UCL
Université catholique de Louvain

INGI2145: CLOUD COMPUTING (Fall 2015)

The Cloud

17 September 2015

Lecture slides adapted from Upenn NETS212 by A. Haeberlen, Z. Ives
Reproduced with permission

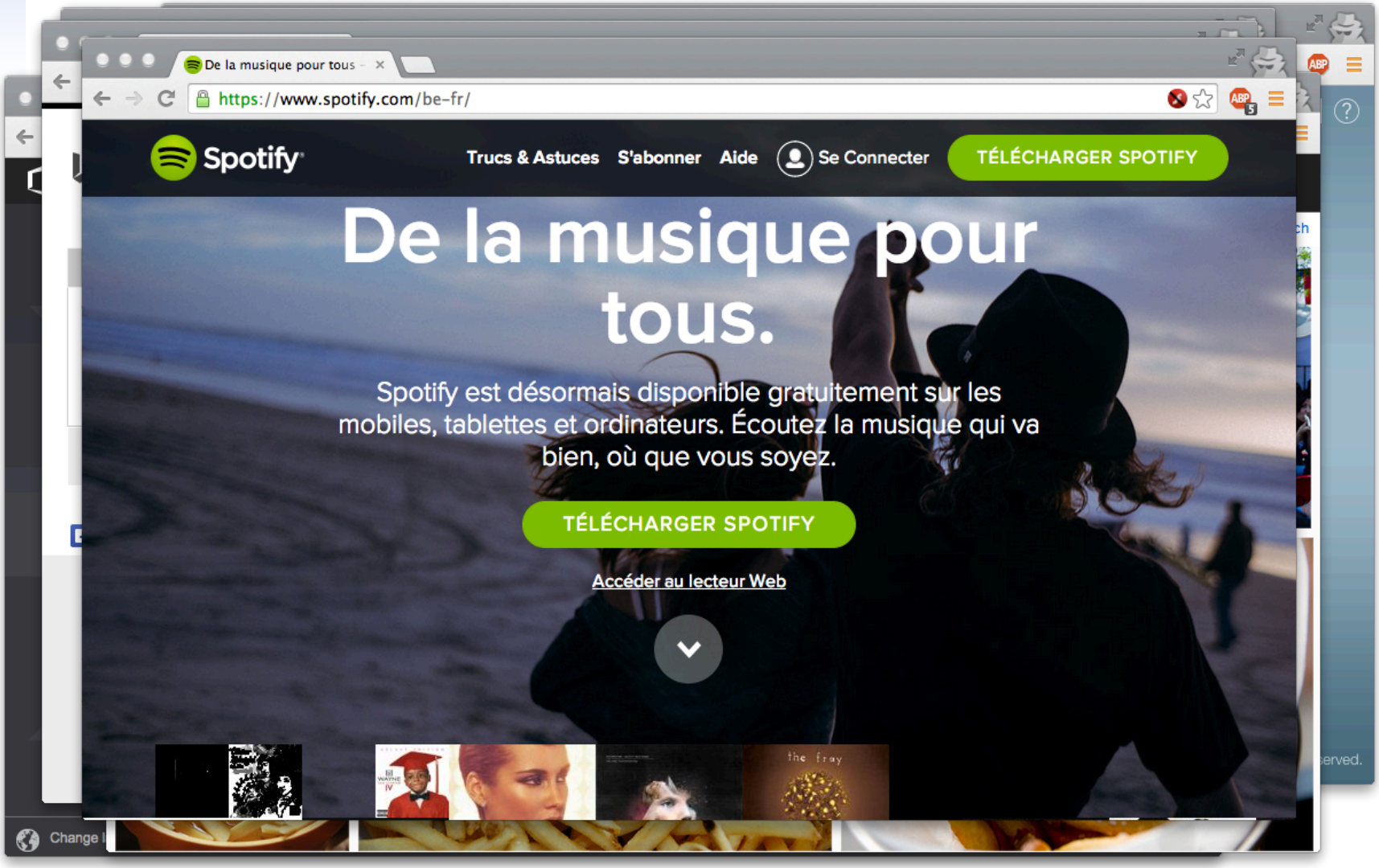
© 2015 M. Canini

Welcome!

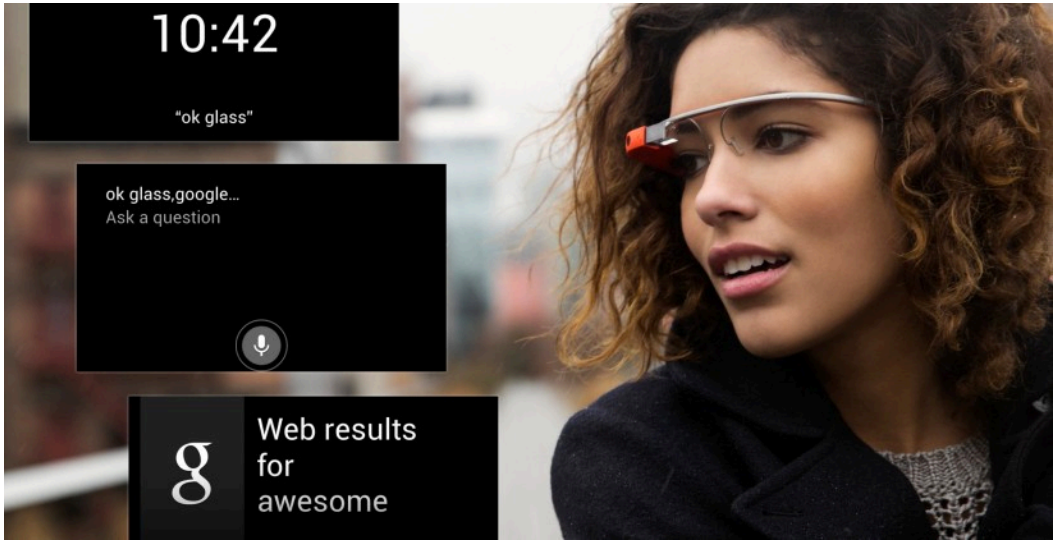
- My name: Marco Canini

- Faculty member at INGI
- Research interests:
  - Large-scale Distributed Systems
  - Cloud Computing
  - Computer Networks
  - See also: http://perso.uclouvain.be/marco.canini/

Have you used 'the cloud' before?



The cloud in your pocket and beyond



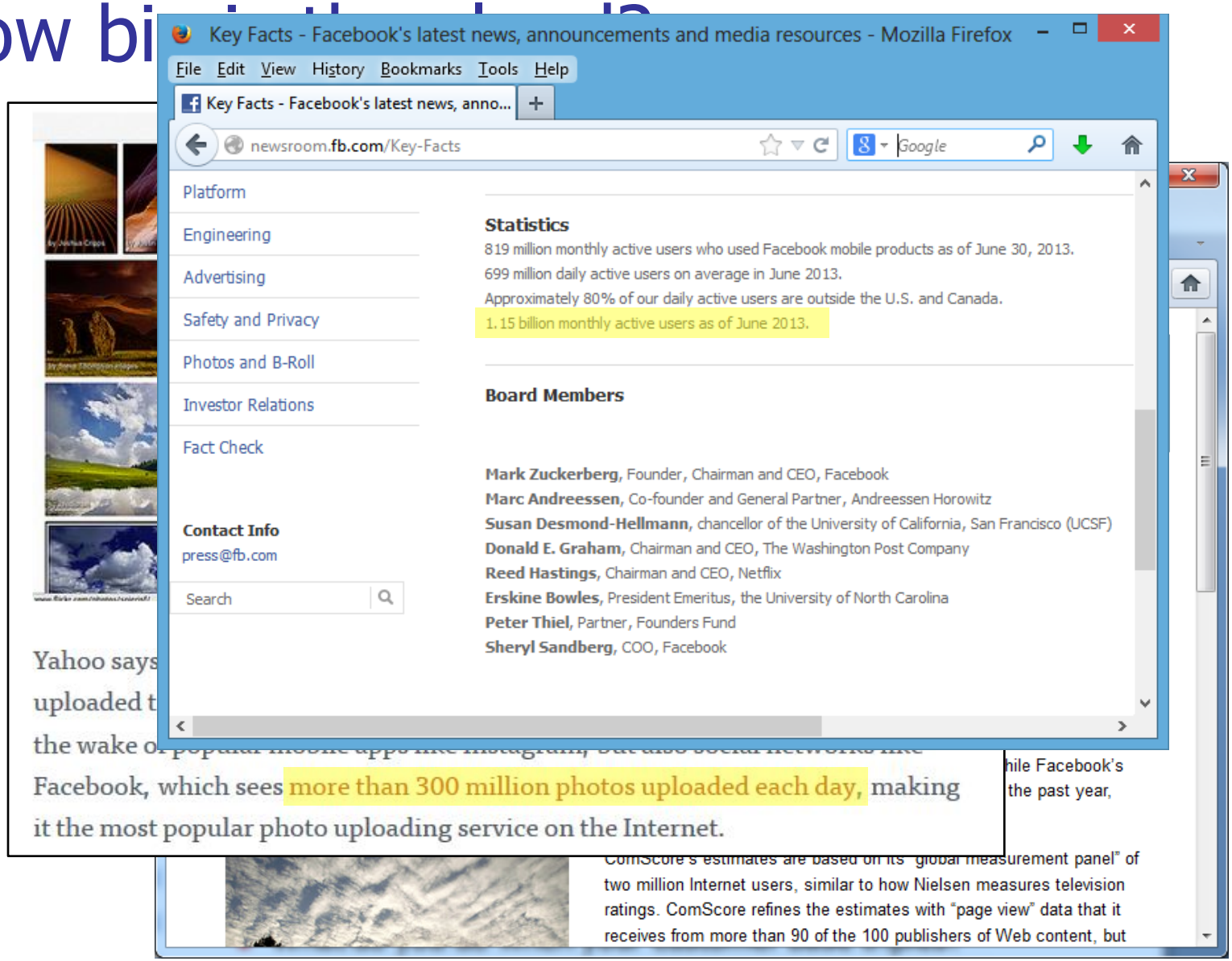
Why should I care?

