CSE335: Cloud Computing

The Cloud

Jaehong Kim

(jaehong.kim@khu.ac.kr)

Annoucements

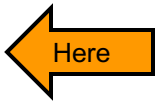
- Homework1 will soon be available on the ecampus
  - If you run into problems, please post a message to the discussion group.
    - Please avoid sending questions via email – if you post your question on ecampus, everyone can discuss the question and see the answer to the question
- Please prepare to discuss everything in ecampus

Scale?

- What is Scalable?
  - A computer system is called scalable
    - If it can scale up to accommodate ever-increasing performance and functionality demand and/or scale down to reduce cost
  - A software is called scalable
    - If it can maintains its functionality and efficiency while the underlying computer system and problem scale up and/or down

Plan for today

- Large-scale computing
  - The need for scalability
  - Scale of current services
  - Scaling up : From PCs to data centers
  - Problems with 'classical' scaling techniques

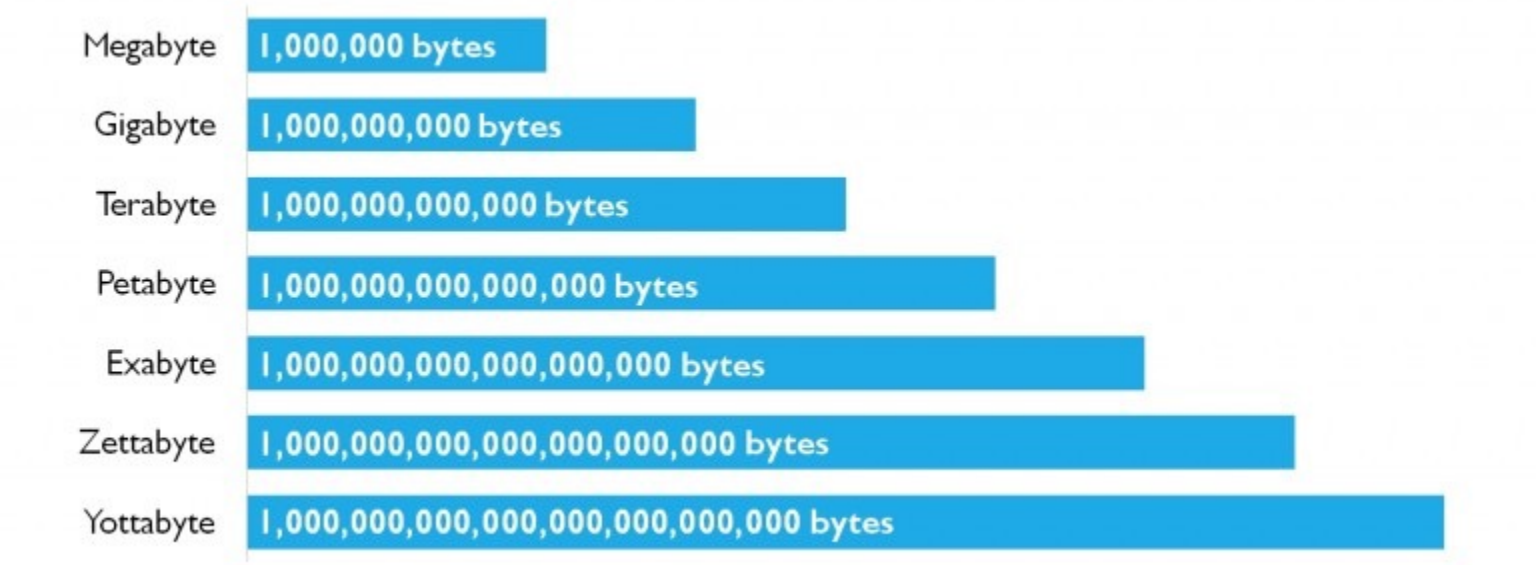


- Utility computing and cloud computing
  - What are utility computing and cloud computing?
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization : How does cloud computing actually work?
  - Some cloud computing challenges

How many users and objects?

- A company has > 10 billion photos
- B company has > 2 billion active users
- C company is serving 2 trillion + queries per year, on more than 27 billion items
- >1 billion hours of videos/day watched on D company, by > 1 billion users

Storage Units Review



Sources: : 31stfebruary

How much data?

- Modern applications use massive data:
  - Rendering 'Avatar' movie required >1 petabyte of storage
  - German Climate computing center dimensioned for 60 petabytes of climate data
  - eBay has >90 petabytes of user data
  - CERN recently passed the 200-petabyte milestone
  - Google is estimated to have ~10 exabytes of data
  - NSA Utah Data Center is said to have 5 zettabyte (?)

- How much is a zettabyte?
  - 1,000,000,000,000,000,000,000 bytes
  - A stack of 1TB hard disks that is 25,400 km high

How much computation?

- No single computer can process that much data
  - Need many computers!

- How many computers do modern services need?
  - Facebook is thought to have more than 60,000 servers
  - 1&1 Internet has over 70,000 servers
  - Akamai has 95,000 servers in 71 countries
  - Intel has ~100,000 servers in 97 data centers
  - Microsoft reportedly had at least 200,000 servers in 2008
  - Google is thought to have more than 1 million servers, is planning for 10 million (according to Jeff Dean)

Why should I care?

- Suppose you want to build the next Google
- How do you
  - …download and store billions of web pages and images?
  - …quickly find the pages that contain a given set of terms?
  - …find the pages that are most relevant to a given search?
  - …answer billions of queries of this type every day?

- Suppose you want to build the next Facebook
- How do you
  - …store the profiles of over 2 billion users?
  - …avoid losing any of them?
  - … find out which users might want to be friends?

Plan for today

- Large-scale computing
  - The need for scalability
  - Scale of current services
  - Scaling up : From PCs to data centers
  - Problems with 'classical' scaling techniques

Here

- Utility computing and cloud computing
  - What are utility computing and cloud computing?
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization : How does cloud computing actually work?
  - Some cloud computing challenges

Scaling up

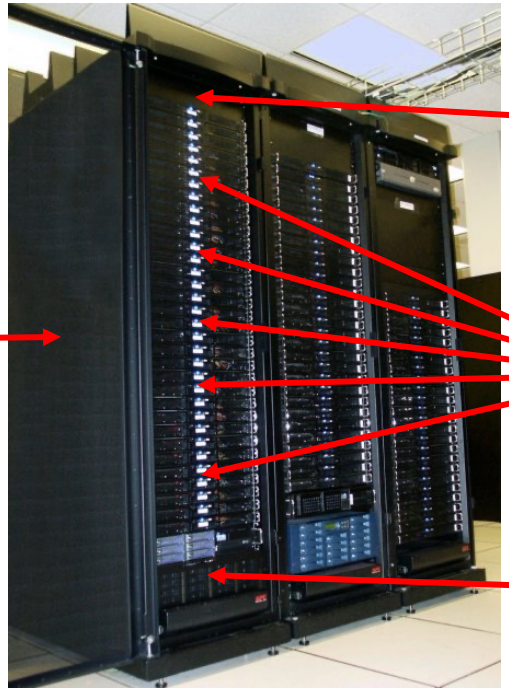
PC

Server

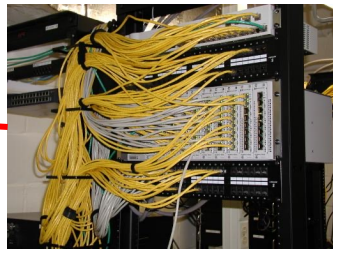
Cluster

- What if one computer is not enough?
  - Buy a bigger (server-class) computer
- What if the biggest computer is not enough?
  - Buy many computers

Clusters



Rack



Network switch
(connects nodes with each other and with other racks)

Many nodes/blades
(often identical)

Storage device(s)

