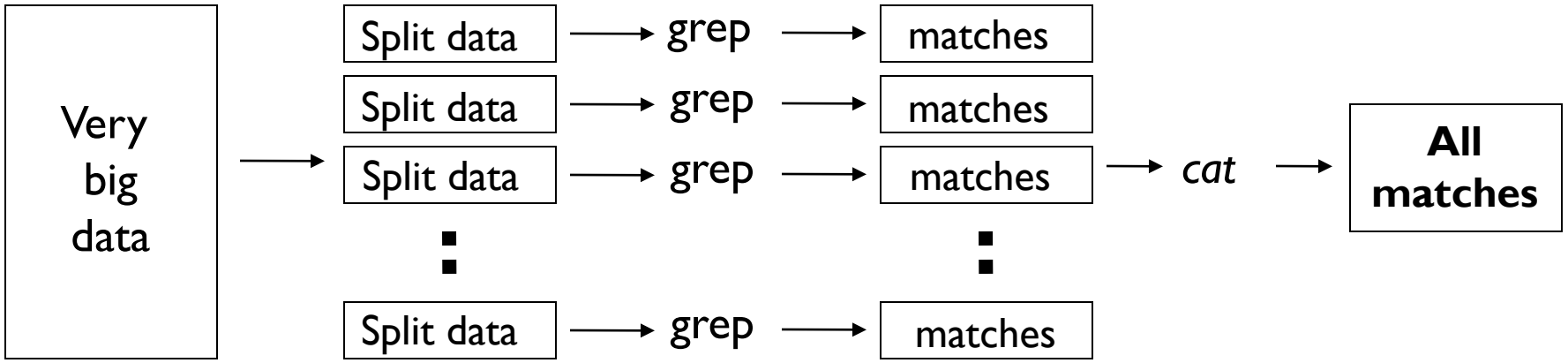
Big Data processing in a Cloud

- MapReduce and Big Data Processing
  - Programming model for data-intensive computing on commodity clusters
  - Pioneered by Google
    - Processes 20 PB of data per day
  - **Scalability** to large data volumes
    - Scan 100 TB on 1 node @ 50 MB/s = 24 days
    - Scan on 1000-node cluster = 35 minutes
  - It is expected that more than half the world's data will be processed by Hadoop – Hortonworks