- Understand what's underneath the Cloud
  - How does it work? What are its strengths? Its shortcomings?
  - Technologies: virtualization, MapReduce, key-value store, NoSQL, Ajax, ...
- Understand the underlying principles
  - How do you build something that is so scalable, robust, etc.?
  - Lots of clever algorithms needed - very different environment!
- Be able to use the right approach when designing new protocols and web systems
  - How would you go about building the next Facebook?
  - Need to scale, be efficient, avoid failures, ...

Why should I care? (continued)

- Gain practical experience with cloud technologies
  - Often, the best way to understand it is to build one yourself
  - In this course, you will build on top of Amazon Web Services and Apache Hadoop, Spark, and more…
- Understand the impact on society
  - Vulnerabilities, privacy concerns, data survivability, ...
  - Need to understand the current state of the technology!
- Anticipate what's possible in the future

How bi


Key Facts - Facebook's latest news, announcements and media resources - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Key Facts - Facebook's latest news, anno...
newsroom.fb.com/Key-Facts
Google
Platform
Engineering
Advertising
Safety and Privacy
Photos and B-Roll
Investor Relations
Fact Check
Contact Info
press@fb.com
Search
Statistics
819 million monthly active users who used Facebook mobile products as of June 30, 2013.
699 million daily active users on average in June 2013.
Approximately 80% of our daily active users are outside the U.S. and Canada.
1.15 billion monthly active users as of June 2013.
Board Members
Mark Zuckerberg, Founder, Chairman and CEO, Facebook
Marc Andreessen, Co-founder and General Partner, Andreessen Horowitz
Susan Desmond-Hellmann, chancellor of the University of California, San Francisco (UCSF)
Donald E. Graham, Chairman and CEO, The Washington Post Company
Reed Hastings, Chairman and CEO, Netflix
Erskine Bowles, President Emeritus, the University of North Carolina
Peter Thiel, Partner, Founders Fund
Sheryl Sandberg, COO, Facebook
Yahoo says
uploaded t
the wake o
Facebook, which sees more than 300 million photos uploaded each day, making
it the most popular photo uploading service on the Internet.
hile Facebook's
the past year,
two million Internet users, similar to how Nielsen measures television
ratings. ComScore refines the estimates with "page view" data that it
receives from more than 90 of the 100 publishers of Web content, but

Data-centric computing

- Trend towards data-centric computing
  - Two words: "Big data"
- Today's currency on the Internet is data!
  - You "pay" for using Google, Facebook, etc. by letting them record your every action, link, search, etc.

- But data's value is not just economic:
  - It allows us to better answer questions, understand what's important, validate hypotheses about social interactions, …
  - Example: Online Social Network research

© 2015 M. Canini

Data-centric computing is pervasive

- Today, Google and Friends aren't the only "Big Data" players
  - Not just Google & friends - banks, financial firms, academia, the government, companies, military, startups, ...
  - All need to store and analyze huge data volumes

- This is being enabled with a new generation of hardware “hosting” services – “the cloud” – and programming models

© 2015 M. Canini

What is INGI2145 about?

- How do we build effective data-centric applications, and serve them to the entire Internet?
  - You've learned procedural programming on a single machine – we'll look at data-centric programming across thousands of machines
  - We'll understand the issues in breaking up problems, global coordination, failures, and so on
  - We'll study many of the systems and algorithms used by real Internet services
- How do we take advantage of "the cloud" – the vision of computing as a utility (like the power grid)?
  - You'll understand what lies underneath the cloud computing hype, and how to use the cloud
  - You'll build real projects hosted "on the cloud"

© 2015 M. Canini

Towards understanding a larger trend

- Internet services are increasingly integrated into the fabric of our society
  - Communication – Twitter, Facebook, Skype, IM, …
  - Media – iTunes, Spotify, Netflix, …
  - Markets – Amazon, eBay, stock exchanges, advertising
  - Utilities, commodities – smart power grids, exch. markets
- Cloud computing likely to have profound implications on economical, social, ethical and legal matters
  - Data-centric, quantitative methods are revolutionizing advertising, sales but also science

© 2015 M. Canini

Plan for today

- Introduction 
- Course logistics 

- The Cloud

© 2015 M. Canini

Course Staff

- Marco Canini
  - Réaumur A-049, x7-4832

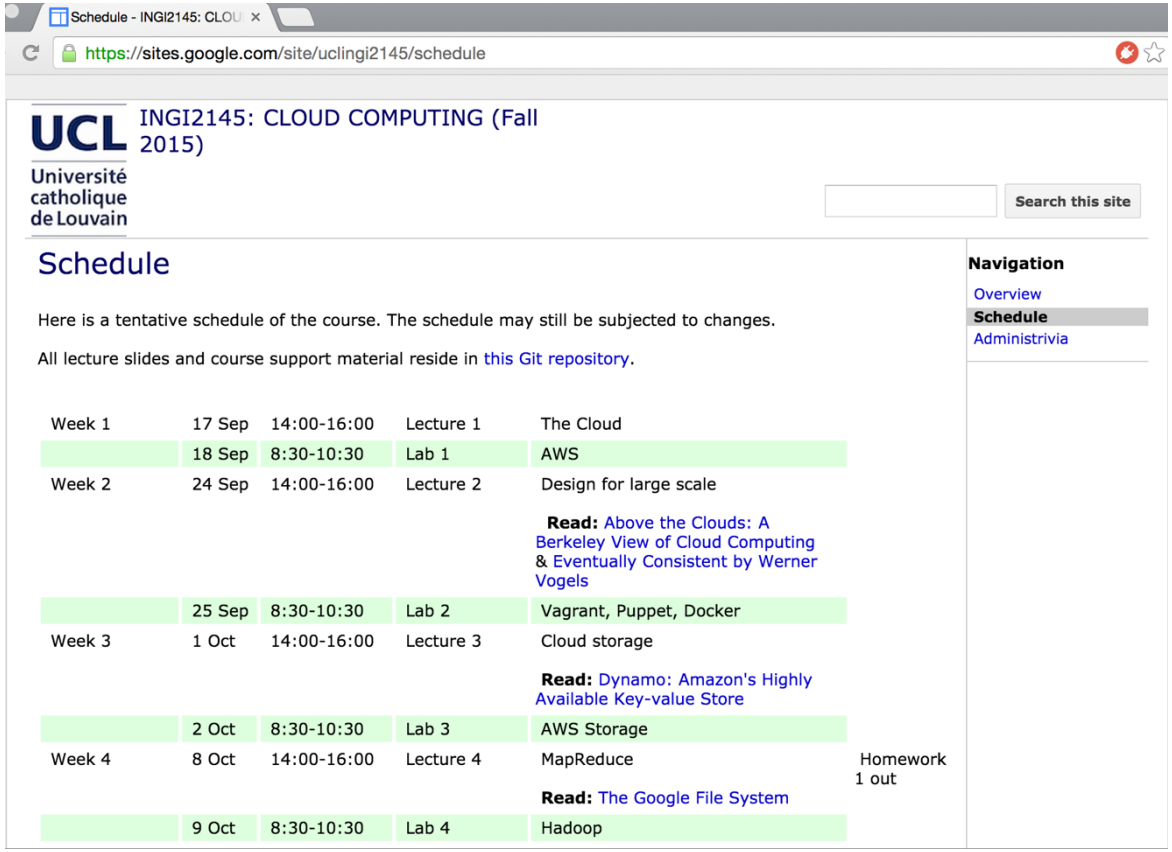
- Bilal (TA)
  - Réaumur A-027

- Thanh Nguyen (TA)
  - Réaumur A-043

Course website

https://sites.google.com/site/uclingi2145/

- All official communication is there



Course discussion group

- We will be using piazza for discussions related to this course
  - Examples: Questions about homework assignments
  - The TAs and I will read the posts and respond to questions
- Piazza will also be used for
  - Announcements, e.g., cancelled classes (if necessary)
  - Supplemental materials, e.g., links to relevant papers
  - Corrections/clarifications, e.g., bugs in homework handouts
  - Please check the group frequently!
- Please sign up at
  - https://piazza.com/uclouvain.be/fall2015/ingi2145

Lectures

- Regularly Thursdays 14:00 to 16:00, BARB 00
  - Could exceptionally be on Fridays 8:30 to 10:30, BARB 21
  - Check online schedule often
- Be on time!
- Course slides published in the git repo at
  - https://github.com/mcanini/INGI2145-2015/
- Most lecture have assigned readings
  - Must be prepared in time for the lecture
  - Evaluated as part of participation grade

© 2015 M. Canini

Lab sessions

- We will organize ~7-8 lab sessions
- Complement the lectures with more practical information and hands-on exercises
- A sketch solution of the exercises will be presented during the session
- Labs on Fridays 8:30 to 10:30, BARB 21
- Attendance is highly encouraged

- First lab is tomorrow: intro to AWS!