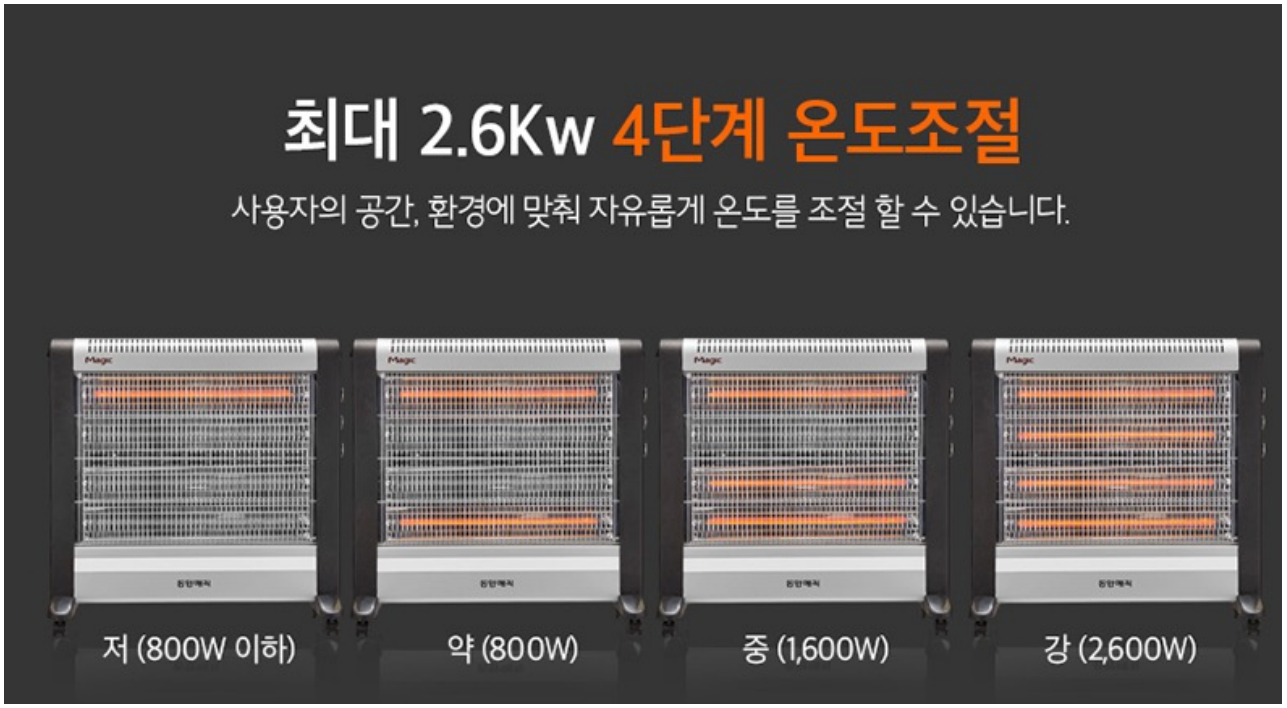
- Characteristics of a cluster:
  - Many similar machines, close interconnection
  - Often special, standardized hardware (racks, blades)
  - Usually owned and used by a single organization

Power and cooling

- Clusters need lots of power
  - Example: 140 Watts per server
  - Rack with 32 servers: 4.5kW (needs special power supply!)
  - Most of this power is converted into heat

- 4.5kW heater??

Power and cooling



- Large clusters need massive cooling
  - 4.5kW is about 1.7 heaters
  - And that's just one rack!

Scaling up

PC

Server

Cluster

Data center

- What if your cluster is too big (hot, power hungry) to fit into your office building?
  - Build a separate building for the cluster
  - Building can have lots of cooling and power
  - Result: Data center

What does a data center look like?

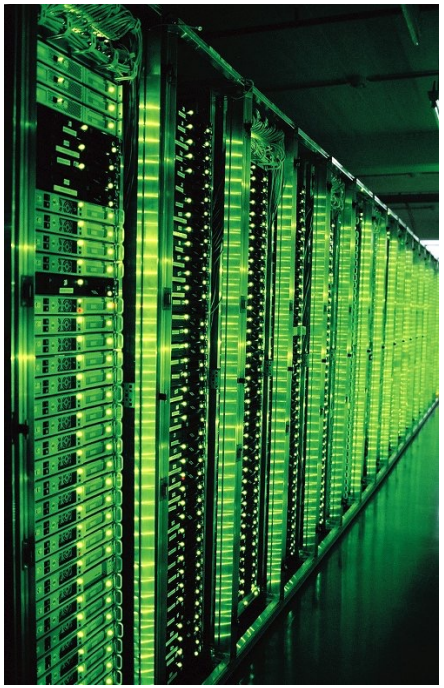
Cooling plant

Data centers (size of a football field)

Google data center in The Dalles, Oregon

- A warehouse-sized computer
  - A single data center can easily contain 10,000 racks with 100 cores in each rack (1,000,000 cores total)

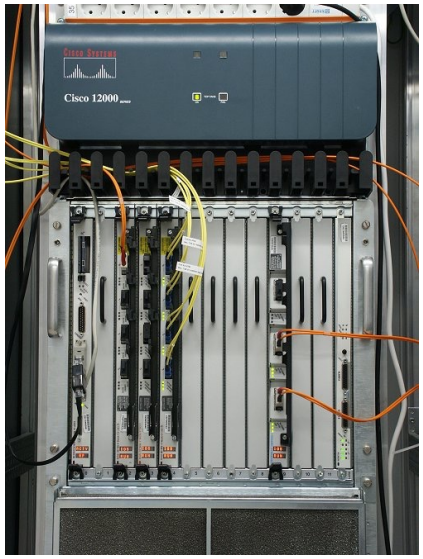
What's in a data center?



Sources: : 1&1

- Hundreds or thousands of racks

What's in data center?



- Massive networking

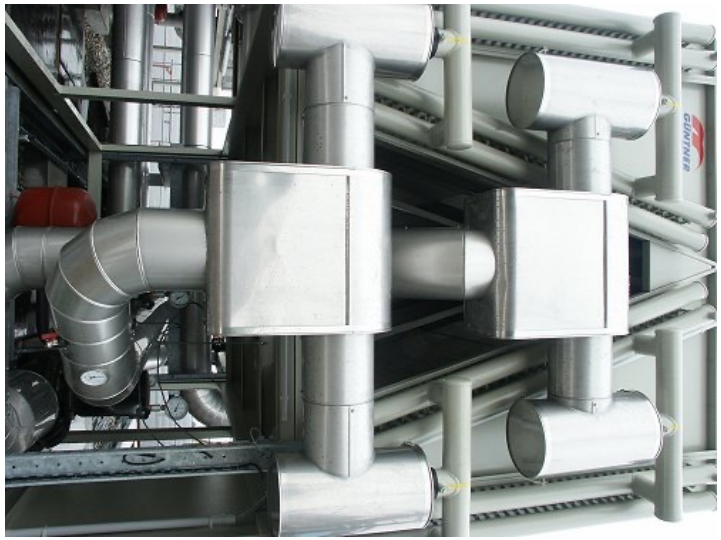
Sources: : 1&1

What's in a data center?

Sources: : 1&1

- Emergency power supplies

What's in a data center?



- Massive cooling

Sources: : 1&1

Energy matters!

Company	Servers	Electricity	Cost
eBay	16K	~$0.6*10^5$ MWh	~$3.7M/yr
Akamai	40K	~$1.7*10^5$ MWh	~$10M/yr
Rackspace	50K	~$2*10^5$ MWh	~$12M/yr
Microsoft	>200K	>$6*10^5$ MWh	>$36M/yr
Google	>500K	>$6.3*10^5$ MWh	>$38M/yr
USA (2006)	10.9M	$610*10^5$ MWh	$4.5B/yr

Source: Qureshi et al., SIGCOMM 2009

- Data centers consumes a lot of energy
  - Makes sense to build them near sources of cheap electricity
  - Example: Price per KWh is 3.6ct in Idaho (near hydroelectric power), 10ct in California (long distance transmission), 18ct in Hawaii (must ship fuel)
  - Most of this is converted into heat -> cooling is a big issue!

Scaling up

PC

Server

Cluster

Data center

Network of data centers

- What if even a data center is not big enough?
  - Build additional data centers
  - Where? How many?

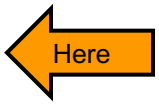
Global distribution

- Data centers are often globally distributed
  - Examples above: Google data center locations
- Why?
  - Need to be close to users (physics!)
  - Cheaper resources
  - Protection against failures

http://www.google.com/about/datacenters/inside/locations/index.html

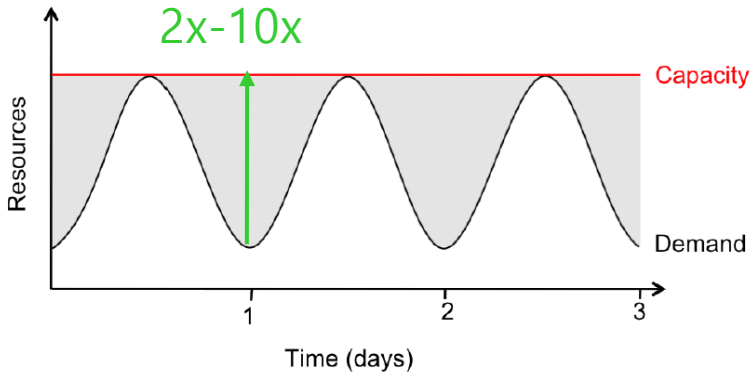
Plan for today

- Large-scale computing
  - The need for scalability
  - Scale of current services
  - Scaling up : From PCs to data centers
  - Problems with 'classical' scaling techniques