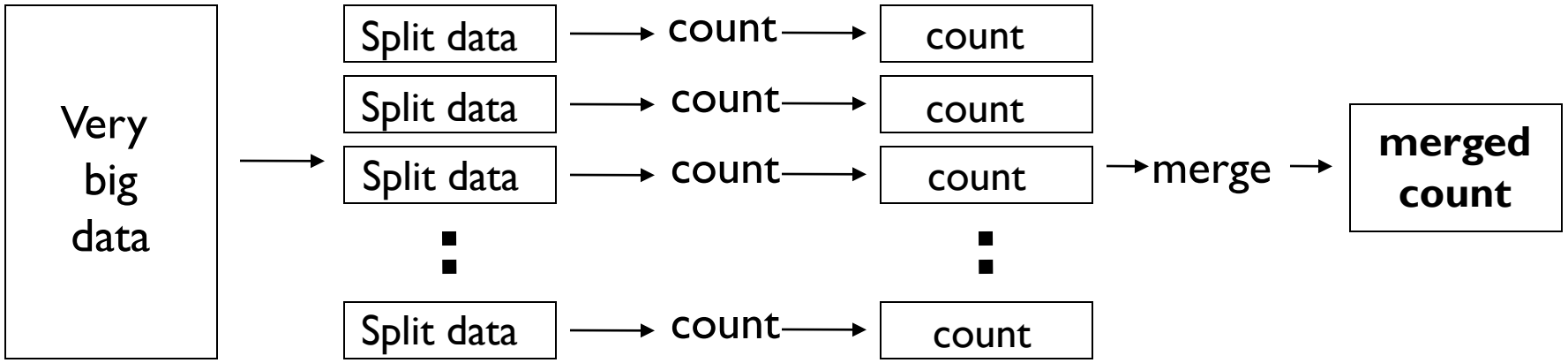
MapReduce Execution Overview



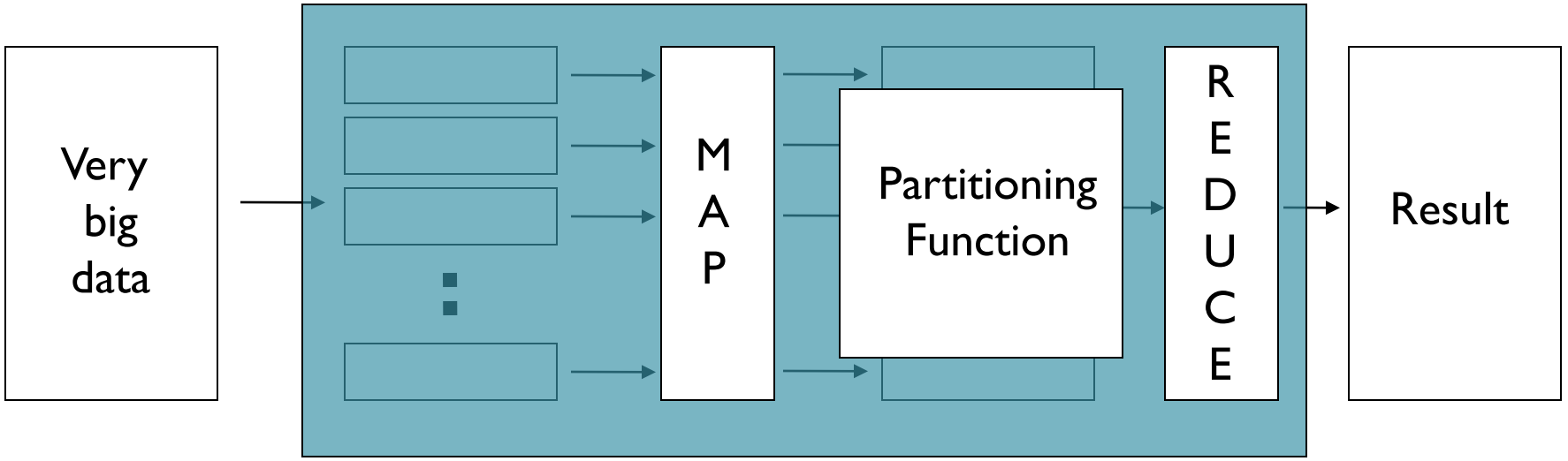
Distributed Grep



Distributed Word Count

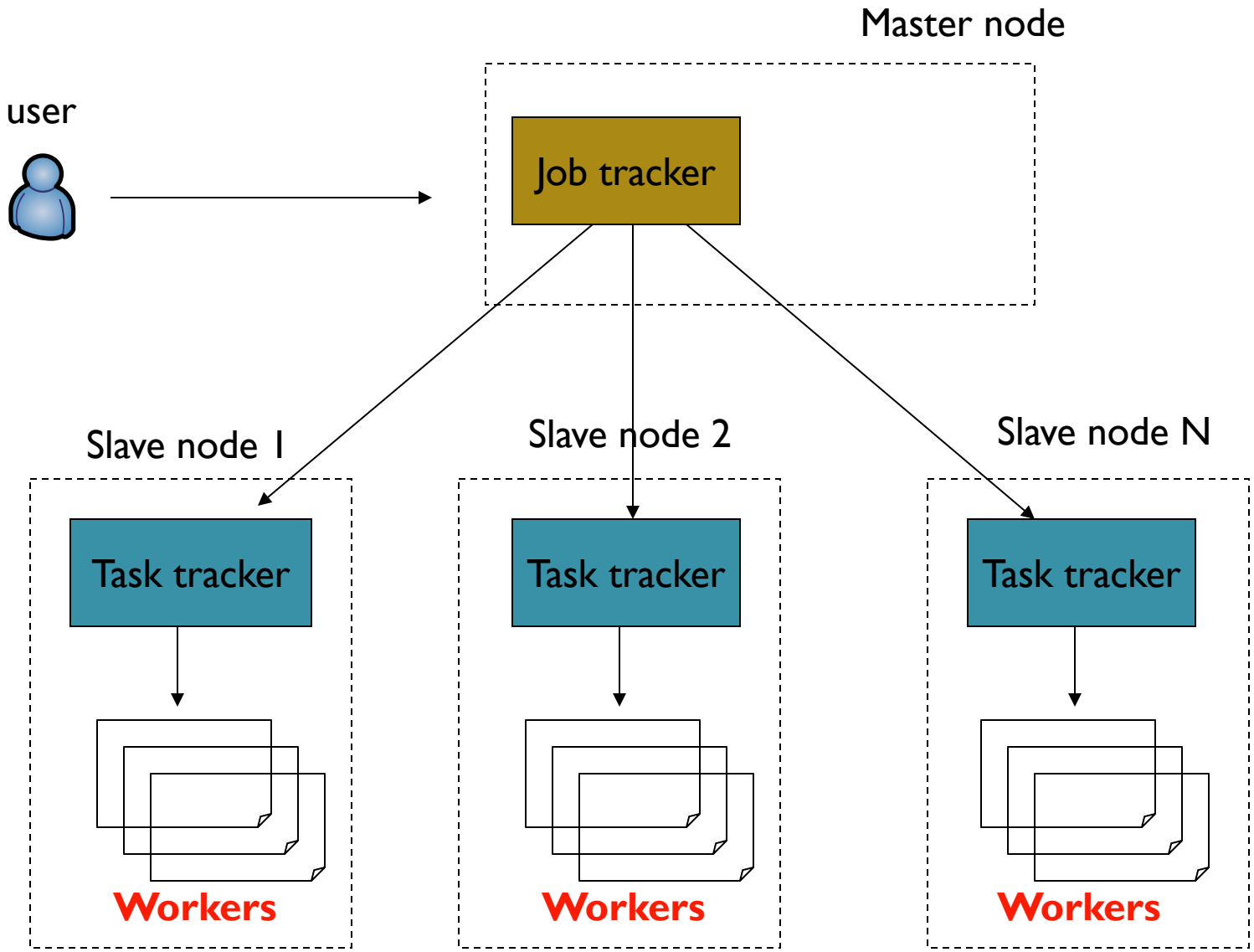


Map+Reduce



- Map:
  - Accepts *input* key/value pair
  - Emits *intermediate* key/value pair
- Reduce :
  - Accepts *intermediate* key/value* pair
  - Emits *output* key/value pair

Architecture overview



Example: Count word occurrences of each word in a large collection of documents

```
map(String input_key, String input_value):
  // input_key: document name
  // input_value: document contents
  for each word w in input_value:
    EmitIntermediate(w, "1");


reduce(String output_key, Iterator
  intermediate_values):
  // output_key: a word
  // output_values: a list of counts
  int result = 0;
  for each v in intermediate_values:
    result += ParseInt(v);
 Emit(AsString(result));
```