Homework assignments

- We will organize 2-3 homework assignments
  - They are all mandatory
- No extensions granted
  - Automatic lateness penalty applies: 10% deducted on the assignment grade for each day late or fraction thereof
    - If an assignment is 1 minute late, it is one day late
- Start to work on them early
  - Do not wait until the last day before homework is due
- Deploy and run code on Amazon EC2
  - But credit is limited. Don't blow it up!
- We will offer a standardized development system that is based on a virtual machine

© 2015 M. Canini

The INGI2145 Virtual Machine

- We will provide a virtual machine with the necessary software
  - Linux, Hadoop, Amazon CLI, …
- Use with VMware Player, VirtualBox, etc.
  - Safe to experiment with
  - Standardized environment makes it easier to support
  - We will not support custom environments
- Second lab session
  - Demonstrate how to provision the virtual machine through Vagrant and Puppet, or Docker

Grading

- Final exam: **60%**
- Homework: **30%**
- Participation: **10%**

- Note: Homework grade and participation grade carry to the second examination session (September) as they cannot be repeated
- Grading policy is the same for first and second examination session

Participation counts towards grade

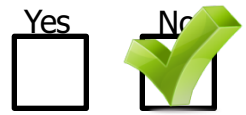
- Prepare mandatory readings for classes and complete paper evaluation quizzes
- Engage in discussion, make insightful questions or comments during class
  - (also on Piazza)
- May organize in-class quizzes on material covered in prior classes

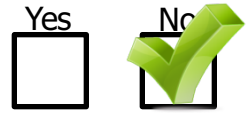
Policies: Collaboration

- All assignments must be done individually
  - All the code you submit has to be your own
    - Only exception: Code we have provided or explicitly authorized
  - UCL's regulations applies
    - No cheating, plagiarism, fabrication, multiple submissions, gaining an unfair advantage, or facilitating (!) academic dishonesty
    - It's not worth it!! Penalties can be severe: http://www.uclouvain.be/enseignement-reglements.html
- Zero tolerance policy to ensure fairness
  - We will use various tools to actively look for cheating

Policies: Collaboration

- Can we work on assignments together?
- Can I discuss the assignment with others (in general terms)?
- Can I use code I copied from the web?
- Can I ask questions about the assignments on Piazza?
- I just happened to leave my password on my table, and XYZ just happened to find it. Will I be penalized for this?





Expected 'payoff'

- You will acquire a set of skills that is in very high demand right now
  - At Google, Facebook, and at many other places
- You will gain interesting insights
- You will have a good basis for other courses

A disclaimer…

- This is a "bleeding edge" course!
  - UCL is one of a handful of places offering these topics
  - The subject of this course is still evolving: no established curriculum, no classical textbooks yet
- Some of the material in the course will result in hair loss
  - Debugging distributed code is hard!
- We will be using some immature technology
  - We will do the best we can to smooth over the bugs
- I hope it will be a fun course, though…
  … and an interesting one!

© 2015 M. Canini

Plan for today

- Introduction 
- Course logistics 
- The Cloud 

© 2015 M. Canini

Plan for today

- Computing at scale
  - 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?
  - Evolution of software business models
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization: How clouds work 'under the hood'
  - Some cloud computing challenges

© 2015 M. Canini

How many users and objects?

- Flickr has >8 billion photos
- Facebook has 1.15 billion active users
- Google is serving >1.2 billion queries/day on more than 27 billion items
- >2 billion videos/day watched on YouTube

© 2015 M. Canini

How much data?

- Modern applications use massive data:
  - Rendering 'Avatar' movie required >1 petabyte of storage
  - eBay has >6.5 petabytes of user data
  - CERN's LHC will produce about 15 petabytes of data per year
  - In 2008, Google processed 20 petabytes per day
  - German Climate computing center dimensioned for 60 petabytes of climate data
  - Google now designing for 1 exabyte of storage
  - 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 1.2 billion queries of this type every day?
- Suppose you want to build the next Facebook
- How do you...
  - ... store the profiles and photos of over 1 billion users?
  - ... avoid losing any of them?
  - ... find out which users might want to be friends?
- Stay tuned!

© 2015 M. Canini

Plan for today

- Computing at scale
  - 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?
  - Evolution of software business models
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization: How clouds work 'under the hood'
  - Some cloud computing challenges

© 2015 M. Canini

Scaling up

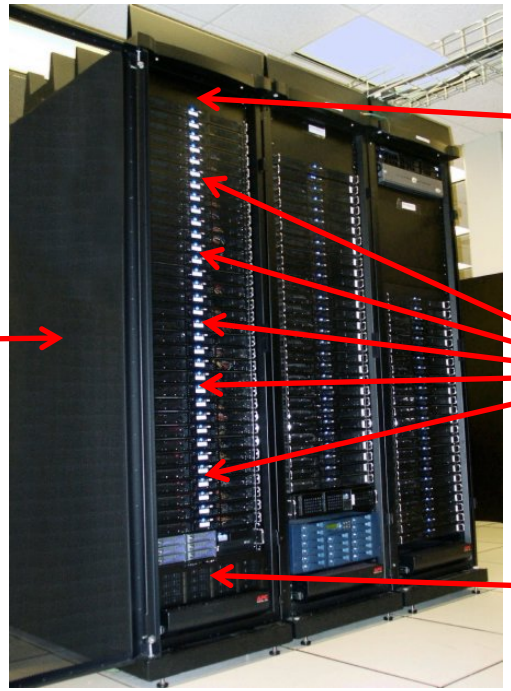
PC

Server

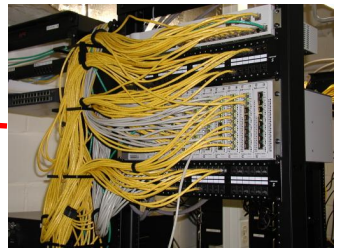
Cluster

- 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 (same room?)
  - 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

- Large clusters need massive cooling
  - 4.5kW is about 3 space heaters
  - And that's just one rack!

Scaling up

PC

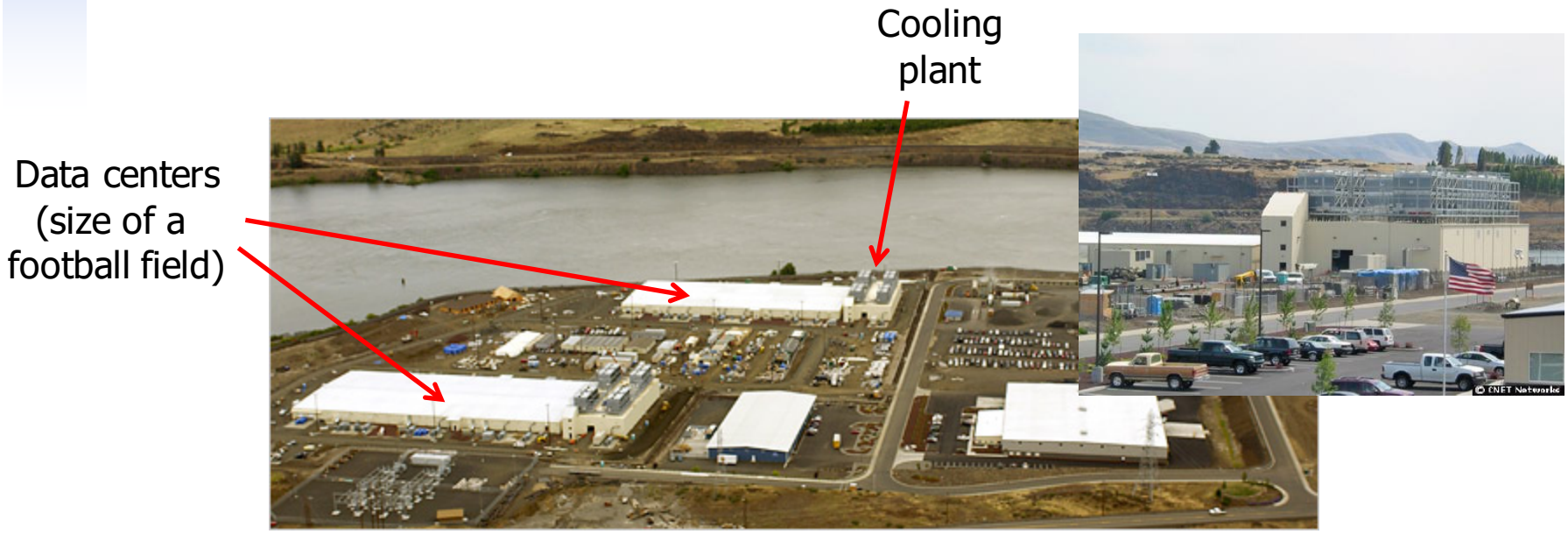
Server

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

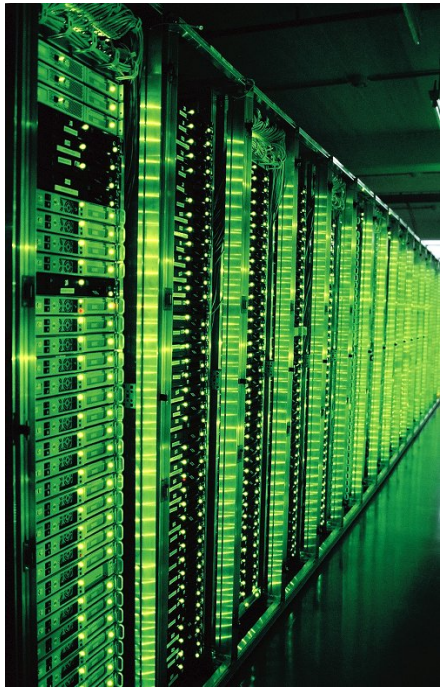
What does a data center look like?