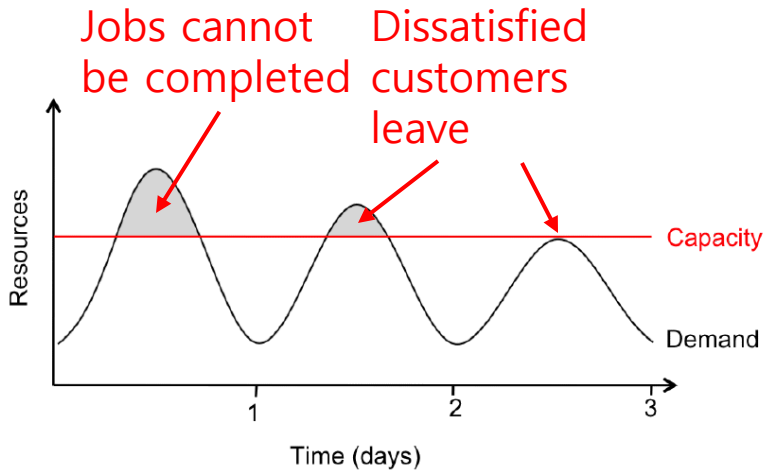


- Utility computing and cloud computing
  - What are utility computing and cloud computing?
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization : How does cloud computing actually work?
  - Some cloud computing challenges

Problem #1 : Difficult to dimension



Provisioning for the peak load



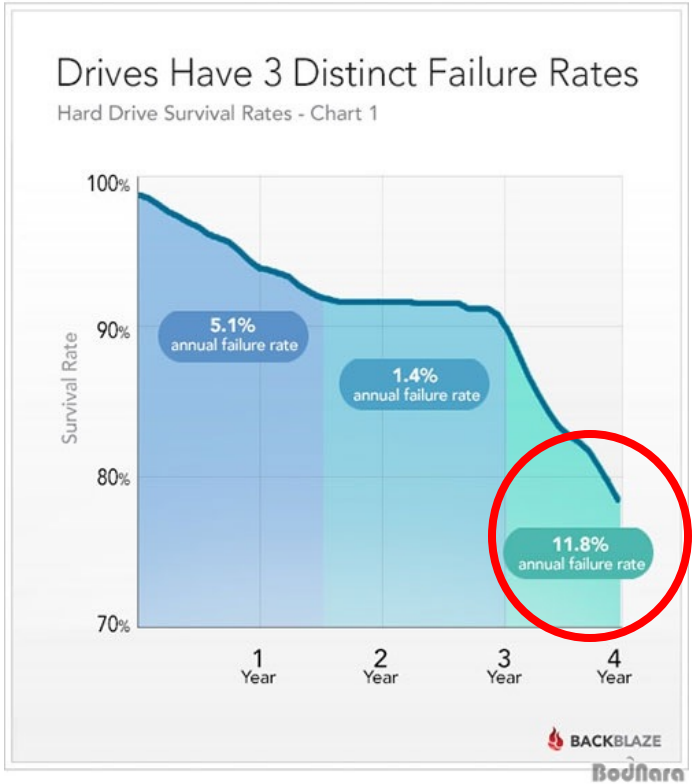
Provisioning below the peak

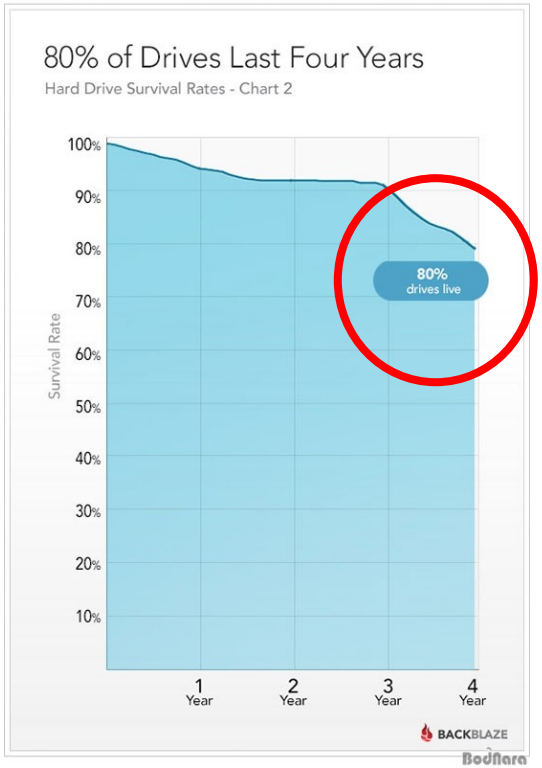
- Problem: Load can vary considerably
  - Peak load can exceed average load by factor 2x-10x [Why?]
  - But: Few users deliberately provision for less than the peak
  - Result: Server utilization in existing data centers ~5%-20%!!
  - Dilemma: Waste resources or lose customers!

Problem #2 : Expensive

- Need to invest many $$$ in hardware
  - Even a small cluster can easily cost $100,000
  - The NSA operates a data center in Utah that reportedly cost $1.5 billion to bulid
- Need expertise
  - Planning and setting up a large cluster is highly nontrivial
  - Cluster may require special software, etc.
- Need maintenance
  - Someone needs to replace faulty hardware, install software upgrades, maintain user accounts, …

About maintenance





Sources: : http://www.bodnara.co.kr

- Need maintenance
  - Disk has a sudden failure after three years
  - In four years, 20 out of 100 disks fail.
  - If you run 100,000 disks for four years, 20,000 disks fail.
  - Since four years are about 1500 days, an average of thirteen (20,000 / 1500) units a day can be disabled.

Problem #3 : Difficult to scale

- Scaling up is difficult
  - Need to order new machines, install them, integrate with existing cluster – can take weeks
  - Large scaling factors (massive expansion factors) may require major redesigns such as new storage system, new interconnect, new buildings, and so on
- Scaling down is difficult
  - What should we do with unnecessary hardware?
  - Server idle power is about 60% of peak
    - Energy is consumed even when no work is being done
  - Many fixed costs, such as construction

Recap: Large-scale computing

- Modern applications require huge amounts of processing and data
  - Measured in exabytes, billions of users, billions of objects
  - Need special hardware, algorithms, tools to work at this scale

- Clusters and data centers can provide the resources we need
  - Main difference : Scale (room-sized vs. building-sized)
  - Special hardware; power and cooling are big concerns

- Clusters and data centers are not perfect
  - Difficult to dimension; expensive; difficult to scale

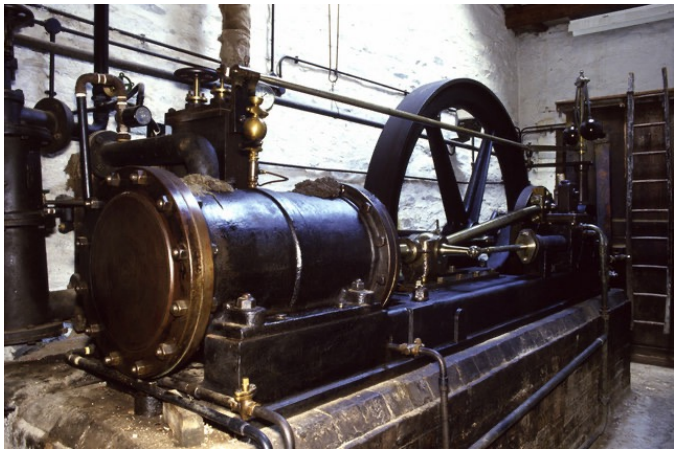
Plan for today

- Large-scale computing
  - The need for scalability
  - Scale of current services
  - Scaling up : From PCs to data centers
  - Problems with 'classical' scaling techniques

- Utility computing and cloud computing
  - What are utility computing and cloud computing? Here
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization : How does cloud computing actually work?
  - Some cloud computing challenges

The power plant analogy

Waterwheel at the Neuhausen ob Eck Open-Air Museum



Steam engine at Stott Park Bobbin Mill

- It used to be that everyone had their own power source
  - Challenges are similar to the cluster: Needs large up-front investment, expertise to operate, difficult to scale up/down…

Scaling the power plant




- Then people started to build large, centralized power plants with very large capacity…

Metered usage model

- Power plants are connected to customers by a network
- Usage is metered, and everyone (basically) pays only for what they actually use

Why is this a good thing?

- Electricity
- Computing

- Economies of scale
  - Cheaper to run one big power plant than many small ones
  - Cheaper to run one big data center than many small ones
- Statistical multiplexing
  - High utilization!
  - High utilization!
- No up-front commitment
  - No investment in generator; pay-as-you-go model
  - No investment in datacenter; pay-as-you-go model
- Scalability
  - Thousands of kilowatts available on demand; add more within seconds
  - Thousands of computers available on demand; add more with seconds

What is cloud computing?

