An Introduction to Cloud Computing

Nilanjan Banerjee

Courtesy: Prof. Armando Fox @ Berkeley

Overview

- What is distributed computing?
- What is warehouse-scale computing?
- What is cloud computing?
- Why should you care?
- What are the challenges?

What is Distributed Computing?

- Distributed computing is a method of computer processing in which different parts of a program run simultaneously on two or more computers that are communicating with each other over a network.

Big Computers c. 1996

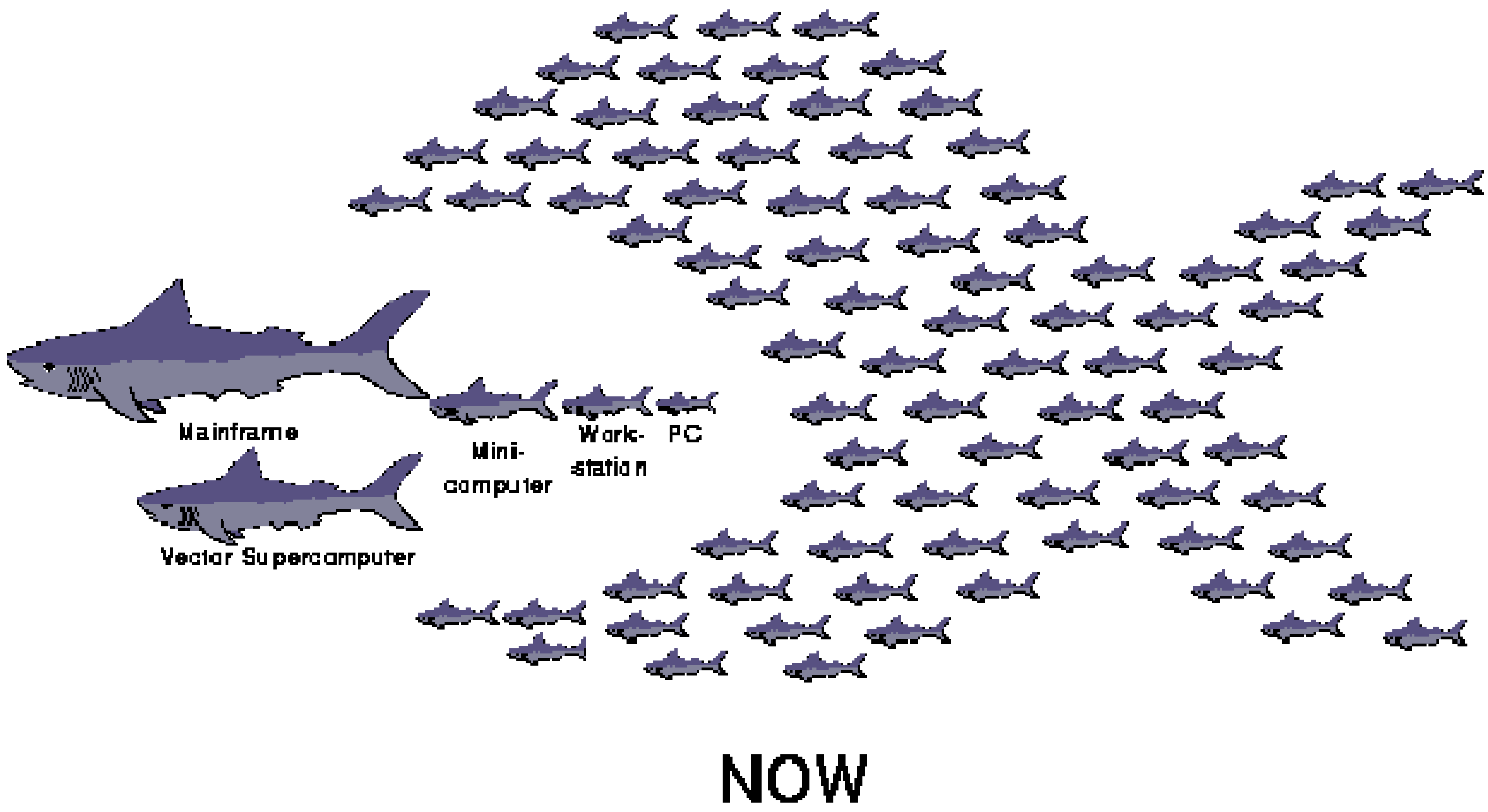
Sun E-10000 “supermini”

- Up to 64 processors @250MHz
- Up to 64 GB RAM
- Up to 20 TB Disk
- Used by eBay, among others