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)

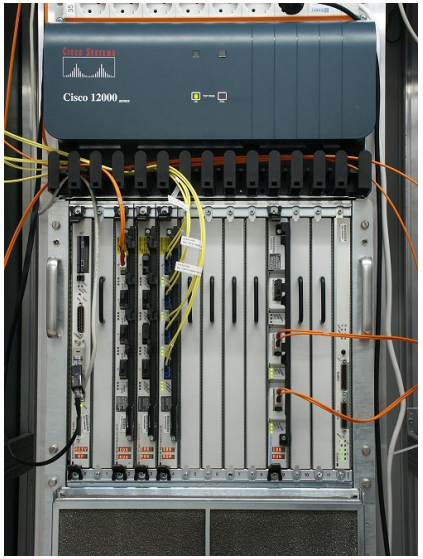
What's in a data center?



Source: 1&1

- Hundreds or thousands of racks

What's in a data center?



Source: 1&1

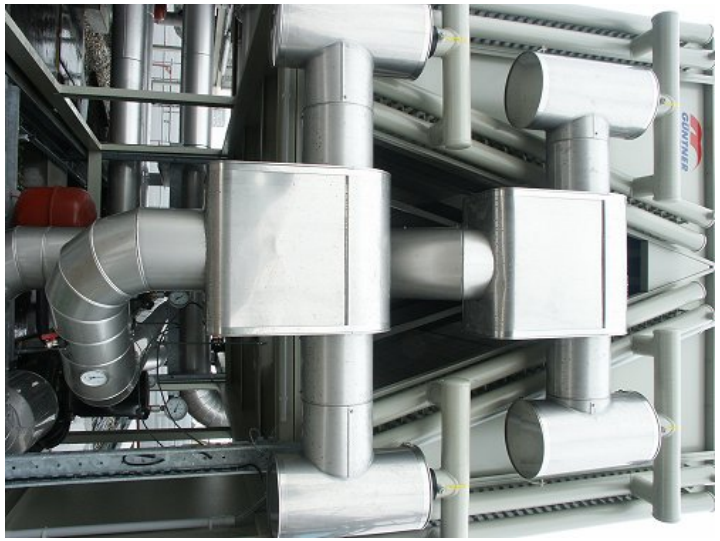
- Massive networking

What's in a data center?

Source: 1&1

- Emergency power supplies

What's in a data center?



Source: 1&1

- Massive cooling

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 consume 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!

© 2015 M. Canini

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
  - Example above: Google data center locations (inferred)
- Why?
  - Need to be close to users (physics!)
  - Cheaper resources
  - Protection against failures

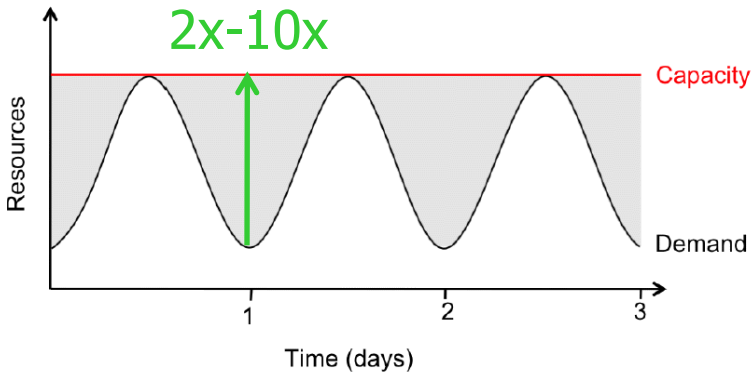
Plan for today

- Computing at scale
  - The need for scalability; scale of current services 
  - Scaling up: From PCs to data centers 
  - Problems with 'classical' scaling techniques 

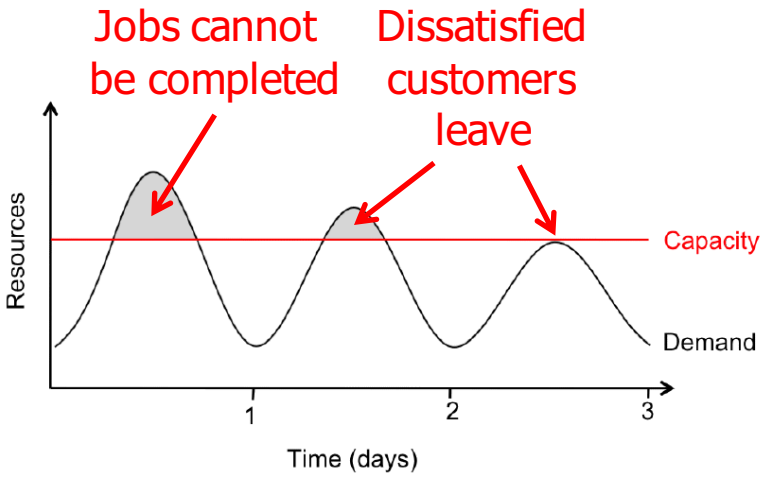
- Utility computing and cloud computing
  - What are utility computing and cloud computing?
  - Evolution of software business models
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization: How clouds work 'under the hood'
  - Some cloud computing challenges

© 2015 M. Canini

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
  - Microsoft recently invested $499 million in a single data center
- 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, ...

© 2015 M. Canini

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 may require major redesign, e.g., new storage system, new interconnect, new building (!)
- Scaling down is difficult
  - What to do with superfluous 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: Computing at scale

- Modern applications require huge amounts of processing and data
  - Measured in petabytes, millions 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

© 2015 M. Canini

Plan for today

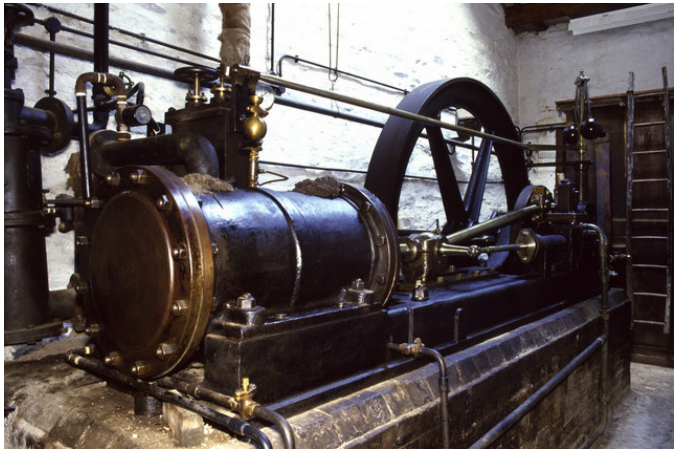
- Computing at scale
  - The need for scalability; scale of current services 
  - Scaling up: From PCs to data centers 
  - Problems with 'classical' scaling techniques 

- Utility computing and cloud computing
  - What are utility computing and cloud computing? 

  - Evolution of software business models
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization: How clouds work 'under the hood'
  - 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...