The interesting thing about Cloud Computing is that we've redefined Cloud Computing to include everything that we already do.... I don't understand what we would do differently in the light of Cloud Computing other than change the wording of some of our ads.

Larry Ellison, quoted in the Wall Street Journal

A lot of people are jumping on the [cloud] bandwagon, but I have not heard two people say the same thing about it. There are multiple definitions out there of "the cloud".

Andy Isherwood, quoted in ZDnet News

So what is it, really?

- According to NIST:
  - Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., network, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.
- Essential characteristics:
  - On-demand self service
  - Broad network access
  - Resource pooling
  - Rapid elasticity
  - Measured service

What is cloud computing?

- Cloud computing
  - On-demand delivery of IT resources and applications via the internet with pay-as-you-go pricing (from AWS)

Other terms you may have heard

- Utility computing
  - The service being sold by a cloud
  - Focuses on the business model (pay-as-you-go), similar to classical utility companies
- The Web
  - The Internet's information sharing model
  - Some web services run on clouds, but not all
- The Internet
  - A network of networks
  - Used by the web; connects (most) clouds to their customers

Plan for today

- Large-scale computing
  - The need for scalability
  - Scale of current services
  - Scaling up : From PCs to data centers
  - Problems with 'classical' scaling techniques

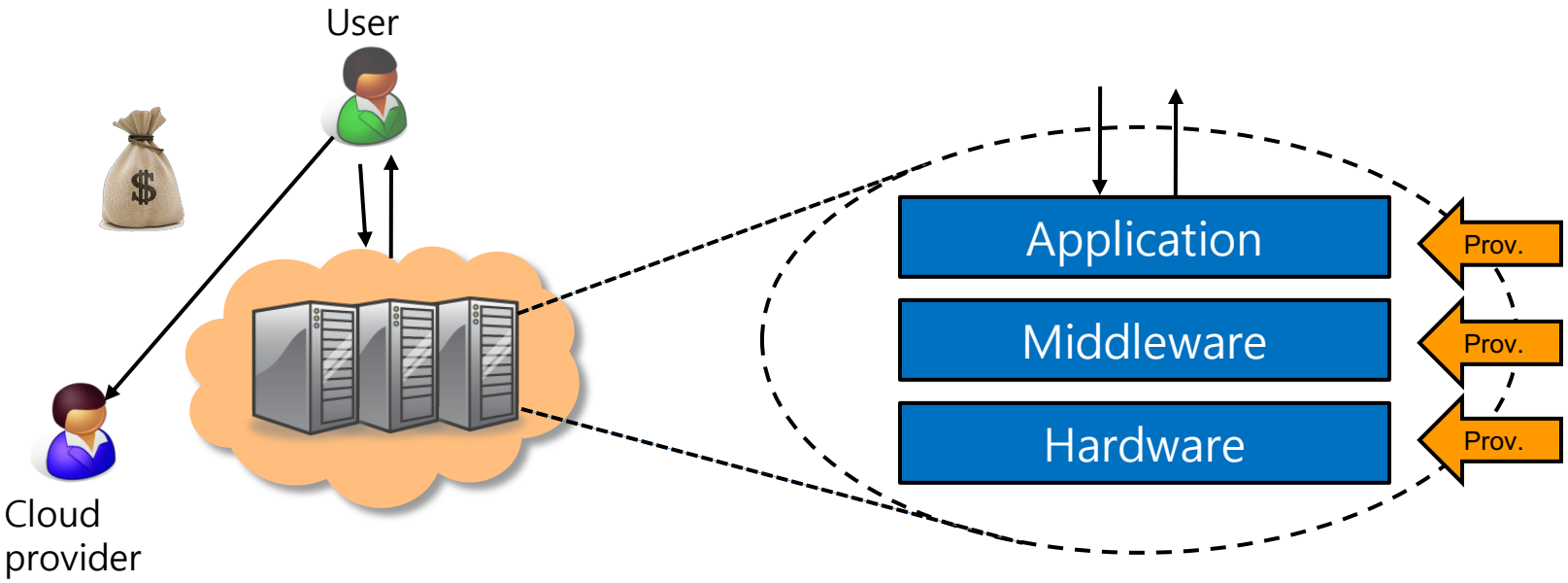
- Utility computing and cloud computing
  - What are utility computing and cloud computing?
  - What kinds of clouds exist today? Here
  - What kinds of applications run on the cloud?
  - Virtualization : How does cloud computing actually work?
  - Some cloud computing challenges

Everything as a Service (XaaS)

- What kind of service does the cloud provide?
  - Does it offer an entire application, or just resources?
  - If resources, what kind / level of abstraction?

- Three main types commonly distinguished:
  - Software as a service (SaaS)
    - Analogy: Restaurant. Prepares & servers entire meal, does the dishes, …
  - Platform as a service (PaaS)
    - Analogy: Take-out food. Prepares meal, but does not serve it
  - Infrastructure as service (IaaS)
    - Analogy: Grocery store. Provides raw ingredients.
  - Other XaaS types have been defined, but are less common
    - Desktop, Backend, Communication, Network, Monitorinng, Video – even Groceries, Transportation, …

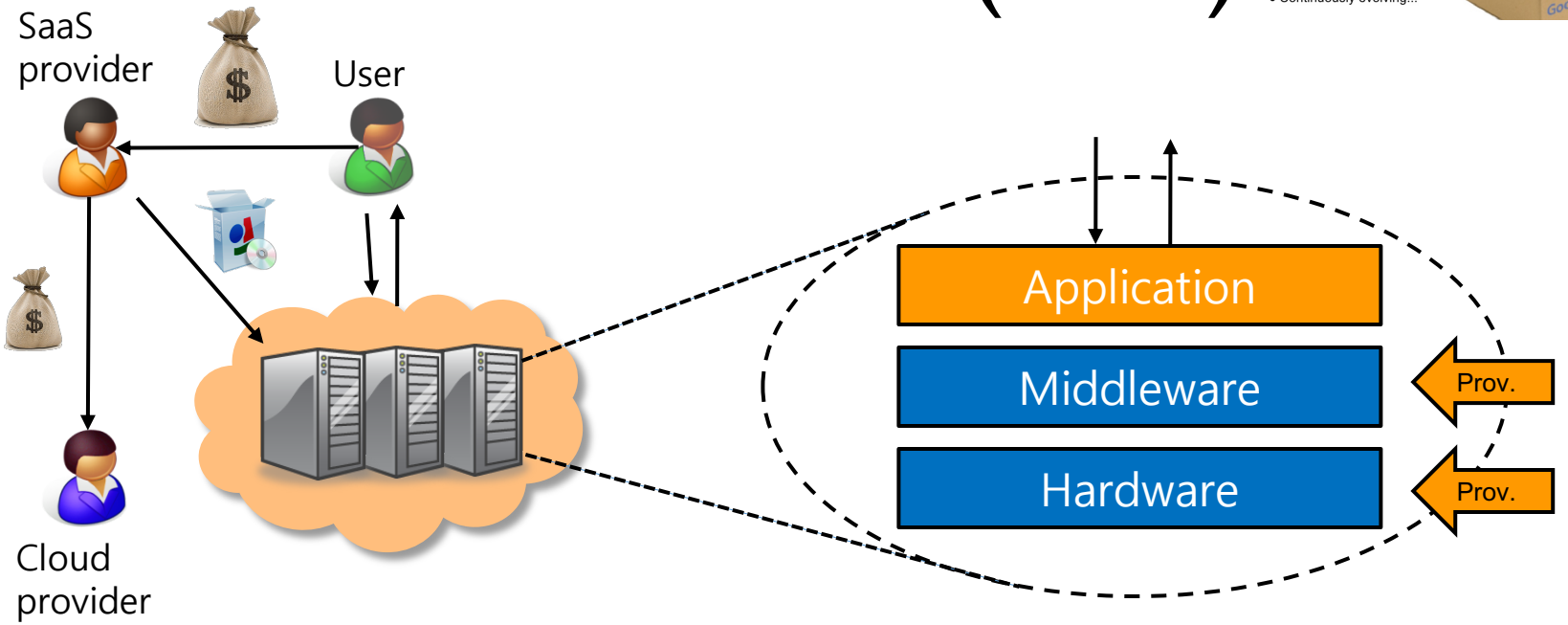
Software as a Service (SaaS)



- Cloud provides an entire application
  - Word processor, spreadsheet, CRM software, calendar…
  - Customer pays cloud provider
  - Example: Google Apps, Salesforce.com