PC

- 200 MHz CPU, 32MB RAM, 4 GB disk

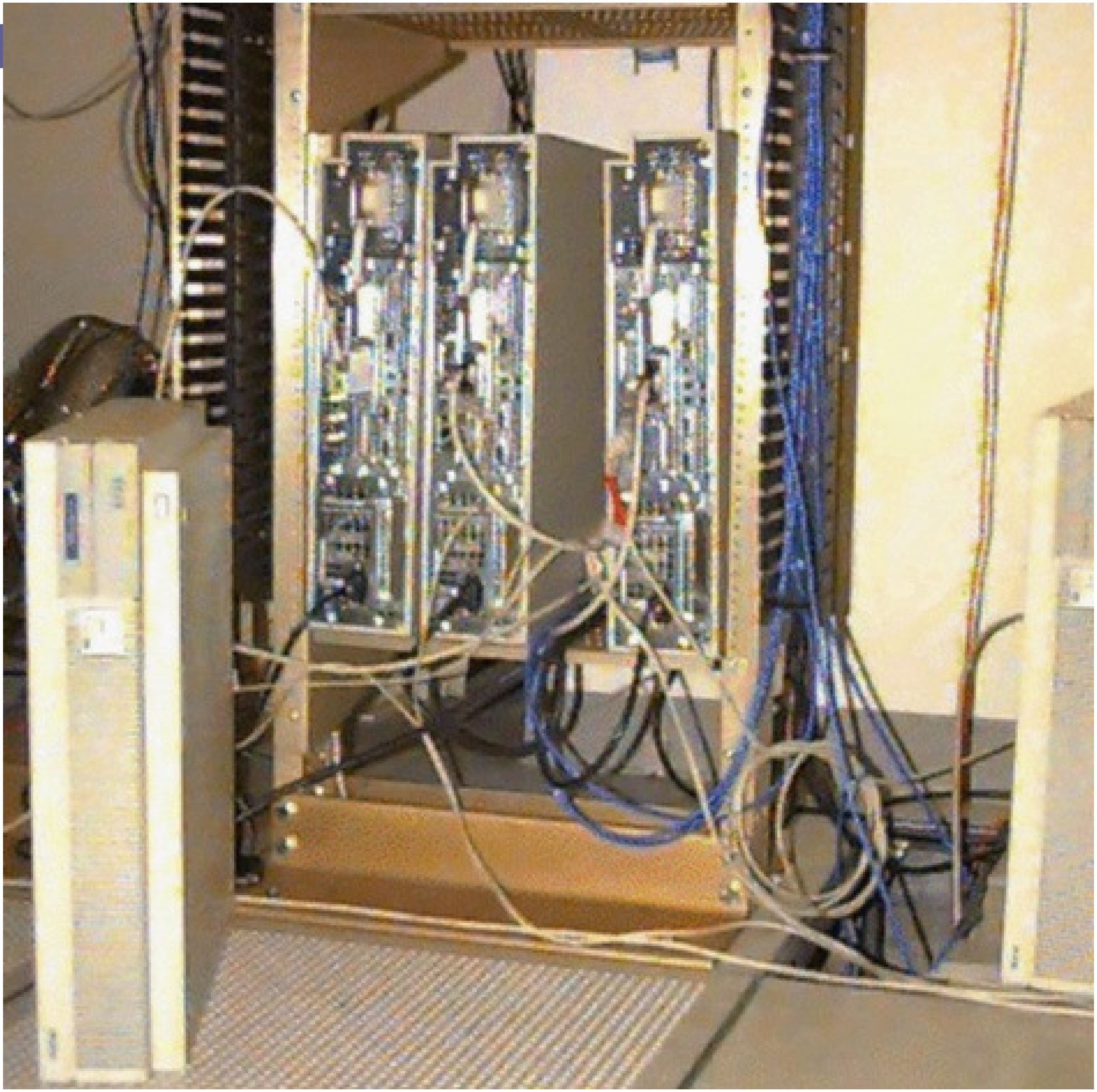
UC Berkeley *Networks Of Workstations* (1994-1999)



NOW-0

1994

Four HP-735's



NOW-1

1995

32 Sun SPARC-stations

NOW-2

1997

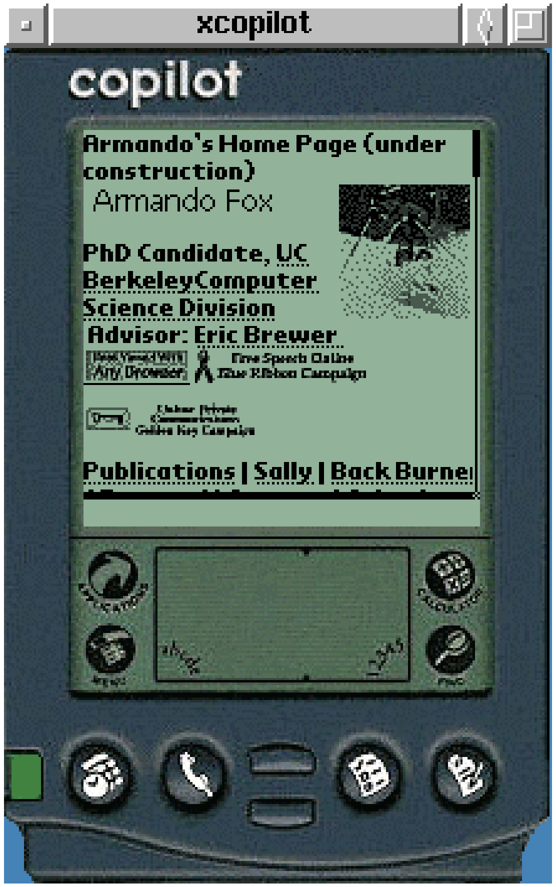
60 Sun SPARC-2

Challenge: how do you program a NOW? (or: what is it good for?)

Access Is the Killer App!

UC Berkeley, 1994-1999

- Project Daedalus: Profs. Katz & Brewer
- Data, services in ~~infrastructure~~ cloud
  - search, email, personal comms, productivity...
- Mobile access *anywhere, anytime*
- Many “firsts”:
  - server architecture with auto-scaling
  - cluster-based Internet service: Inktomi
  - mobile Web: TopGun Wingman on Palm


xcopilot
copilot
Best Viewed With Any Browser
Free Speech Online Blue Ribbon Campaign
Golden Key Campaign

Armando's Home Page (under construction)

Armando Fox

PhD Candidate, UC BerkeleyComputer Science Division

Advisor: Eric Brewer

Publications | Sally | Back Burne

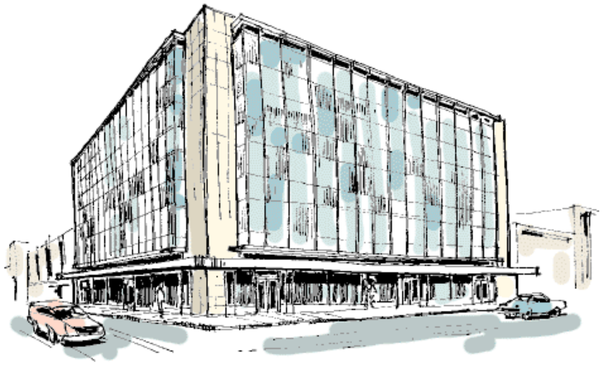
Challenge: deploying the service!

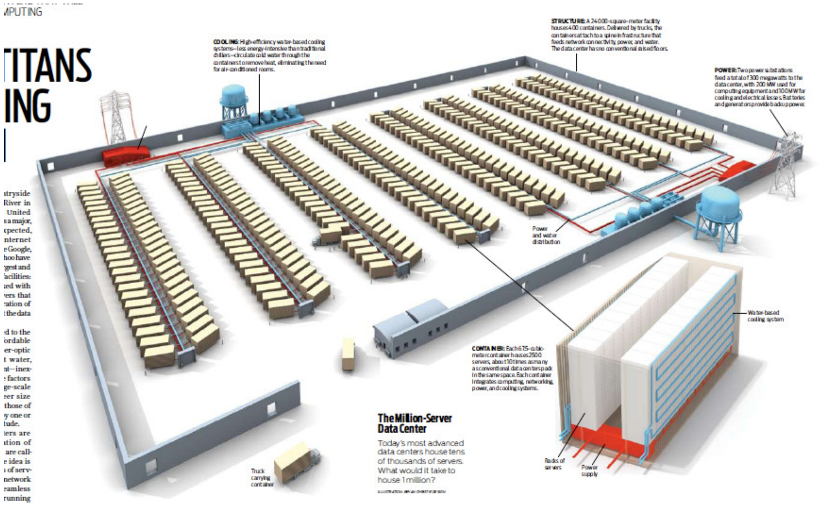
The Killer App for NOWs

- Prof. Eric Brewer, Armando Fox, Steve Gribble, Paul Gauthier, Yatin Chawathe: *Cluster-Based Scalable Network Servers* in Symposium on Operating Systems Principles, 1997
- *Non*-goal: build best/fastest search engine
  - But led to Inktomi, first *truly scalable* search engine that took advantage of NOW ideas
- Goal: show general techniques for programming NOW's for Internet services