© 2015 M. Canini

Scaling the power plant




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

© 2015 M. Canini

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

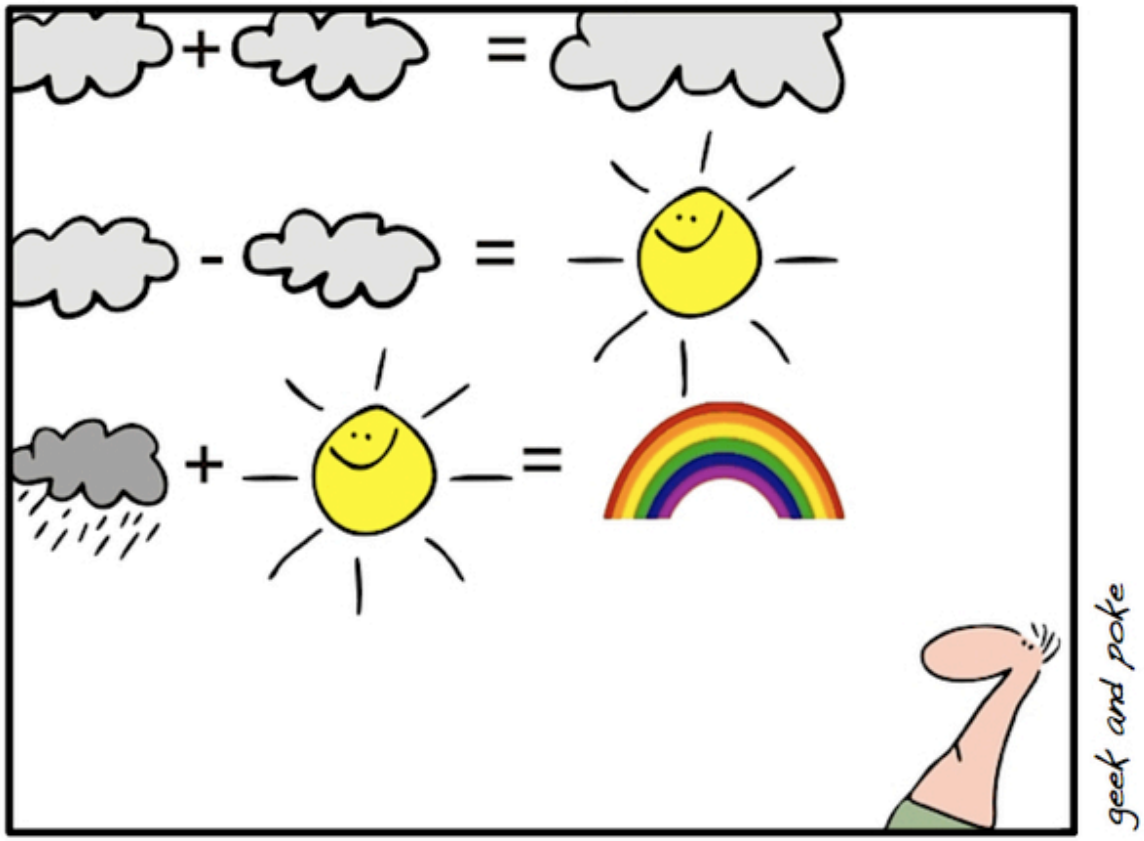
© 2015 M. Canini

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 data center; pay-as-you-go model
Scalability	
Thousands of kilowatts available on demand; add more within seconds	Thousands of computers available on demand; add more within seconds

© 2015 M. Canini

What is cloud computing?



SIMPLY EXPLAINED – PART 17:
CLOUD COMPUTING

http://geekandpoke.typepad.com/geekandpoke/images/2008/05/16/se17.jpg

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, September 26, 2008

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, December 11, 2008

So what is it, really?

- According to NIST:

  Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, 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

© 2015 M. Canini

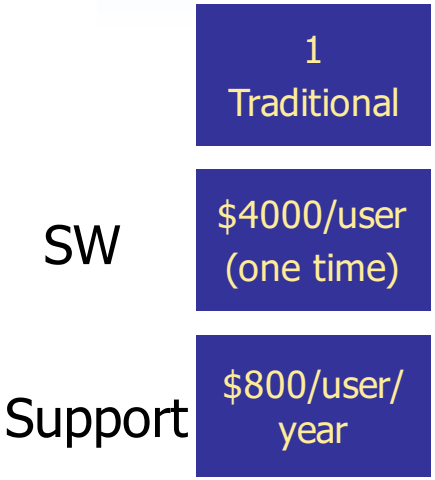
Other terms you may have heard

- 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

- Computing at scale
  - 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?
  - Evolution of software business models
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization: How clouds work 'under the hood'
  - Some cloud computing challenges

Model 1



- For large software companies (MS, SAP, Oracle) is a good model
- Buy perpetual license, and upgrades, bug fixes, phone support

Model 2

	1 Traditional	2 Open Source
SW	$4000/user (one time)	$0/user
Support	$800/user/ year	$1600/user/ year

- Not as many success stories (biggest one probably RedHat)
- But open source is fueling much success for cloud computing

Management Costs

- Cost to manage software is 4x the purchase price per year

© 2015 M. Canini

Model 3

	1 Traditional	2 Open Source	3 Outsourcing
SW	$4000/user (one time)	$0/user	$4000/user (one time)
Support	$800/user/ year	$1600/user/ year	$800/user /year
Mgmt			Bid <1300/user/ month
			At home / at customer

- Outsourcing became popular in the ’90s as Internet got widespread
- Cost reduction through low-cost labor is hard to sustain
- Human error is #1 cause of computer system failures

Model 4

	1 Traditional	2 Open Source	3 Outsourcing	4 Hybrid
SW	$4000/user (one time)	$0/user	$4000/user (one time)	$4000/user (one time)
Support	$800/user/ year	$1600/user/ year	$800/user/ year	$800/user/ year
Mgmt			Bid <1300/user/ month	$150/user/ month
			At home / at customer	At home / at customer

- How to decrease costs by 10x?
- Standardization, specialization, repetition → in time automate

© 2015 M. Canini

Model 5

	1 Traditional	2 Open Source	3 Outsourcing	4 Hybrid	5 Hybrid+
SW	$4000/user (one time)	$0/user	$4000/user (one time)	$4000/user (one time)	$300/user/ month
Support	$800/user/ year	$1600/user/ year	$800/user/ year	$800/user/ year	
Mgmt			Bid <1300/user/ month	$150/user/ month	
			At home / at customer	At home / at customer	At home / at customer

- Pure subscription model
- Simply a change of the payment term compared to Model 4

© 2015 M. Canini

Model 6

	1 Traditional	2 Open Source	3 Outsourcing	4 Hybrid	5 Hybrid+	6 Cloud
SW	$4000/user (one time)	$0/user	$4000/user (one time)	$4000/user (one time)	$300/user/ month	<$100/user/ month
Support	$800/user/ year	$1600/user/ year	$800/user/ year	$800/user/ year		
Mgmt			Bid <1300/user/ month	$150/user/ month		
			At home / at customer	At home / at customer	At home / at customer	

- Software companies gone public since 2000 deliver Cloud services
- 10x cost reduction in exchange for higher degree of standardization

© 2015 M. Canini

Model 7

	1 Traditional	2 Open Source	3 Outsourcing	4 Hybrid	5 Hybrid+	6 Cloud	7 Internet
SW	$4000/user (one time)	$0/user	$4000/user (one time)	$4000/user (one time)	$300/user/ month	<$100/user/ month	Ads Transactions Embedded (<$10/user/ month)
Support	$800/user/ year	$1600/user/ year	$800/user/ year	$800/user/ year			
Mgmt			Bid <1300/user/ month	$150/user/ month			
			At home / at customer	At home / at customer	At home / at customer		

- No direct charge; monetization model is asymmetric
- Software is paid through ads, transactions fees, …
- Consumer apps have dramatically lower costs than enterprise apps

Cloud computing is a business model

AWS us-west-2 prices

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
t2.micro	1	Variable	1	EBS Only	$0.013 per Hour
t2.small	1	Variable	2	EBS Only	$0.026 per Hour
t2.medium	2	Variable	4	EBS Only	$0.052 per Hour
t2.large	2	Variable	8	EBS Only	$0.104 per Hour
m4.large	2	6.5	8	EBS Only	$0.126 per Hour
m4.xlarge	4	13	16	EBS Only	$0.252 per Hour
m4.2xlarge	8	26	32	EBS Only	$0.504 per Hour

Compute
Per service hour
$0.16/hour
(averaged on instance sizes)

Storage
Per GB stored
SSD: $0.10 HDD: $0.05
GB/month

© 2015 M. Canini

Plan for today

- Computing at scale
  - 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?
  - Evolution of software business models
  - What kinds of clouds exist today? NEXT
  - What kinds of applications run on the cloud?
  - Virtualization: How clouds work 'under the hood'
  - Some cloud computing challenges