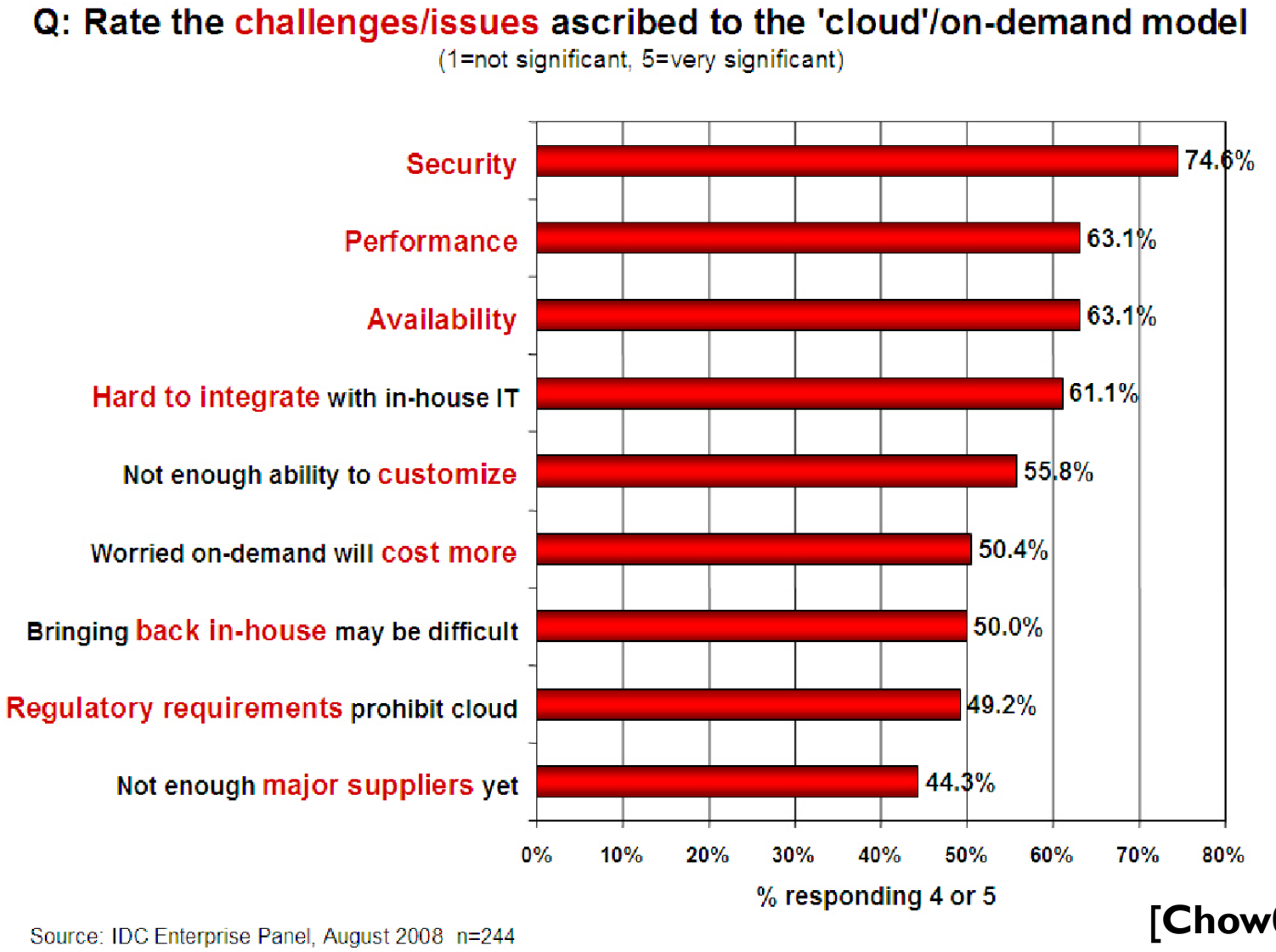
Open Source Implementation

- Apache Hadoop – open source Map/reduce software platform, automatically provides framework for developing map/reduce applications,
- Handles mapping and reducing logistics, programmer just provides custom functionality,
- http://lucene.apache.org/hadoop,

So, if cloud computing is so great, why aren't everyone doing it?

Clouds are still subject to traditional data confidentiality, integrity, availability, and privacy issues, plus some additional attacks

Companies are still **afraid** to use clouds



[Chow09ccsw]

Anatomy of **fear** …

Confidentiality

- Will the sensitive data stored on a cloud remain confidential? Will cloud compromises leak confidential client data (i.e., fear of loss of control over data)

- Will the cloud provider itself be honest and won't peek into the data?

Anatomy of **fear** …

Integrity

- How do I know that the cloud provider is doing the computations correctly?
- How do I ensure that the cloud provider really stored my data without tampering with it?

Anatomy of **fear** …

Availability

- Will critical systems go down at the client, if the provider is attacked in a Denial of Service attack?
- What happens if cloud provider goes out of business?

Anatomy of **fear** …

Privacy issues raised via massive data mining

- Cloud now stores data from a lot of clients, and can run data mining algorithms to get large amounts of information on clients

Anatomy of **fear** …

Increased attack surface

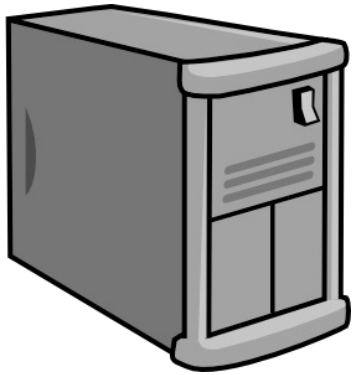
- Entity outside the organization now stores and computes data, and so
- Attackers can now target the communication link between cloud provider and client
- Cloud provider employees can be phished

Anatomy of **fear** …

Auditability and forensics

- Difficult to audit data held outside organization in a cloud
- Forensics also made difficult since now clients don't maintain data locally

Traditional systems security vs Cloud Computing Security

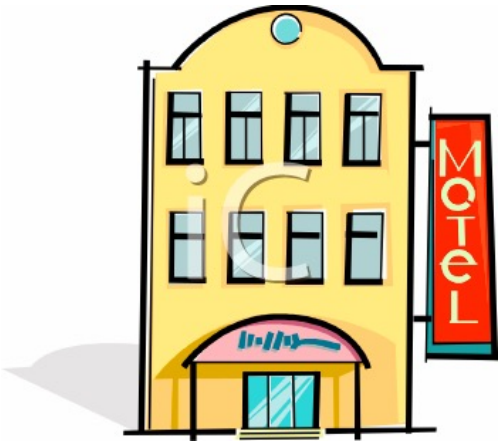


Securing a traditional system

Securing a cloud

Traditional systems security vs Cloud Computing Security

Analogy



Securing a house

Owner and user are often the same entity

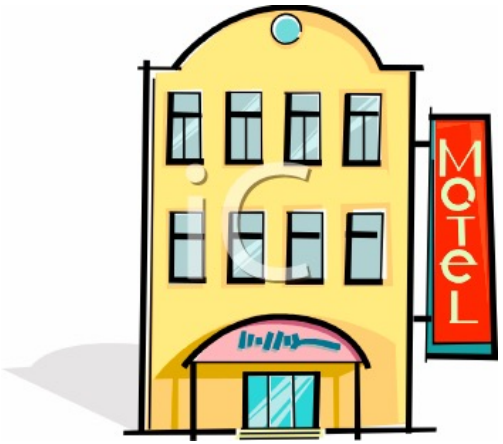
Securing a motel

Owner and users are almost invariably distinct entities

Traditional systems security vs Cloud Computing Security

Securing a house

Biggest user concerns
Securing perimeter
Checking for intruders
Securing assets



Securing a motel

Biggest user concern
Securing room against (the bad guy in next room | hotel owner)

- Next class
  - Introduction to Datacenters and Datacenter File systems