2005: *Datacenter* is new "server"

- *"Program"* => Web search, email, map/GIS, …
- *"Computer"* => 1000's computers, storage, network
- Warehouse-sized facilities and workloads





COOLING: High-efficiency water-based cooling systems—less energy-intensive than traditional chillers—circulate cold water through the containers to remove heat, eliminating the need for air-conditioned rooms.

STRUCTURE: A 24 000-square-meter facility houses 400 containers. Delivered by trucks, the containers attach to a spine infrastructure that feeds network connectivity, power, and water. The data center has no conventional raised floors.

POWER: Two power substations feed a total of 300 megawatts to the data center, with 200 MW used for computing equipment and 100 MW for cooling and electrical losses. Batteries and generators provide backup power.

CONTAINER: Each 67.5-cubic-meter container houses 2500 servers, about 10 times as many as conventional data centers pack in the same space. Each container integrates computing, networking, power, and cooling systems.

The Million-Server Data Center

Today's most advanced data centers house tens of thousands of servers. What would it take to house 1 million?

photos: Sun Microsystems, CNET, & datacenterknowledge.com

Utility Computing Arrives

- Amazon Elastic Compute Cloud (EC2)
- “Compute unit” rental: $0.08-0.80/hr.
  - □ 1 CU ≈ 1.0-1.2 GHz 2007 AMD Opteron/Xeon core

“Instances”	Platform	Cores	Memory	Disk
Small - $0.10 / hr	32-bit	1	1.7 GB	160 GB
Large - $0.40 / hr	64-bit	4	7.5 GB	850 GB – 2 spindles
XLarge - $0.80 / hr	64-bit	8	15.0 GB	1690 GB – 3 spindles

- Billing rounded to nearest hour; pay-as-you-go storage also available
- A new paradigm for deploying services?
  - □ “Computing as Utility” – A vision from MULTICS c. 1969

But...

What *is* cloud computing, exactly?

“It’s nothing (new)”

“...we’ve redefined Cloud Computing to include everything that we already do... I don’t understand what we would do differently ... other than change the wording of some of our ads.”

– *Larry Ellison, CEO, Oracle (Wall Street Journal, Sept. 26, 2008)*

Above the Clouds: A Berkeley View of Cloud Computing

(February 10, 2009)

abovetheclouds.cs.berkeley.edu

What is it? What's new?

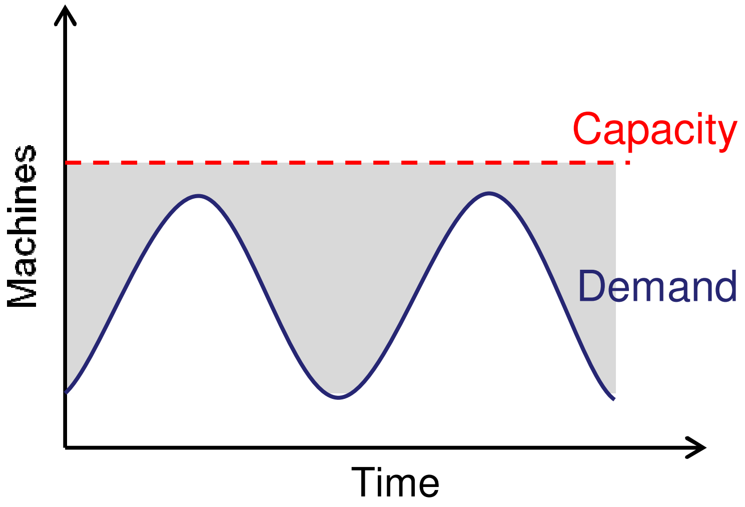
- Old idea: Software as a Service (SaaS)
  - Software hosted in the infrastructure vs. installed on local servers or desktops; dumb (but brawny) terminals
- **New:** pay-as-you-go *utility computing*
  - Illusion of infinite resources on demand
  - Fine-grained billing: release == don't pay
  - Earlier examples: Sun, Intel Computing Services—longer commitment, more $$$/hour, no storage
  - *Public (utility)* vs. *private* clouds

Why Now (not then)?

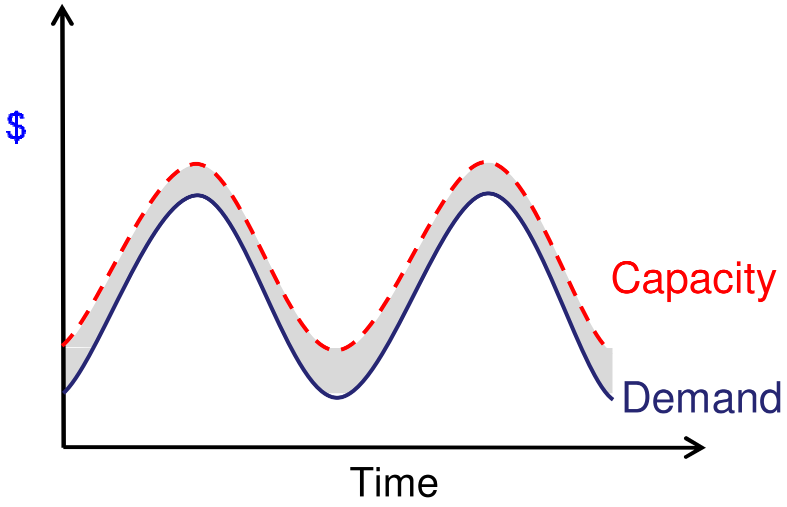
- The Web "Space Race": Build-out of extremely large datacenters (10,000's of ***commodity*** PCs)
- Driven by growth in demand (more users)
  - Infrastructure software: e.g., Google File System
  - Operational expertise
  - Discovered economy of scale: 5-7x cheaper than provisioning a medium-sized (100's machines) facility
- More pervasive broadband Internet
- Free & open source software