© 2015 M. Canini

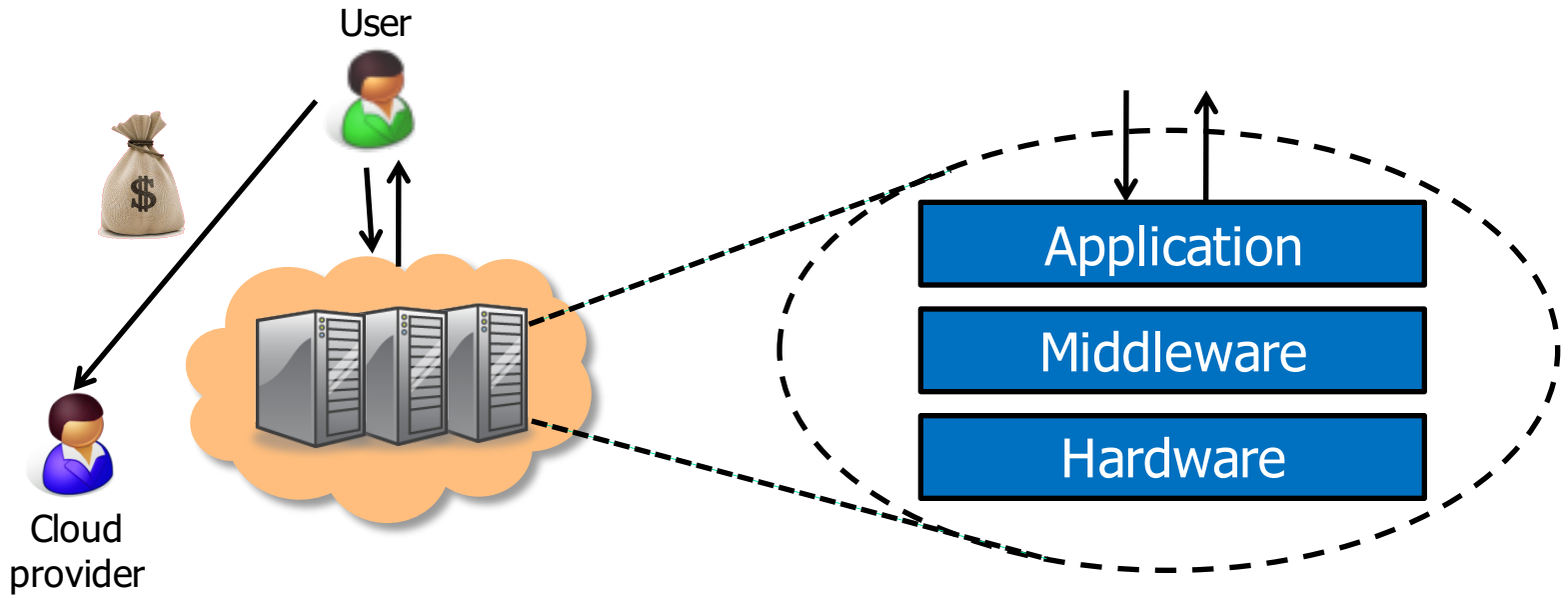
Everything as a Service

- 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 types commonly distinguished:
  - Software as a service (SaaS)
    - Analogy: Restaurant. Prepares&serves entire meal, does the dishes, ...
  - Platform as a service (PaaS)
    - Analogy: Take-out food. Prepares meal, but does not serve it.
  - Infrastructure as a service (IaaS)
    - Analogy: Grocery store. Provides raw ingredients.
  - Other *aaS types have been defined, but are less common
    - Desktop, Backend, Communication, Network, Monitoring, ...

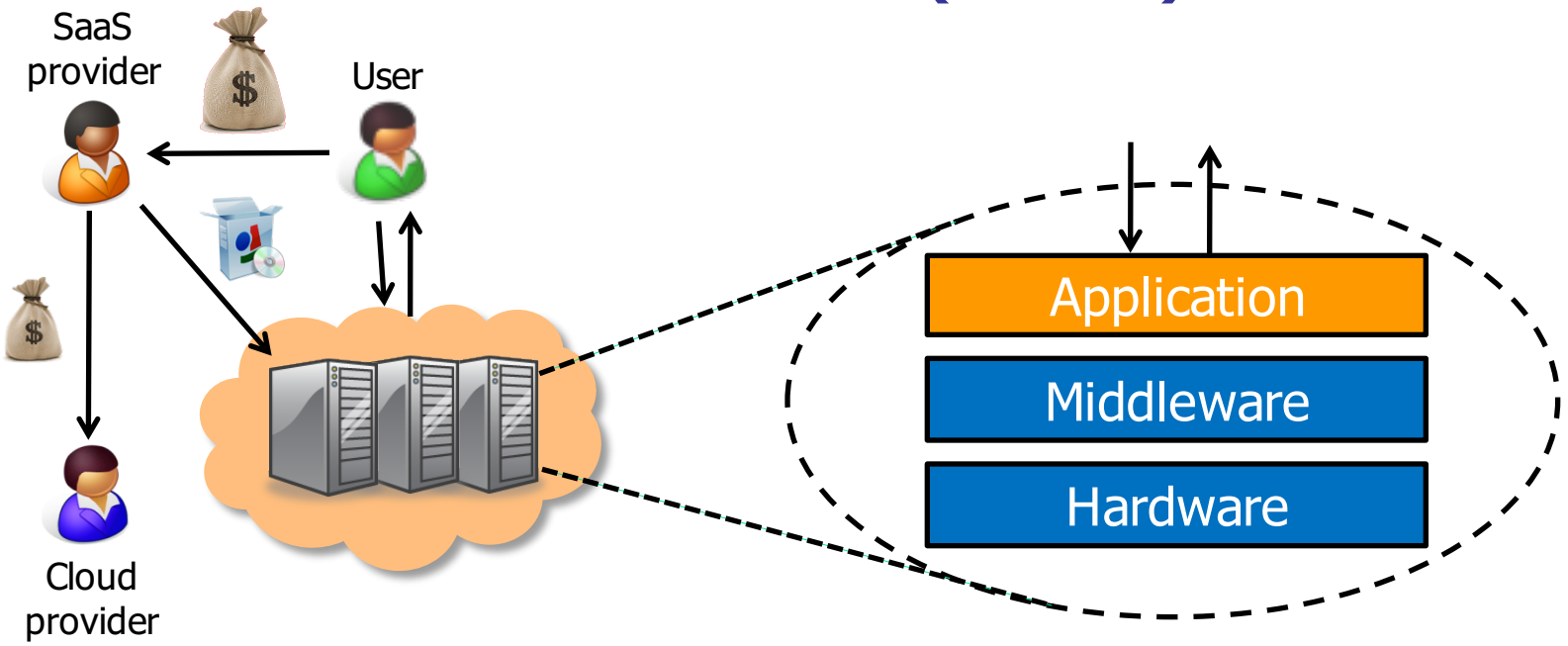
© 2015 M. Canini

Software as a Service (SaaS)



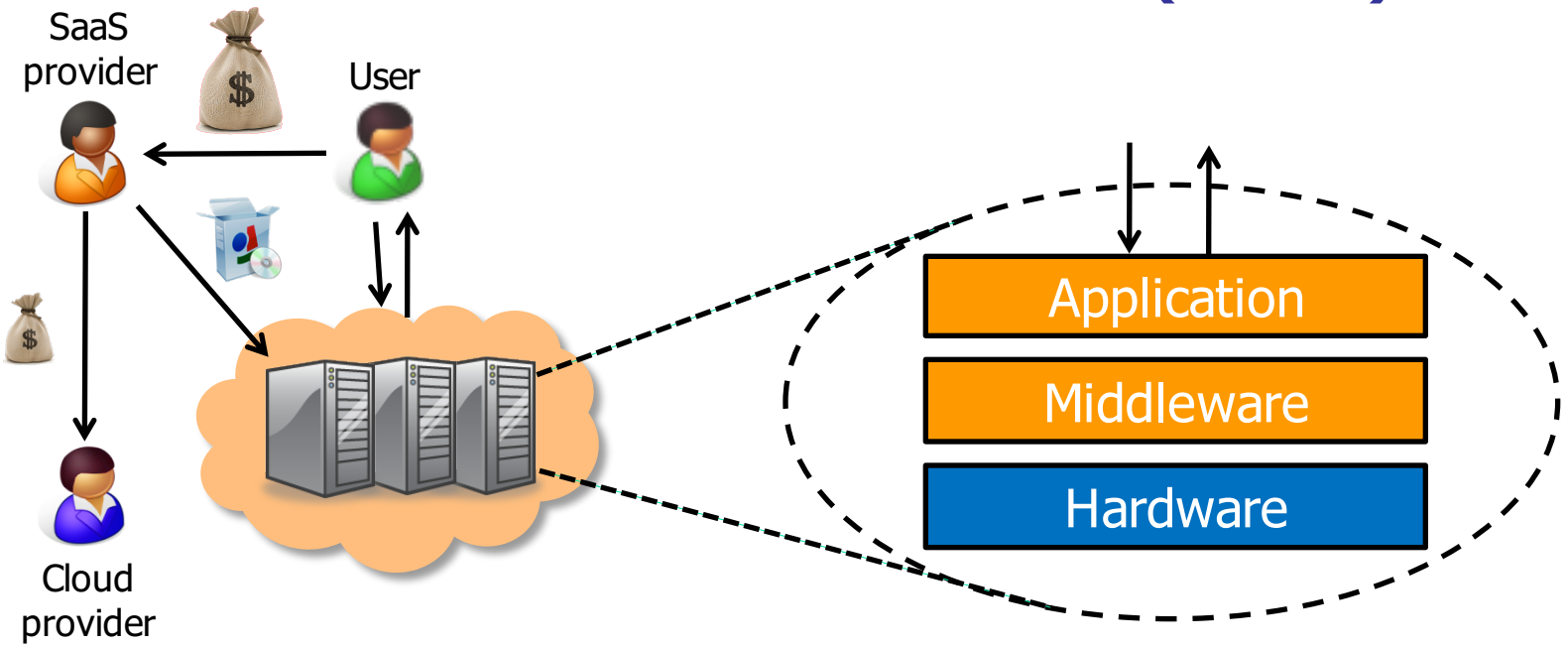
- Cloud provides an entire application
  - Word processor, spreadsheet, CRM software, calendar…
  - Customer pays cloud provider
  - Example: Google Apps, Salesforce.com, Concur, Doodle

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 infrastructure
  - Example: Windows Azure, Google App Engine, Heroku