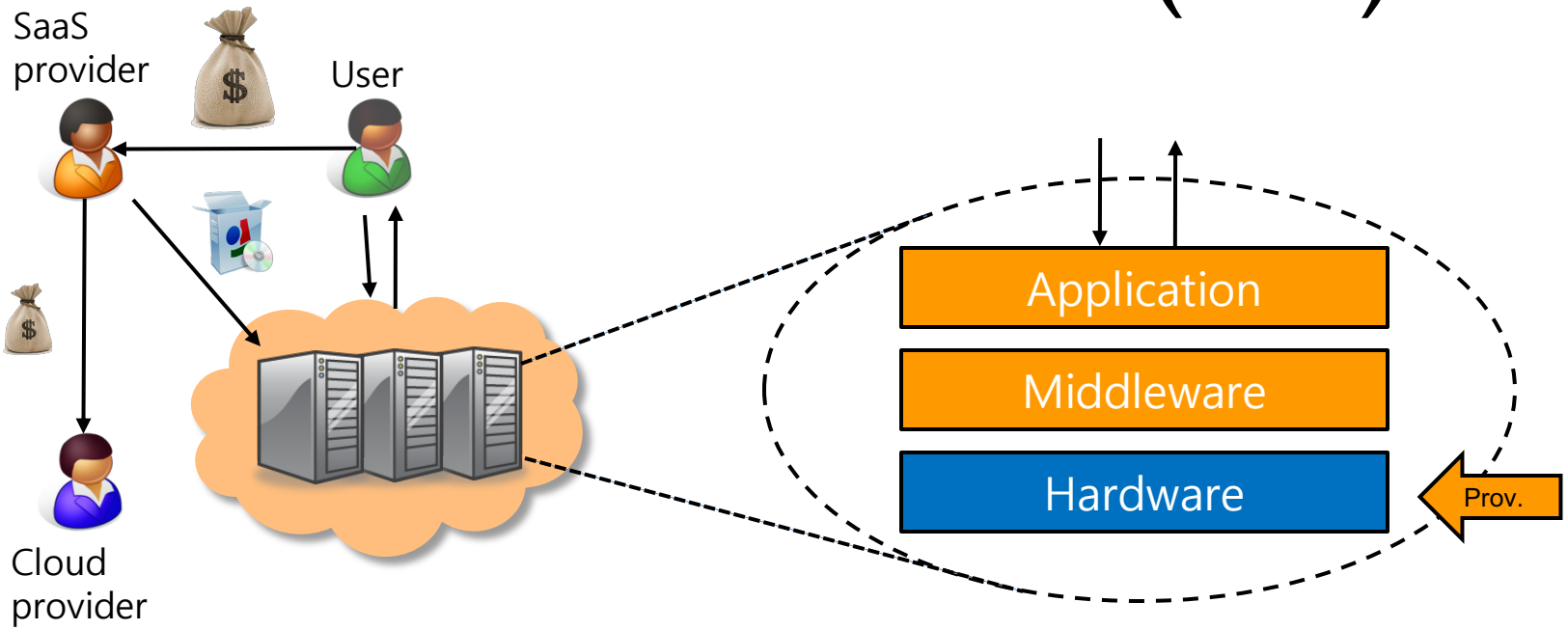
- Downloadable SDK
- Application runtimes
  - Java, Python, Go
- Local development tools
  - Eclipse plugin
- Specialized api services
- Cloud based dashboard
- Automatically scales
- Built in fault tolerance, load balancing
- Continuously evolving...

Platform as a Service (PaaS)



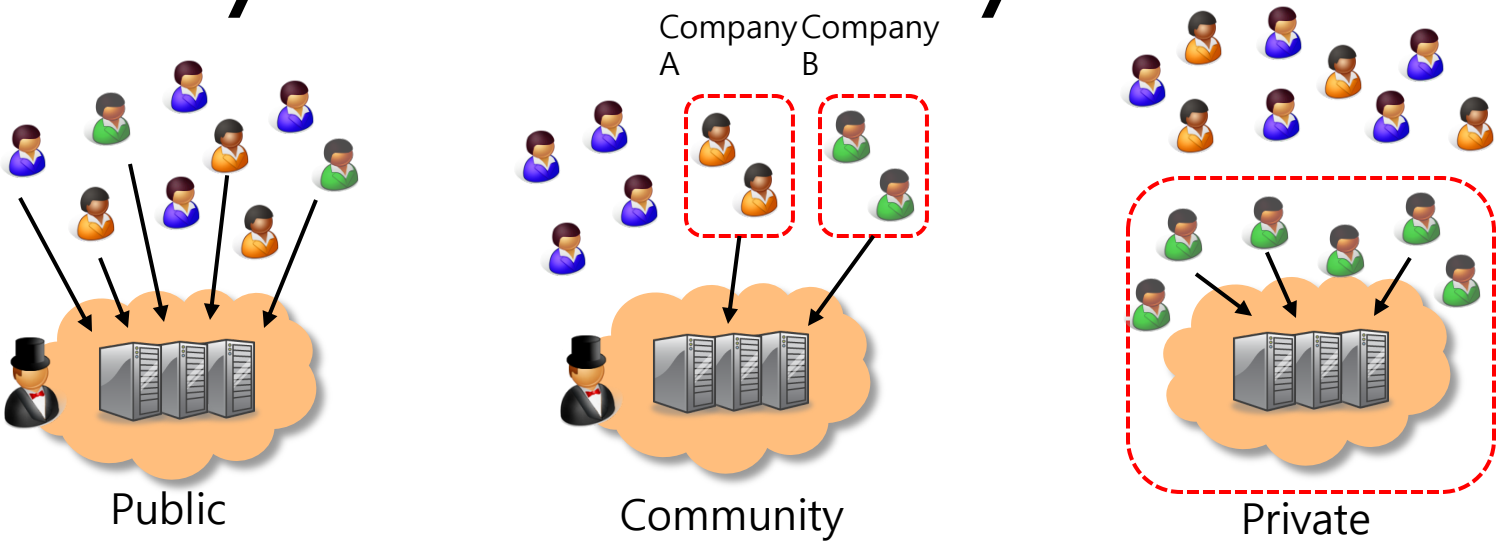
- Cloud provides middleware/infrastructure
  - For example, Microsoft Common Language Runtime (CLR)
  - Customer pays SaaS provider for the service; SaaS provider pays the cloud for the middleware/infrastructure
  - Example: Windows Azure(it is different now), Google App Engine

Infrastructure as a Service (IaaS)



- Cloud provides raw computing resources
  - Virtual machine, blade server, hard disk, …
  - Customer pays SaaS provider for the service; SaaS provider pays the cloud for the hardwares
  - Examples: Amazon Web Services, Rackspace Cloud, etc

Private/hybrid/community clouds



- Who can become a customer of the cloud?
  - Public cloud: Commercial service; open to (almost) anyone. Example: Amazon AWS, Microsoft Azure, Google Cloud Platform
  - Community cloud: Shared by several similar organizations. Example: Google's "Gov Cloud"
  - Private cloud: Shared within a single organization. Example: Internal datacenter of a large company.

Focus of this class →

Plan for today

- Large-scale computing
  - The need for scalability
  - Scale of current services
  - Scaling up : From PCs to data centers
  - Problems with 'classical' scaling techniques

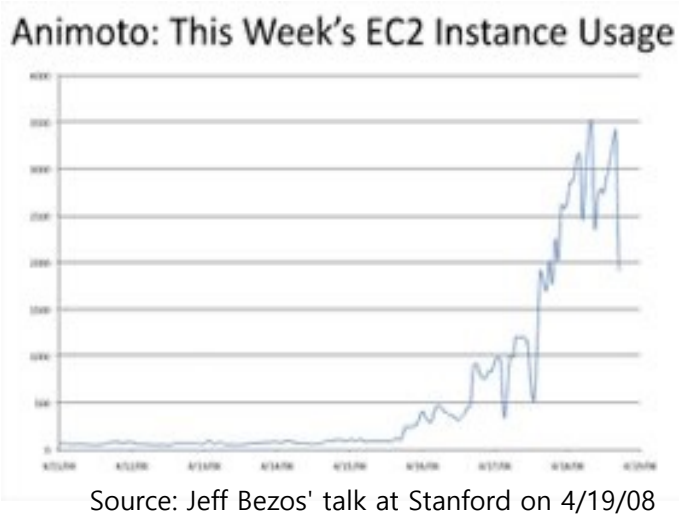
- Utility computing and cloud computing
  - What are utility computing and cloud computing?
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud? Here
  - Virtualization : How does cloud computing actually work?
  - Some cloud computing challenges

Examples of cloud applications

- Application hosting
- Backup and Storage
- Content delivery
- E-commerce
- High-performance computing
- Media hosting
- On-demand workforce
- Search engines
- Big data analytics
- etc