Cloud Economics 101

- Static provisioning for peak - wasteful, but necessary for SLA



"Statically provisioned" data center

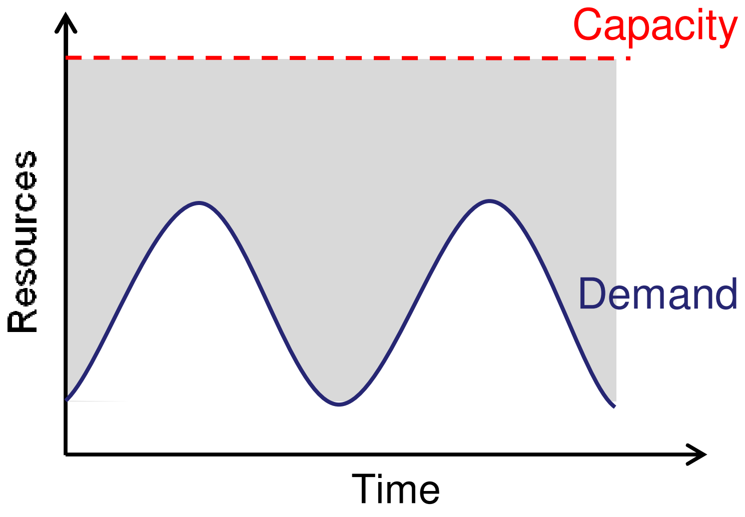


"Virtual" data center in the cloud

▨ Unused resources

Risk of Under Utilization

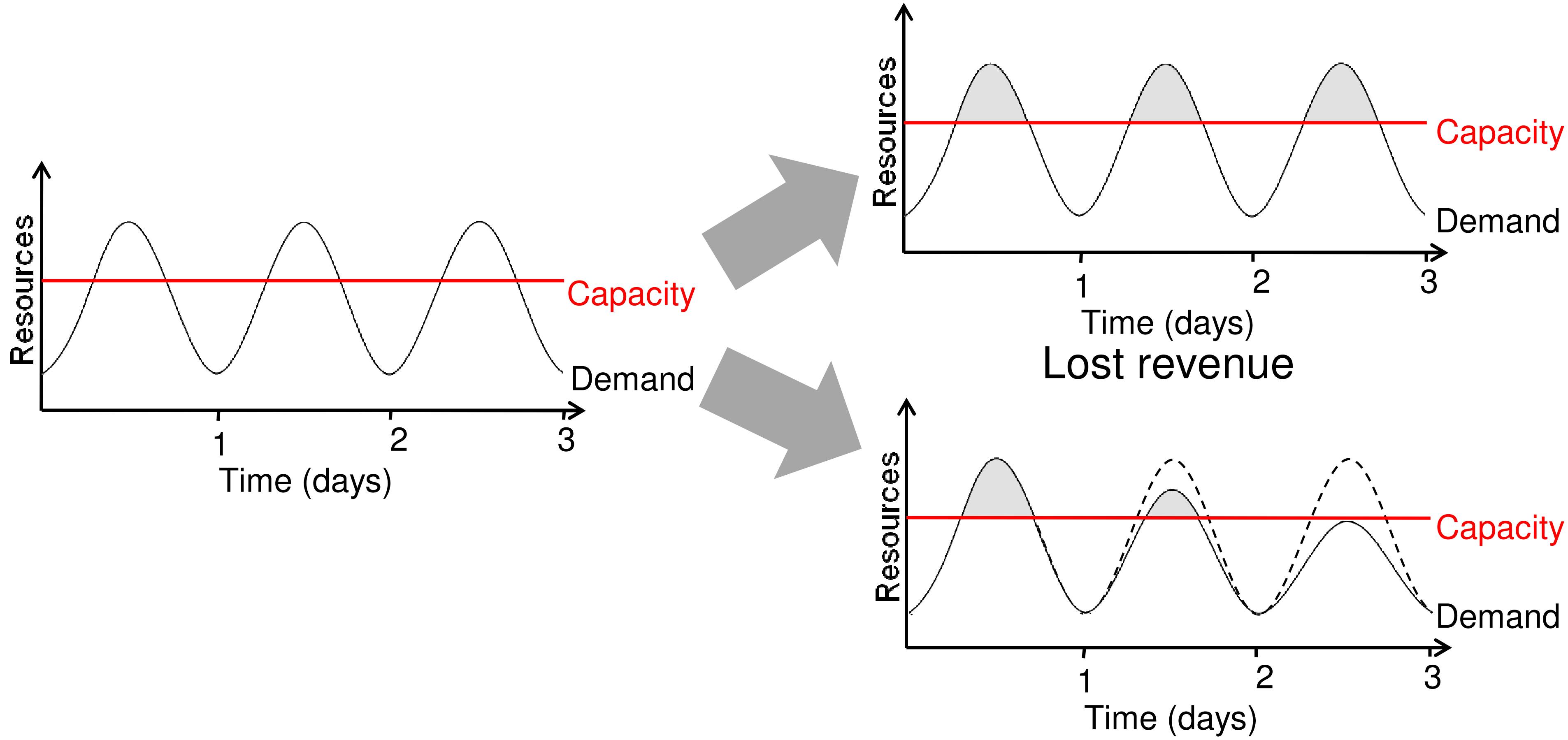
- Underutilization results if “peak” predictions are too optimistic