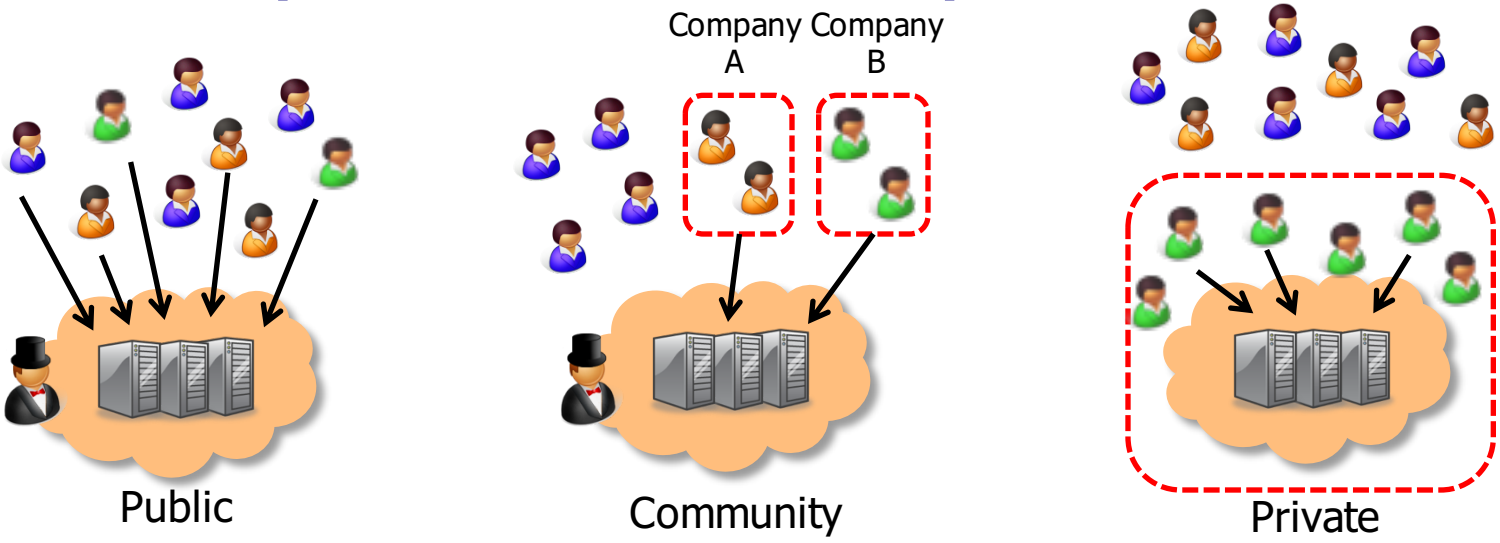
Infrastructure as a Service (IaaS)



- Cloud provides raw computing resources
  - Virtual machine, blade server, storage, network, …
  - Customer pays SaaS provider for the service; SaaS provider pays the cloud for the resources
  - Examples: Amazon Web Services, DigitalOcean, Joyent

© 2015 M. Canini

Private/hybrid/community clouds



- Who can become a customer of the cloud?

Focus of this course →

- Public cloud: Commercial service; open to (almost) anyone.
  Example: Amazon AWS, Microsoft Azure, Google App Engine
- 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.

Plan for today

- Computing at scale
  - 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?
  - Evolution of software business models
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud? NEXT
  - Virtualization: How clouds work 'under the hood'
  - Some cloud computing challenges

Examples of cloud applications

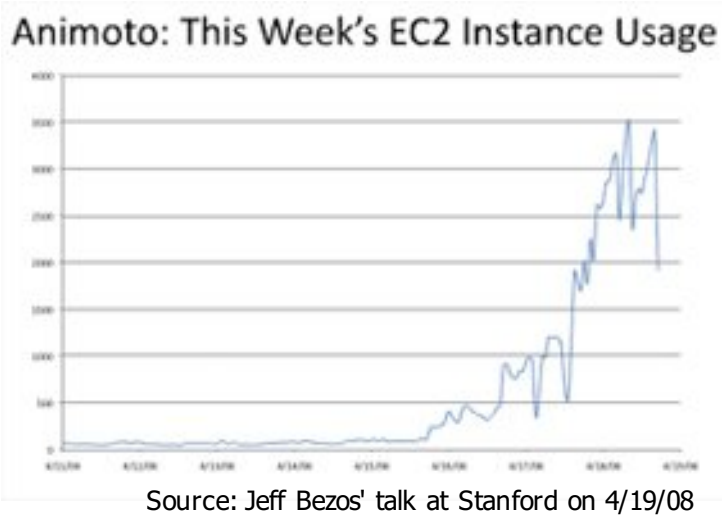
- Application hosting
- Backup and Storage
- Content delivery
- E-commerce
- High-performance computing
- Media hosting
- On-demand workforce
- Search engines
- Web hosting

If interested, glance through
http://aws.amazon.com/solutions/case-studies/
for a list of many use cases

© 2015 M. Canini

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://aws.amazon.com/solutions/case-studies/animoto/

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

Case study:  NOVARTIS

- Novartis Institutes for Biomedical Research is focused on the drug discovery phase of the ~10 year / $1 billion drug development process
- In 2013, NIBR ran a project to screen 10 M compounds against a common cancer target
- Compute requirements >> internal capacity / $
- The project ran across 10,500 EC2 Spot instances (~87,000 cores) for $4,232 in 9 hours
- Equiv. of 39 years of computational chemistry

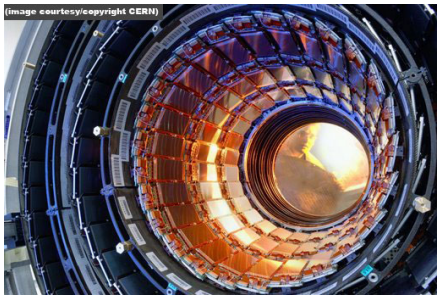
http://aws.amazon.com/solutions/case-studies/novartis/

© 2015 M. Canini

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 economies of scale
  - Elasticity - can quickly scale up/down as demand varies
- Different types of clouds
  - SaaS, PaaS, IaaS; public/private/community 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
  - Auditability requirements
  - Legislative frameworks
- Example: Personal data privacy
  - EU Data Protection law
- Example: Processing medical records (US)
  - HIPAA (Health Insurance Portability and Accountability Act) privacy and security rule

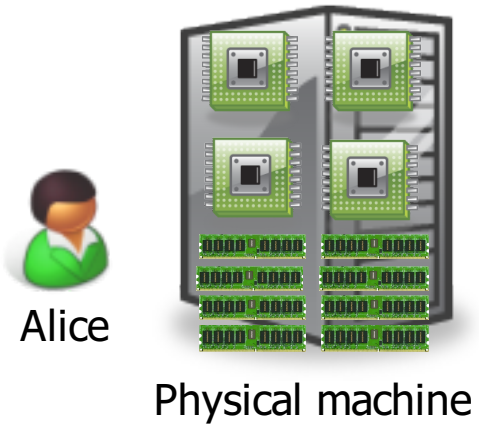
Recap: Cloud applications

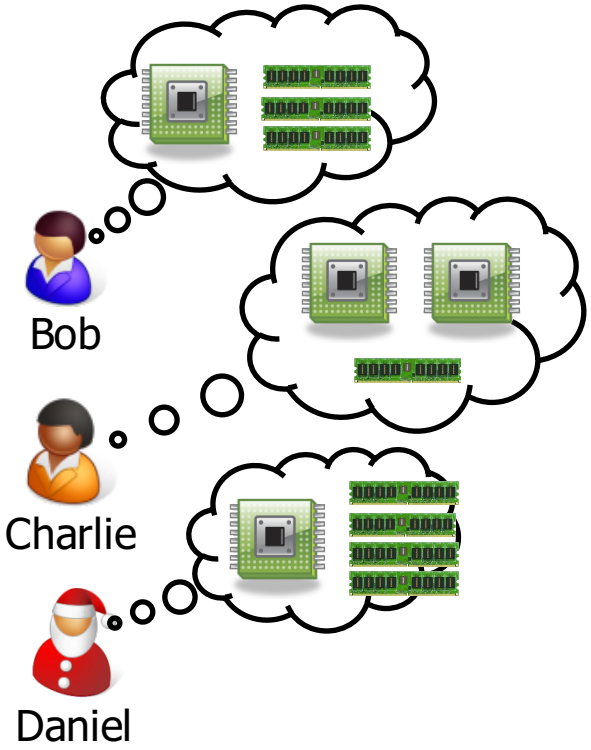
- Clouds are good for many things...
  - Applications that involve large amounts of computation, storage, bandwidth
  - Especially when lots of resources are needed quickly (Novartis example) or load varies rapidly (Animoto example)
- ... but not for all things
  - and might not be the cheapest solution either

Plan for today

- Computing at scale
  - 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?
  - Evolution of software business models
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization: How clouds work 'under the hood'
  - Some cloud computing challenges

What is virtualization?