Case study: animoto

- Animoto: Lets users create videos from their own photos/music
  - Auto-edits photos and aligns them with the music, so it "looks good"



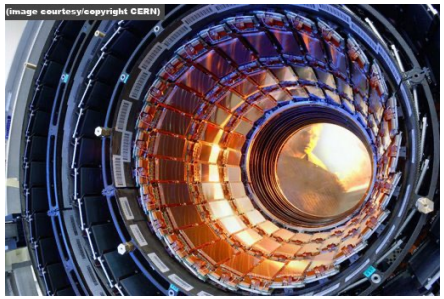
Source: Jeff Bezos' talk at Stanford on 4/19/08

- Built using Amazon EC2+S3+SQS
- Released a Facebook app in mid-April 2008
  - More than 750,000 people signed up within 3 days
  - EC2 usage went from 50 machines to 3,500 (x70 scalability!)

http://animoto.com/blog/company/amazon-com-ceo-jeff-bezos-on-animoto/

Other examples

- DreamWorks is using the Cerelink cloud to render animation movies
  - Cloud was already used to render parts of Shrek Forever After and How to Train your Dragon
- CERN is working on a "Science cloud" to process experimental data
- Virgin atlantic is hosting their new travel portal on Amazon AWS



Recap: Utility/Cloud Computing

- Why is cloud computing attractive?
  - Analogy to 'classical' utilities (electricity, water, …)
  - No up-front investment (pay-as-you-go model)
  - Low price due to economics of scale
  - Elasticity – can quickly scale up/down as demand varies
- Different types of clouds
  - SaaS, PaaS, IaaS; public/private/community/hybrid clouds
- What runs on the cloud?
  - Many potential applications : Application hosting, backup/storage, scientific computing, content delivery, …
  - Not yet suitable for certain applications (sensitive data, compliance requirements)

Is the cloud good for everything?

- No.
- Sometimes it is problematic, e.g., because of auditability requirements
- Example : Processing medical records
  - HIPAA (Health Insurance Portability and Accountability Act) privacy and security rule
- Example : Processing financial information
  - Sarbanes-Oxley act
- Would you put your medical data on the cloud?
  - Why / why not?

Recap: Cloud applications

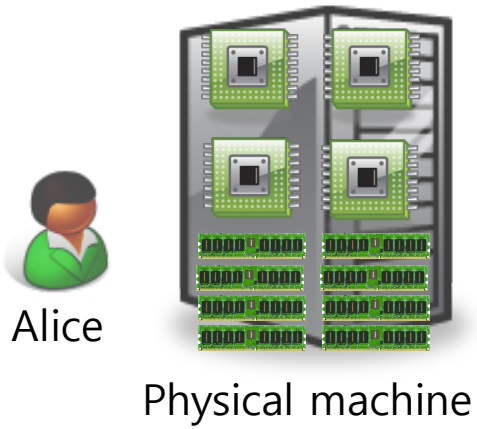
- Clouds are good for many things…
  - Applications that involve large amounts of computation, storage, bandwidth
  - Especially when lots of resources are needed quickly (Washington Post example).
- … But not for all things

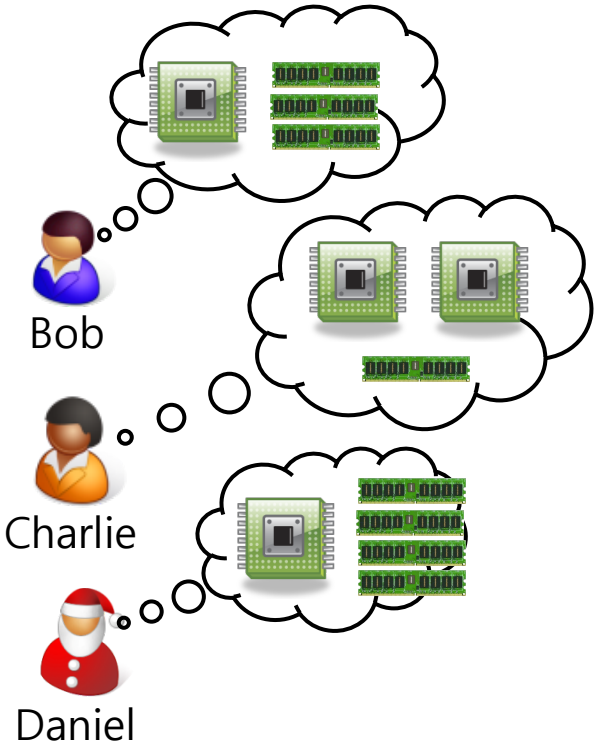
Plan for today

- Large-scale computing
  - The need for scalability
  - Scale of current services
  - Scaling up : From PCs to data centers
  - Problems with 'classical' scaling techniques

- Utility computing and cloud computing
  - What are utility computing and cloud computing?
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization : How does cloud computing actually work? Here
  - Some cloud computing challenges

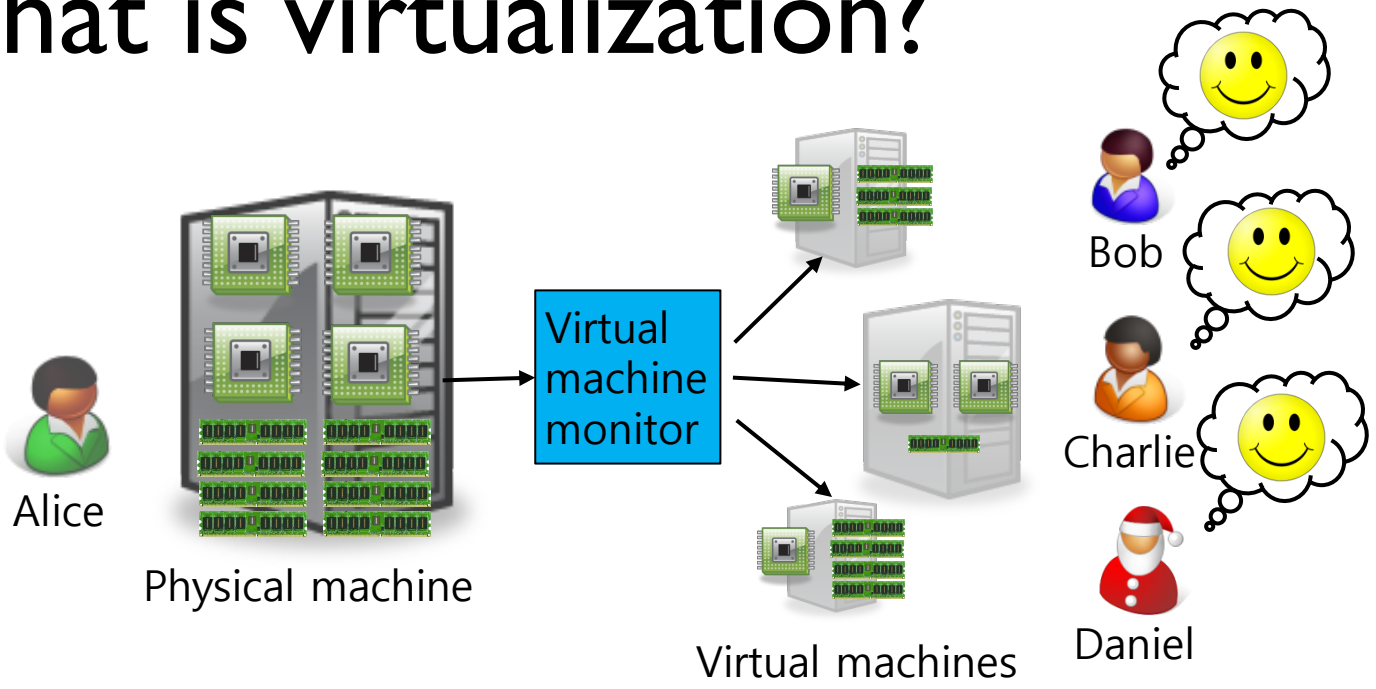
What is virtualization?





- Suppose Alice has a machine with 4 CPUs and 8 GB of memory, and three customers:
  - Bob wants a machine with 1 CPU and 3GB of memory
  - Charlie wants 2 CPUs and 1GB of memory
  - Daniel wants 1 CPU and 4GB of memory
- What should Alice do?

What is virtualization?