▨ Unused resources

Static data center

Risks of Under Provisioning

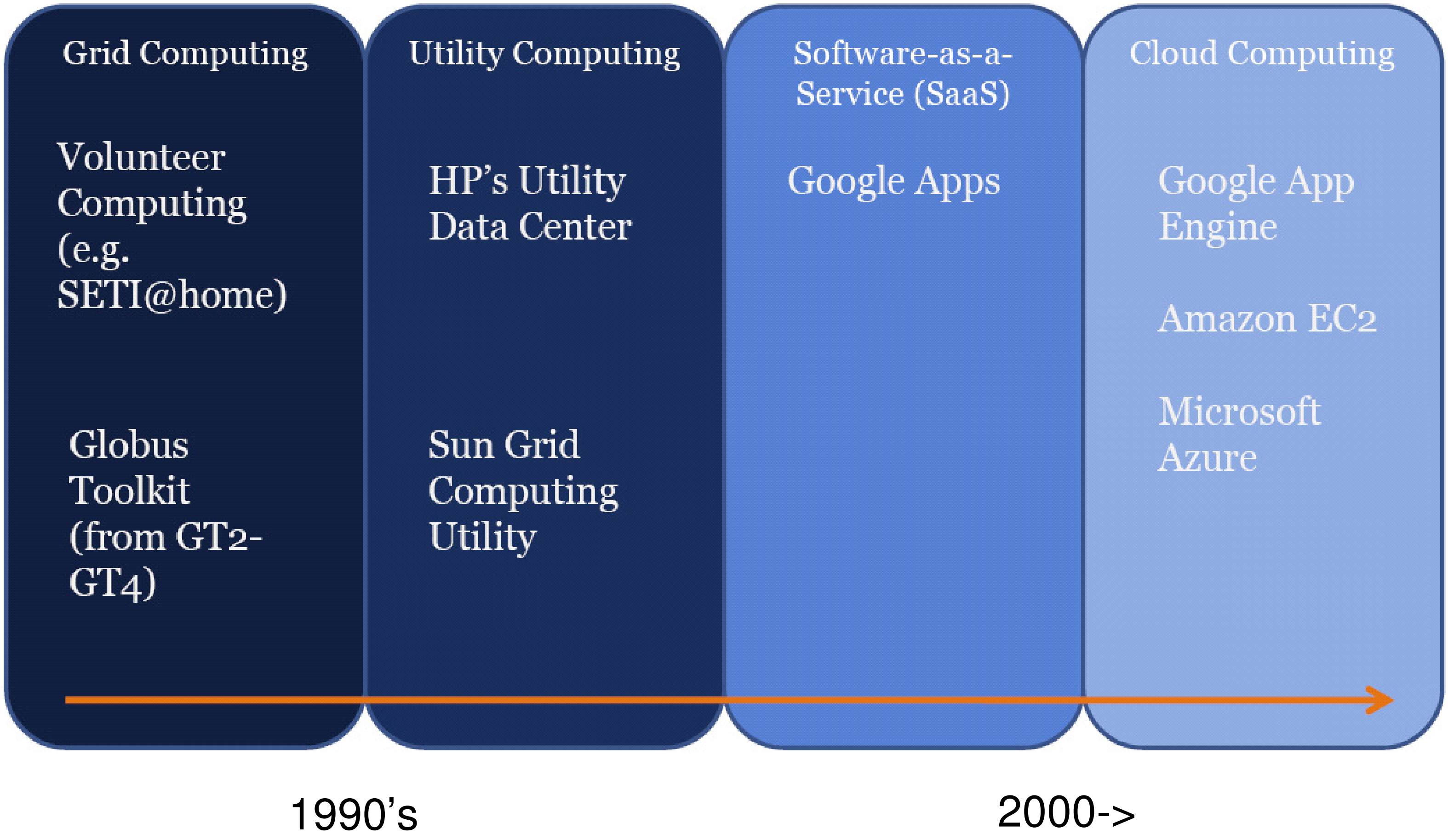


Lost revenue

Lost users

What is Grid Computing then?

Timeline



Grid Computing

Volunteer Computing (e.g. SETI@home)

Globus Toolkit (from GT2-GT4)

Utility Computing

HP's Utility Data Center

Sun Grid Computing Utility

Software-as-a-Service (SaaS)

Google Apps

Cloud Computing

Google App Engine

Amazon EC2

Microsoft Azure

1990's

2000->

Grid Computing

- The term *grid computing* originated in the early 1990s as a metaphor for making computer power as easy to access as an electric power grid in Ian Foster's and Carl Kesselman's seminal work, "The Grid: Blueprint for a new computing infrastructure" (2004).
- Grid computing is a term referring to the combination of computer resources from multiple administrative domains to reach a common goal. The grid can be thought of as a distributed system with non-interactive workloads that involve a large number of files. What distinguishes grid computing from conventional high performance computing systems such as cluster computing is that grids tend to be more loosely coupled, heterogeneous, and geographically dispersed.

Wikipedia.com

Cloud vs. Grid

- Distinctions: not clear maybe because Clouds and Grids share similar visions
  - □ Reducing computing costs
  - □ Increasing flexibility and reliability by using third-party operated hardware
- Grid
  - □ System that coordinates resources which are not subject to centralized control, using standard, open, general-purpose protocols and interfaces to deliver nontrivial qualities of service
  - □ Ability to combine resources from different organizations for a common goal

Cloud vs. Grid – Feature Comparison

- Resource Sharing
  - Grid: collaboration (among Virtual Organizations, for fair share)
  - Cloud: assigned resources not shared
- Virtualization
  - Grid: virtualization of data and computing resources
  - Cloud: virtualization of hardware and software platforms
- Security
  - Grid: security through credential delegations
  - Cloud: security through isolation

Cloud vs. Grid – Feature Comparison (Contd.)

- Self Management
  - Grid: reconfigurability
  - Cloud: reconfigurability and self-healing
- Payment Model
  - Grid: rigid
  - Cloud: flexible

Grid computing is often (but not always) associated with the delivery of cloud computing systems

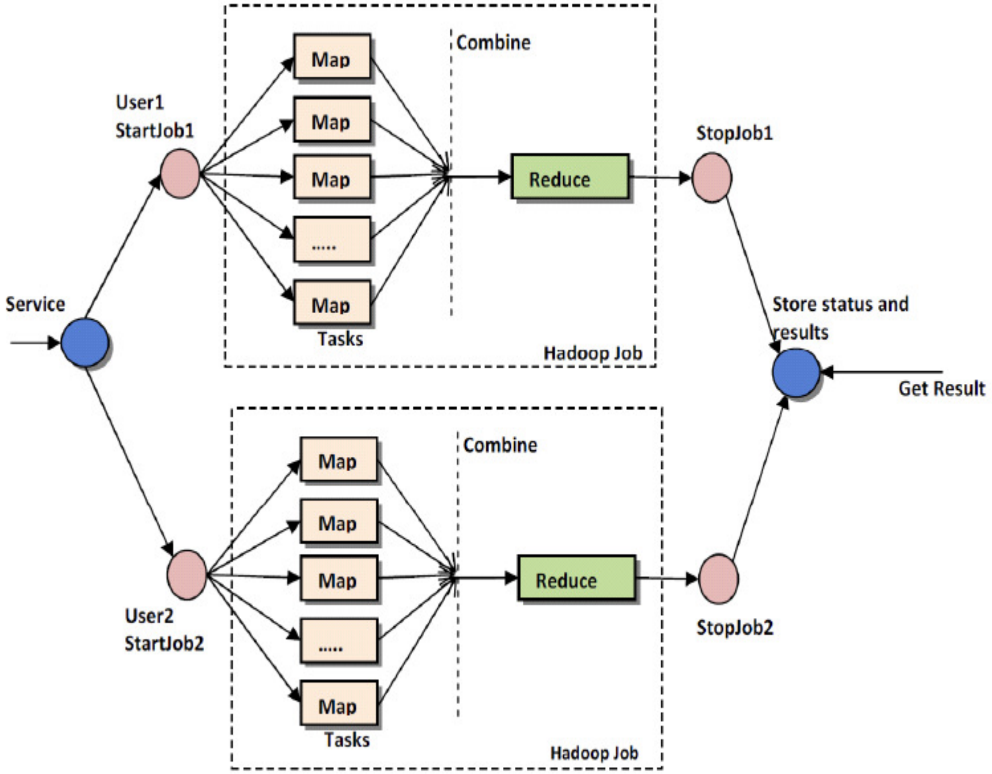
What can you do with Cloud Computing?