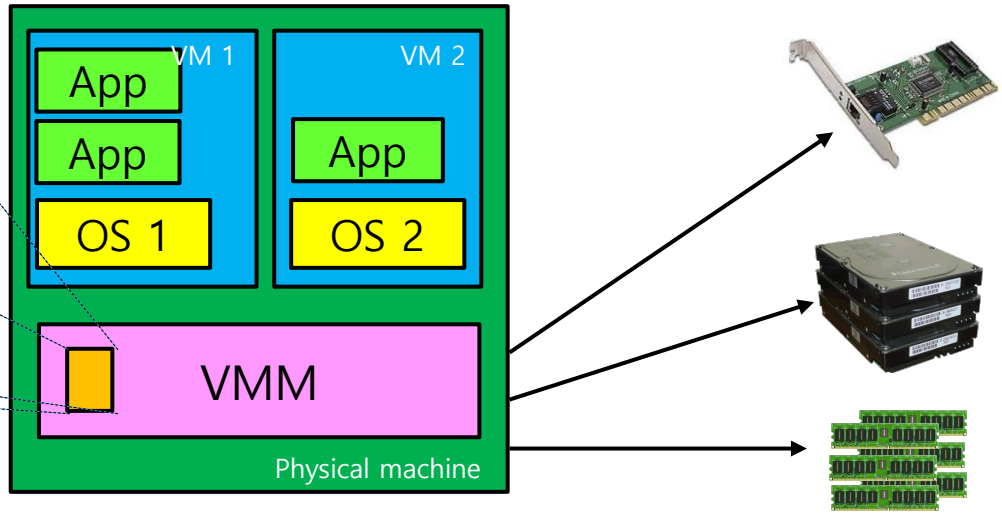


- Alice can sell each customer a virtual machine(VM) with the requested resources
  - From each customer's perspective, it appears as if they had a physical machine all by themselves (isolation)

How does it work?

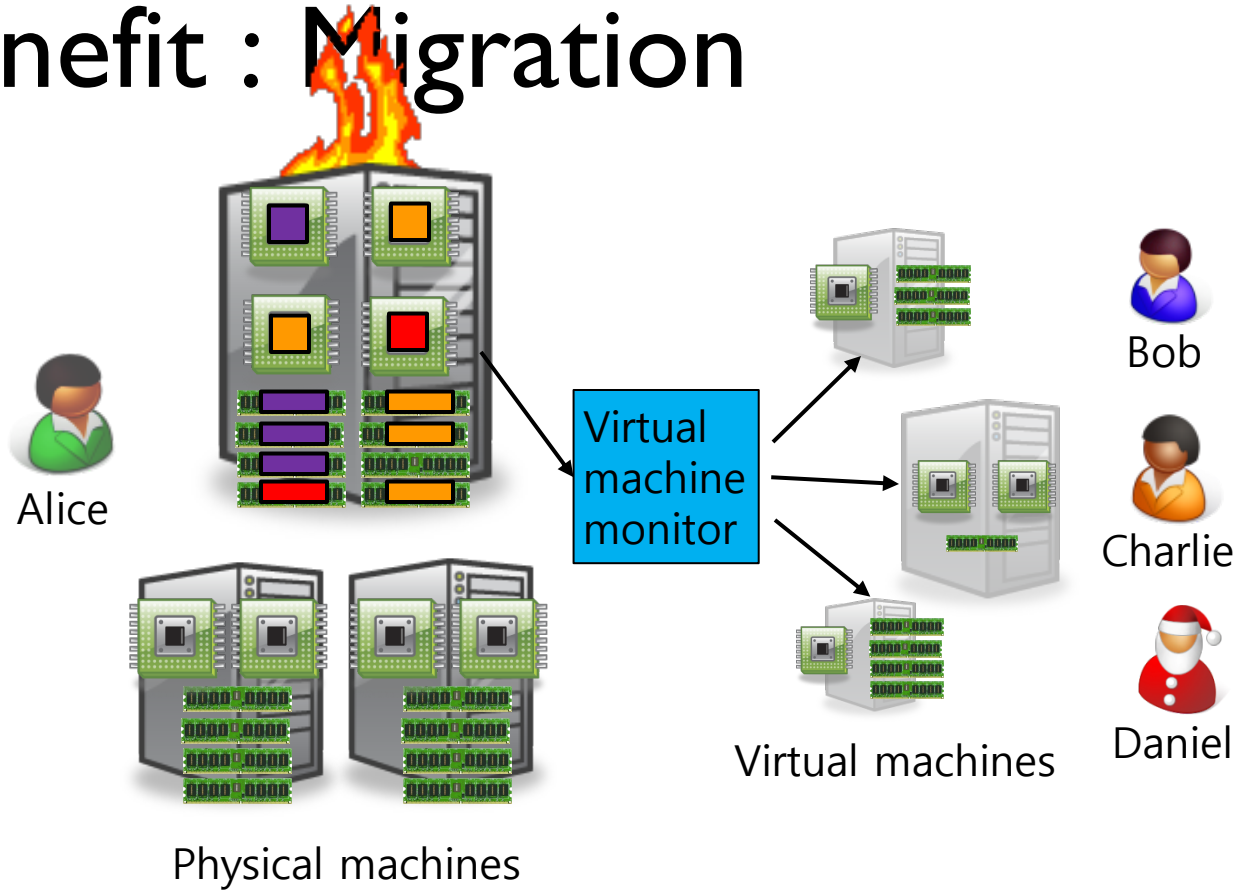
VM	Virt	Phys
1	0-99	0-99
1	299-399	100-199
2	0-99	300-399
2	200-299	500-599
2	600-699	400-499

Translation table



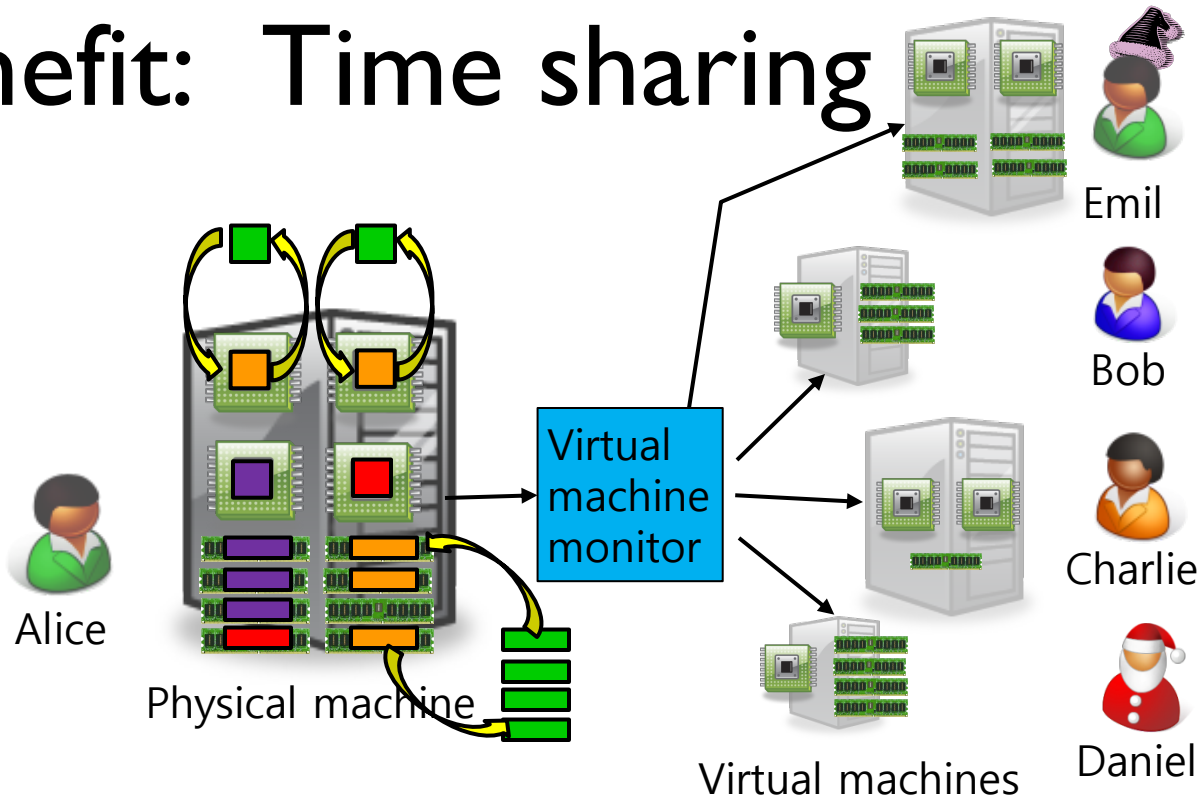
- Resources (CPU, memory, ...) are virtualized
  - VMM ("Hypervisor") has translation tables that map requests for virtual resources to physical resources
  - Example: VM 1 accesses memory cell #323; VMM maps this to memory cell 123.
  - For which resources does this (not) work?
  - How do VMMs differ from OS kernels?

Benefit : Migration



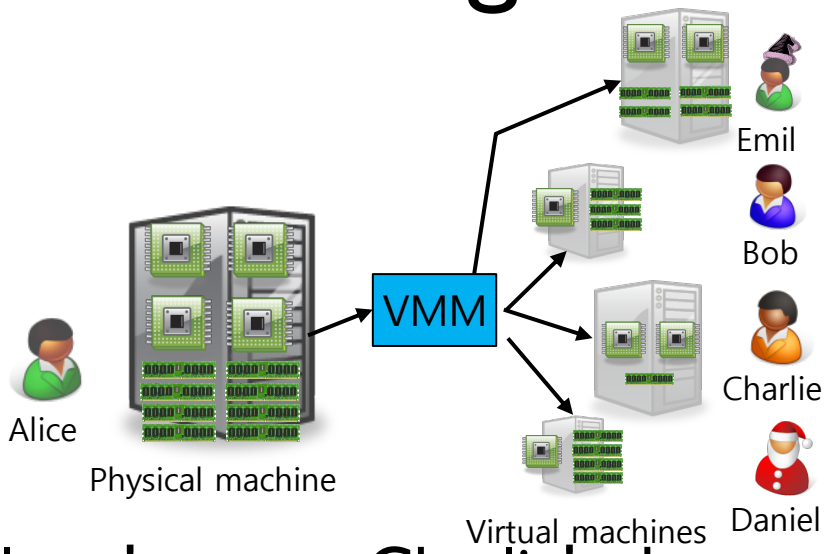
- What if the machine needs to be shut down?
  - e.g., for maintenance, consolidation, ...
  - Alice can migrate the VMs to different physical machines without any customers noticing

Benefit: Time sharing



- What if Alice gets another customer?
  - Multiple VMs can time-share the existing resources
  - Result: Alice has more virtual CPUs and virtual memory than physical resources (but not all can be active at the same time)

Benefit and challenge: Isolation



- Good: Emil can't access Charlie's data
- Bad: What if the load suddenly increases?
  - Example: Emil's VM shares CPUs with Charlie's VM, and Charlie suddenly starts a large compute job
  - Emil's performance may decrease as a result
  - VMM can move Emil's software to a different CPU, or migrate it to a different machine

Recap: Virtualization in the cloud

- Gives cloud provider a lot of flexibility
  - Can produce VMs with different capabilities
  - Can migrate VMs if necessary (e.g., for maintenance)
  - Can increase load by overcommitting resources
- Provides security and isolation
  - Programs in one VM cannot influence programs in another
- Convenient for users
  - Complete control over the virtual 'hardware' (can install own operating system, own applications, ...)
- But: Performance may be hard to predict
  - Load changes in other VMs on the same physical machine may affect the performance seen by the customer

Plan for today

- Large-scale computing
  - The need for scalability
  - Scale of current services
  - Scaling up : From PCs to data centers
  - Problems with 'classical' scaling techniques

- Utility computing and cloud computing
  - What are utility computing and cloud computing?
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization : How does cloud computing actually work?
  - Some cloud computing challenges

10 obstacles and opportunities

1. Availability
   - What happens to my business if there is an outage in the cloud?

2. Data lock-in
   - How do I move my data from one cloud to another?

Service	Duration	Date
S3	6-8 hrs	7/20/08
AppEngine	5 hrs	6/17/08
Gmail	1.5 hrs	8/11/08
Azure	22 hrs	3/13/09
Intuit	36 hrs	6/16/10
EBS	>3 days	4/21/11
EC2	~2 hrs	6/30/12

Some prominent cloud outages

3. Data confidentiality and auditability
   - How do I make sure that the cloud doesn't leak my confidential data?
   - Can I comply with regulations like HIPAA and Sarbanes/Oxley?

10 obstacles and opportunities

4. Data transfer bottlenecks

- How do I copy large amounts of data from/to the cloud?
- Example: 10 TB from UC Berkeley to Amazon in Seattle, WA
- Motivated Import/Export feature on AWS

Method	Time
Internet (20Mbps)	45 days
FedEx	1 day

Time to transfer 10TB

5. Performance unpredictability

- Example: VMs sharing the same disk → I/O interference
- Example: HPC tasks that require coordinated scheduling

Primitive	Mean perf.	Std dev
Memory bandwidth	1.3GB/s	0.05GB/s (4%)
Disk bandwidth	55MB/s	9MB/s (16%)

Performance of 75 EC2 instances in benchmarks

10 obstacles and opportunities

6. Scalable storage
    - Cloud model (short-term usage, no up-front cost, infinite capacity on demand) does not fit persistent storage well
7. Bugs in large distributed systems
    - Many errors cannot be reproduced in smaller configs
8. Scaling quickly
    - Problem: Boot time; idle power

10 obstacles and opportunities

9. Reputation fate sharing
    - One customer's bad behavior can affect the reputation of others using the same cloud
    - Example: Spam blacklisting
10. Software licensing
    - What if licenses are for specific computers?
        - Example: Microsoft Windows
    - How to scale number of licenses up/down?
        - Need pay-as-you-go model as well

Plan for today

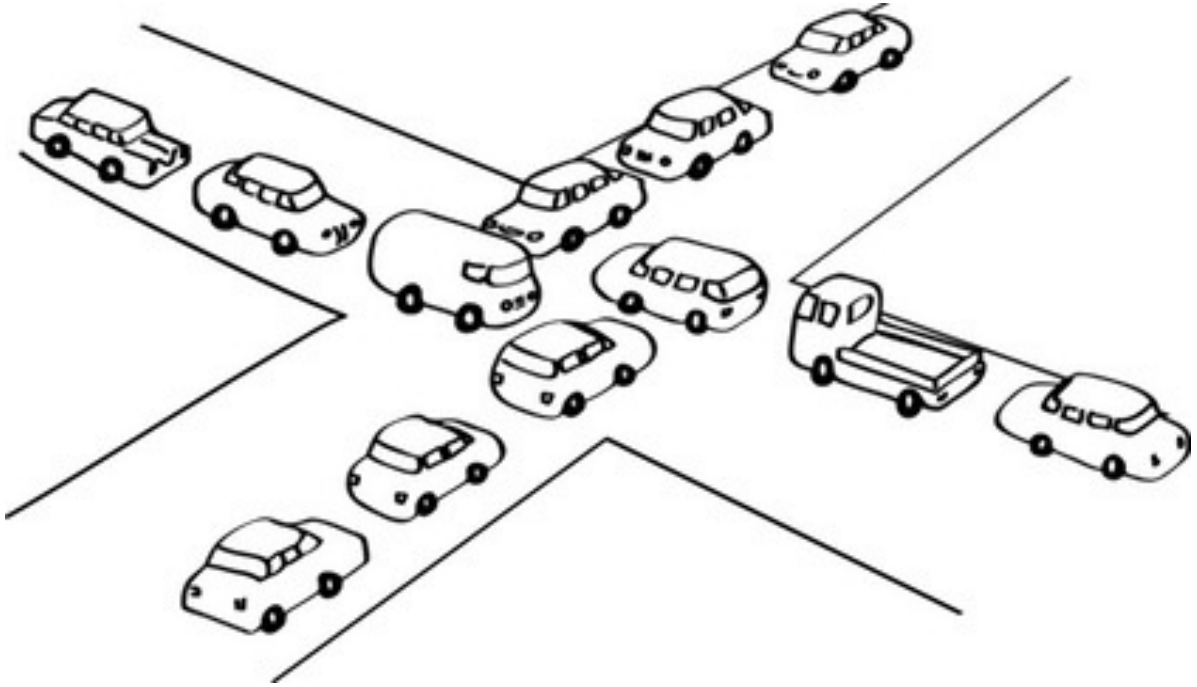
- Large-scale computing
  - The need for scalability ✓
  - Scale of current services ✓
  - Scaling up : From PCs to data centers ✓
  - Problems with 'classical' scaling techniques ✓

- Utility computing and cloud computing
  - What are utility computing and cloud computing? ✓
  - What kinds of clouds exist today? ✓
  - What kinds of applications run on the cloud? ✓
  - Virtualization : How does cloud computing actually work? ✓
  - Some cloud computing challenges ✓

For next time

- Homework submission due policy
  - Submit it by the day before the next class.
  - For example, if the next class is 12.25, submit it by 12.24 11:59 PM.
- Homework # 1
  - Read the Armbrust et al, paper "A View of Cloud Computing"
  - http://bit.ly/2xIdgyv
  - Submit a review of the article into 1 A4-sized paper to ecampus
- Homework # 1-2(no submission)
  - Learn git & python
  - You need to know git and python in order to carry out the project in this course.

Next Time



- Next time you will learn about :
  - Programming at scale; Concurrency and Consistency

Any Questions?

Credits & References

- Credits : A. Haeberlen, Z. Ives (University of Pennsylvania)