Cost Associativity

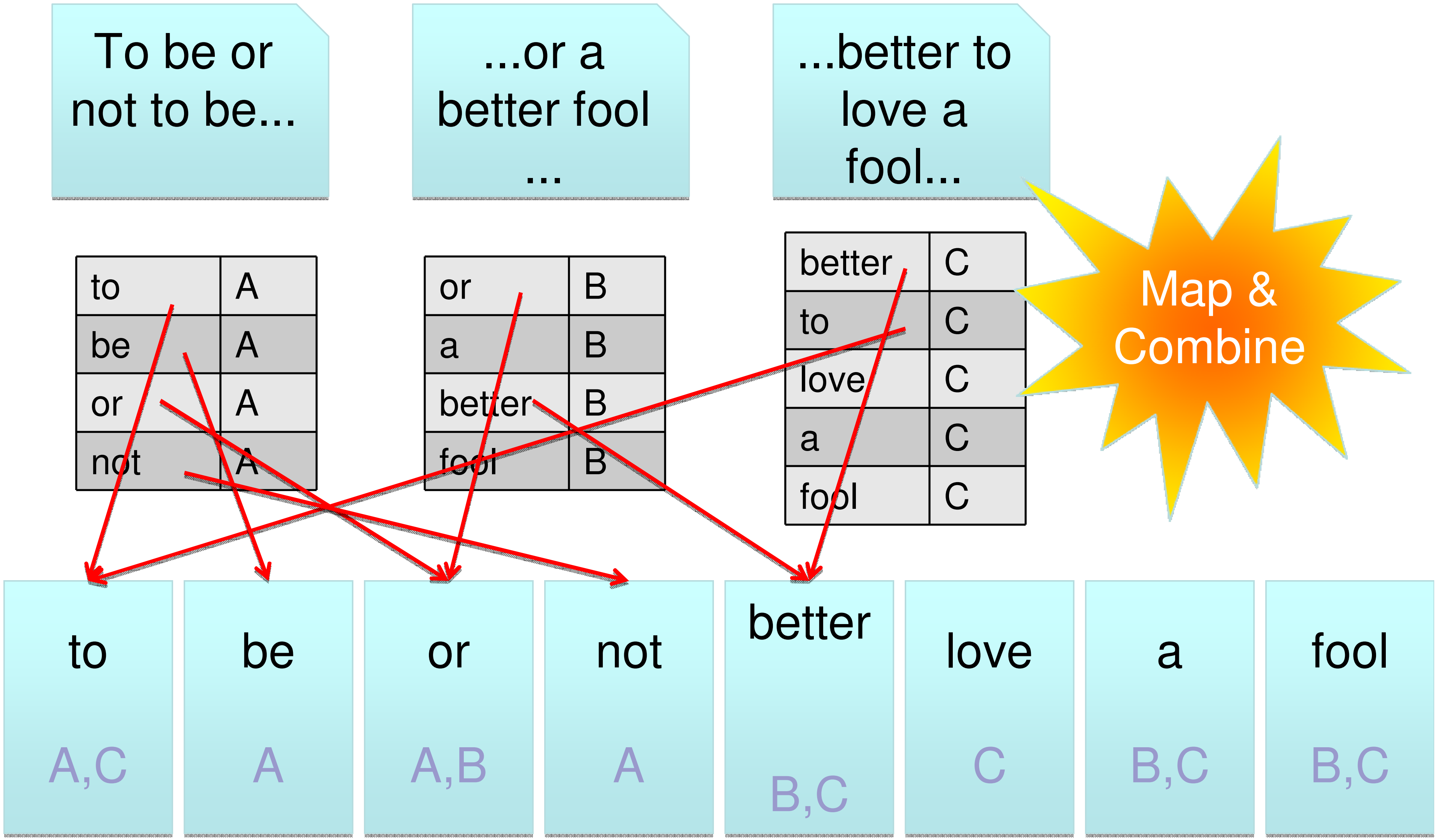
- 1,000 CPUs for 1 hour same price as 1 CPU for 1,000 hours
- Washington Post converted Hillary Clinton's travel documents to post on WWW
  - Conversion time: **<1 day** after released
  - Cost: less than $200
- RAD Lab graduate students demonstrate improved MapReduce scheduling—on 1,000 servers

Map Reduce

- MapReduce is a patented software framework introduced by Google to support distributed computing on large data sets on clusters of computers.
- The framework is inspired by the map and reduce functions commonly used in functional programming
- "Map" step: The master node takes the input, partitions it up into smaller sub-problems, and distributes those to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes that smaller problem, and passes the answer back to its master node.
- "Reduce" step: The master node then takes the answers to all the sub-problems and combines them in some way to get the output — the answer to the problem it was originally trying to solve.
- Hadoop, Apache's free and open source implementation of MapReduce.



Indexing the Web



To be or not to be...

...or a better fool ...

...better to love a fool...

to	A
be	A
or	A
not	A

or	B
a	B
better	B
fool	B

better	C
to	C
love	C
a	C
fool	C

Map & Combine

to	be	or	not	better	love	a	fool
A,C	A	A,B	A	B,C	C	B,C	B,C

MapReduce in Practice

- Example: spam classification
  - training: 10^7 URLs x 64KB data each = 640GB data
  - One heavy-duty server: ~270 hours
  - 100 servers in cloud: ~3 hours (= ~$255)
- Rapid uptake in other scientific research
  - Large-population genetic risk analysis & simulation (Harvard Medical School)
  - Genome sequencing (UNC Chapel Hill Cancer Ctr)
  - many others... so *what's the downside?*

Risk transfer

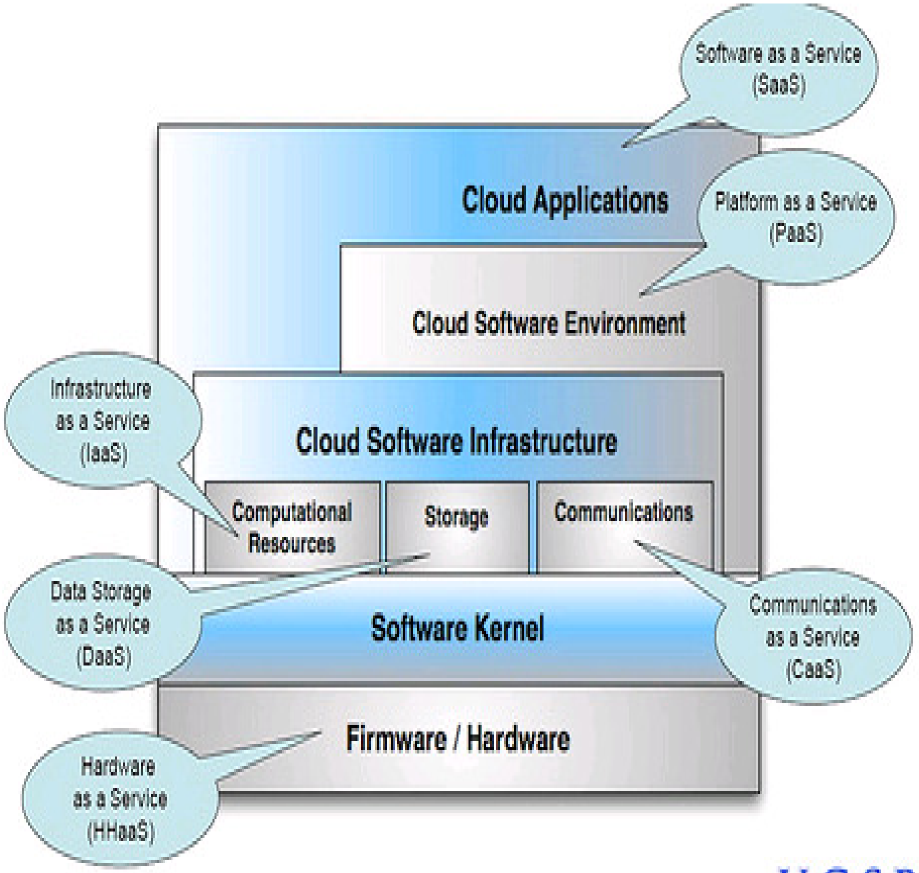
- 2001: CNN home page meltdown on 9/11
  - ~10x traffic increase in ~15 minutes
  - result: site had to go offline
- 2008: Animoto
  - traffic doubled every 12 hours for 3 days when released as Facebook plug-in
  - Scaled from 50 to >3500 servers
  - ***...then scaled back down***

Evolution of Cloud Service Architecture

Cloud Service Taxonomy

- Layer
  - □ Hardware-as-a-Service(HaaS)
  - □ **Infrastructure-as-a-Service (IaaS)**
  - □ **Data Storage-as-a-Service (DaaS)**
  - □ Communication-as-a-Service (CaaS)
  - □ **Platform-as-a-Service (PaaS)**
  - □ **Software-as-a-Service (SaaS)**

- Type
  - □ Public cloud
  - □ Private cloud
  - □ Inter-cloud


Software as a Service (SaaS)
Cloud Applications
Platform as a Service (PaaS)
Cloud Software Environment
Infrastructure as a Service (IaaS)
Cloud Software Infrastructure
Computational Resources
Storage
Communications
Data Storage as a Service (DaaS)
Software Kernel
Communications as a Service (CaaS)
Firmware / Hardware
Hardware as a Service (HHaaS)

UCSB

Infrastructure-as-a-Service

- Definition
  - Provision model in which an organization outsources the equipment used to support operations, including storage, hardware, servers and networking components
    - Infrastructure as a Service is sometimes referred to as Hardware as a Service (HaaS).
      - Virtualization of hardware resource resource
    - The service provider owns the equipment and is responsible for housing, running and maintaining it
    - The client typically pays on a per-use basis

Example: Amazon EC2

Data Storage as a Service (DaaS)

- Definition
  - □ Delivery of data storage as a service, including database-like services, often billed on a utility computing basis
    - Database (Amazon SimpleDB & Google App Engine's BigTable datastore)
    - Network attached storage (MobileMe iDisk & Nirvanix CloudNAS)
    - Synchronization (Live Mesh Live Desktop component & MobileMe push functions)
    - Web service (Amazon Simple Storage Service & Nirvanix SDN)

Communication as a Service (CaaS)

- Definition
  - Delivery of communication software as a service, billed on a utility computing basis
    - The newly developed Avaya Aura™ System Platform technology leverages standard virtualization software to run multiple certified UC applications on certified x86 servers.
      - The Avaya Aura System Platform aggregates Avaya Aura Communication Manager (supporting voice, video, and call center ACD), messaging server, media services, SIP support, and application enablement and utilities (serviceability, management, DHCP/HTTPS) on a single server. The solution supports up to 2,400 users and up to 250 locations, and installation can be completed in about one hour. The midsize solution was built to provide organizations with fewer resources and less capital than large enterprises a means to flexibly and easily deploy a scalable UC solution at an affordable price point.

Platform-as-a-Service (PaaS)

- Definition
  - Platform providing all the facilities necessary to support the complete process of building and delivering customized applications and services, all available over the Internet
  - Entirely virtualized platform that includes one or more servers, operating systems and specific applications

Example: Google App Engine provides the developers with a sandbox for User's Java and Python programs for Web Application development and deployment

Software-as-a-Service (SaaS)

- Definition
  - □ *Software deployed as a hosted service and accessed over the Internet*
- Features
  - □ *Open, Flexible*
  - □ *Easy to Use*
  - □ *Easy to Upgrade*
  - □ *Easy to Deploy*

Example: Gmail, Google Docs……..

Public Cloud

- Definition
  - The standard cloud computing model
    - The SP makes resources, such as applications and storage, available to the general public over the Internet
  - Free or offered on a pay-per-use model
- Examples of public clouds
  - Amazon Elastic Compute Cloud (EC2), IBM's Blue Cloud, Sun Cloud, Google AppEngine and Windows Azure Services Platform.

Private Cloud Services

- Internal cloud or corporate cloud
- Definition
  - Proprietary computing architecture that provides hosted services to a limited number of people behind a firewall
    - Designed to appeal to an organization that needs or wants more control over their data than they can get by using a third-party hosted service

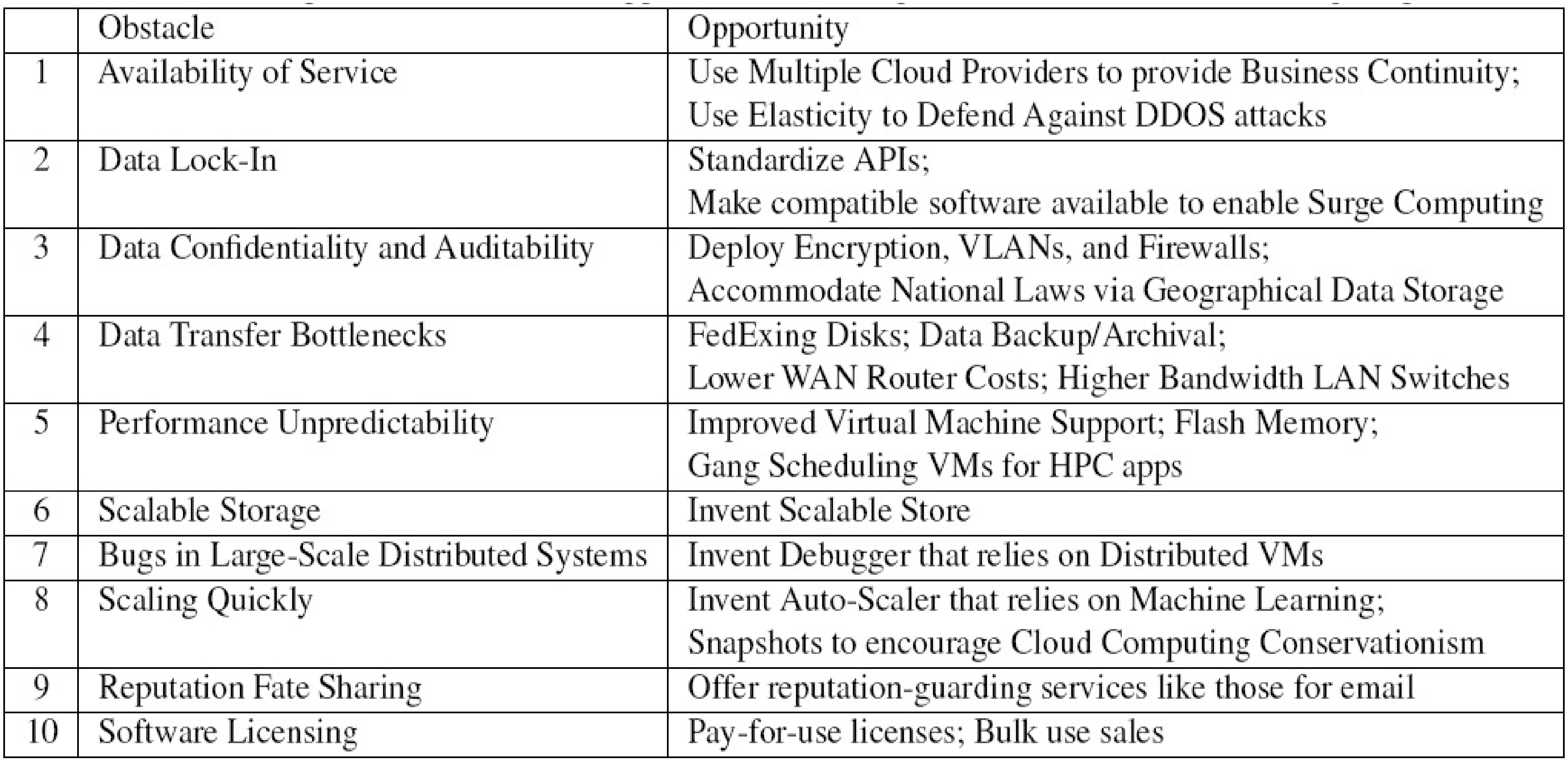
Inter-Cloud

- Definition
  - Federation of clouds based on open standards
  - Concept primarily promoted by Cisco
- Similar Concepts
  - Cloud of Clouds, Cloud Interoperability, etc.

Challenges & Opportunities

- Challenges to adoption, growth, & business/policy models
- Both technical and nontechnical
- Most translate to 1 or more *opportunities*

Top 10 Obstacles and Opportunities for Adoption and Growth of Cloud Computing



	Obstacle	Opportunity
1	Availability of Service	Use Multiple Cloud Providers to provide Business Continuity; Use Elasticity to Defend Against DDOS attacks
2	Data Lock-In	Standardize APIs; Make compatible software available to enable Surge Computing
3	Data Confidentiality and Auditability	Deploy Encryption, VLANs, and Firewalls; Accommodate National Laws via Geographical Data Storage
4	Data Transfer Bottlenecks	FedExing Disks; Data Backup/Archival; Lower WAN Router Costs; Higher Bandwidth LAN Switches
5	Performance Unpredictability	Improved Virtual Machine Support; Flash Memory; Gang Scheduling VMs for HPC apps
6	Scalable Storage	Invent Scalable Store
7	Bugs in Large-Scale Distributed Systems	Invent Debugger that relies on Distributed VMs
8	Scaling Quickly	Invent Auto-Scaler that relies on Machine Learning; Snapshots to encourage Cloud Computing Conservationism
9	Reputation Fate Sharing	Offer reputation-guarding services like those for email
10	Software Licensing	Pay-for-use licenses; Bulk use sales

©Above the clouds

Summary

- Cloud computing *democratizes access* to "supercomputer-class" capability
  - All you need is a credit card
- Puts students, academia on more level playing field to have high impact in industry
- The next Google, eBay, Amazon, etc. can come from a small team of entrepreneurs even *without* heavy dose of $$ up front

Thank You!