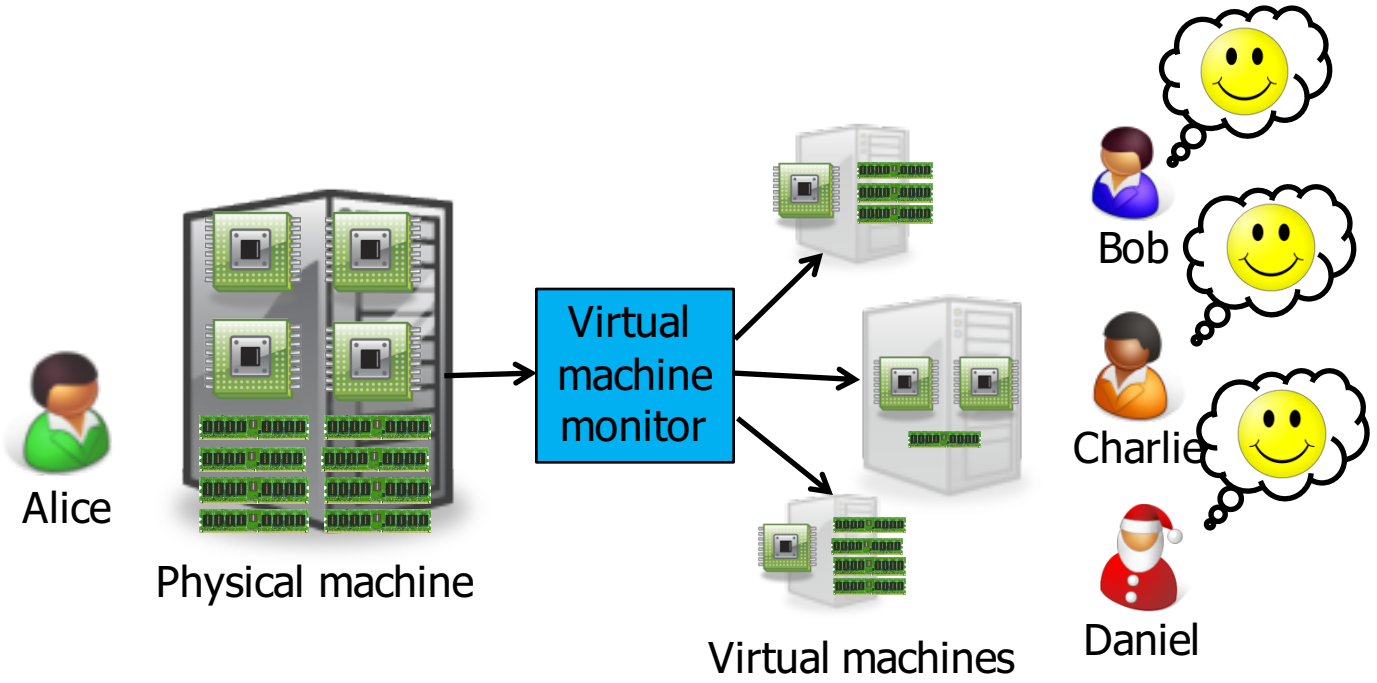




- Suppose Alice has a machine with 4 CPUs and 8 GB of memory, and three customers:
  - Bob wants a machine with 1 CPU and 3GB of memory
  - Charlie wants 2 CPUs and 1GB of memory
  - Daniel wants 1 CPU and 4GB of memory
- What should Alice do?

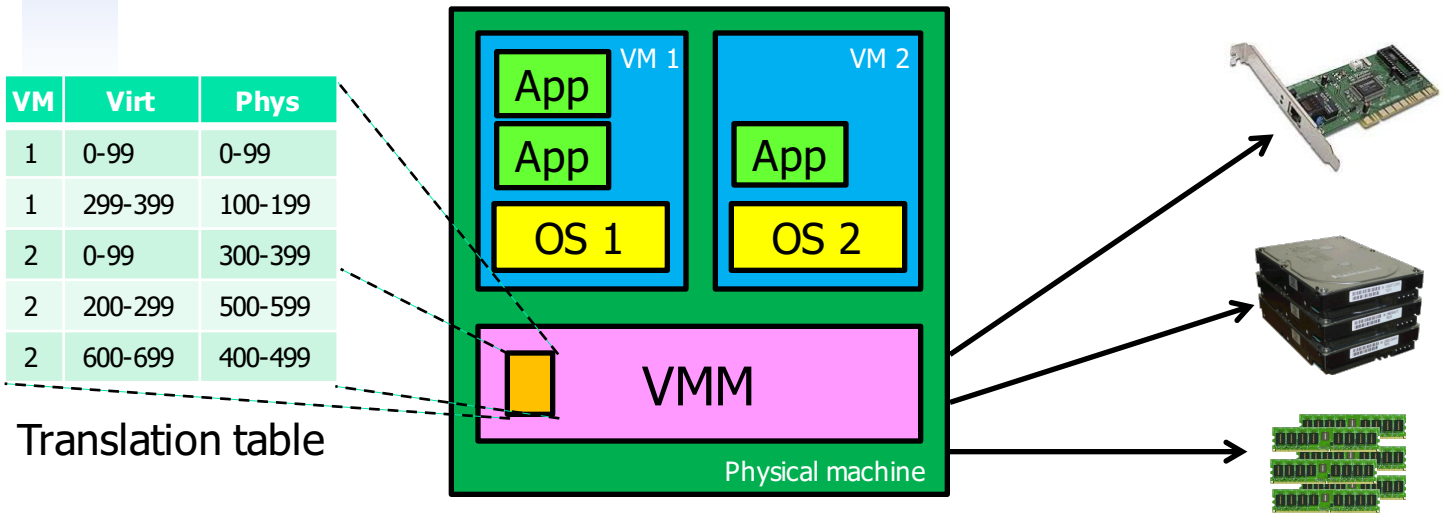
© 2015 M. Canini

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)

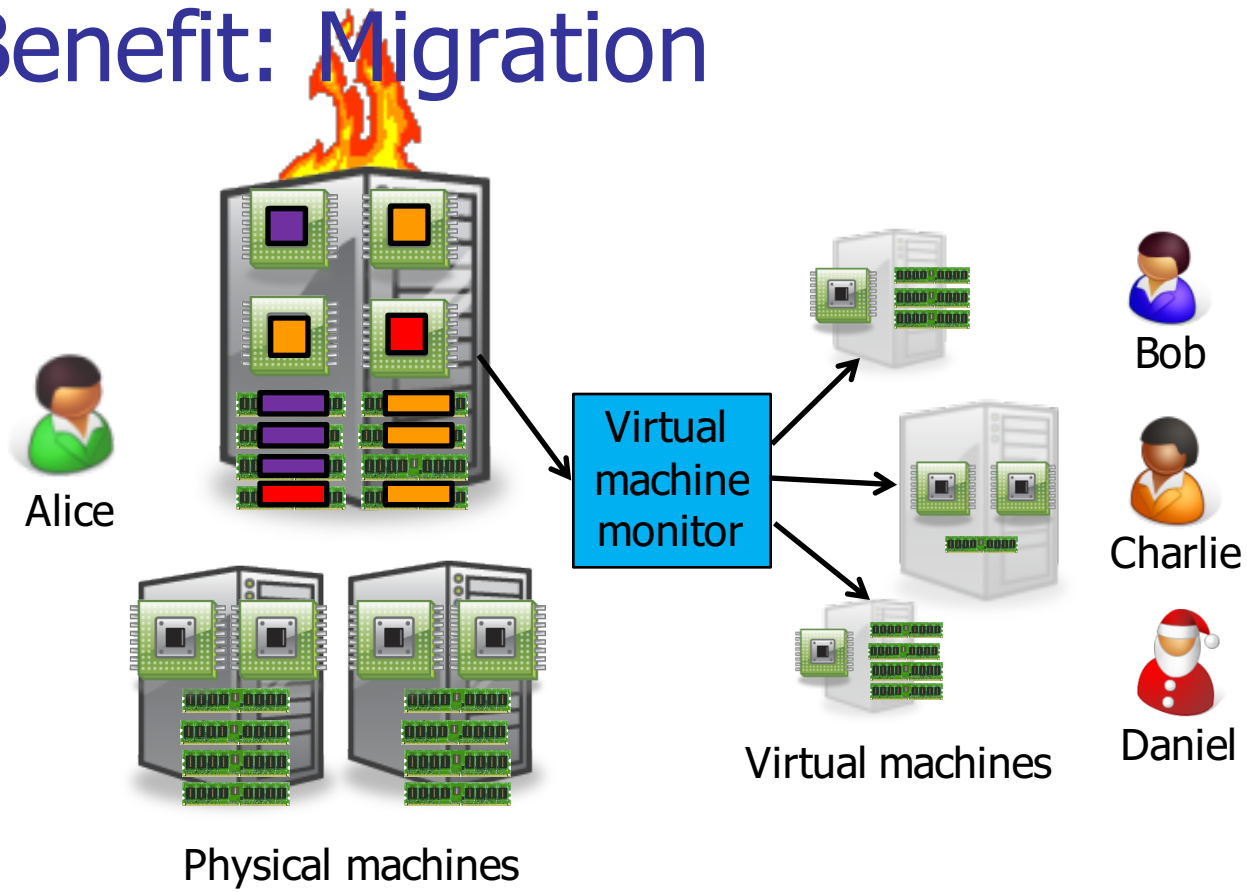
How does it work?



- 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?

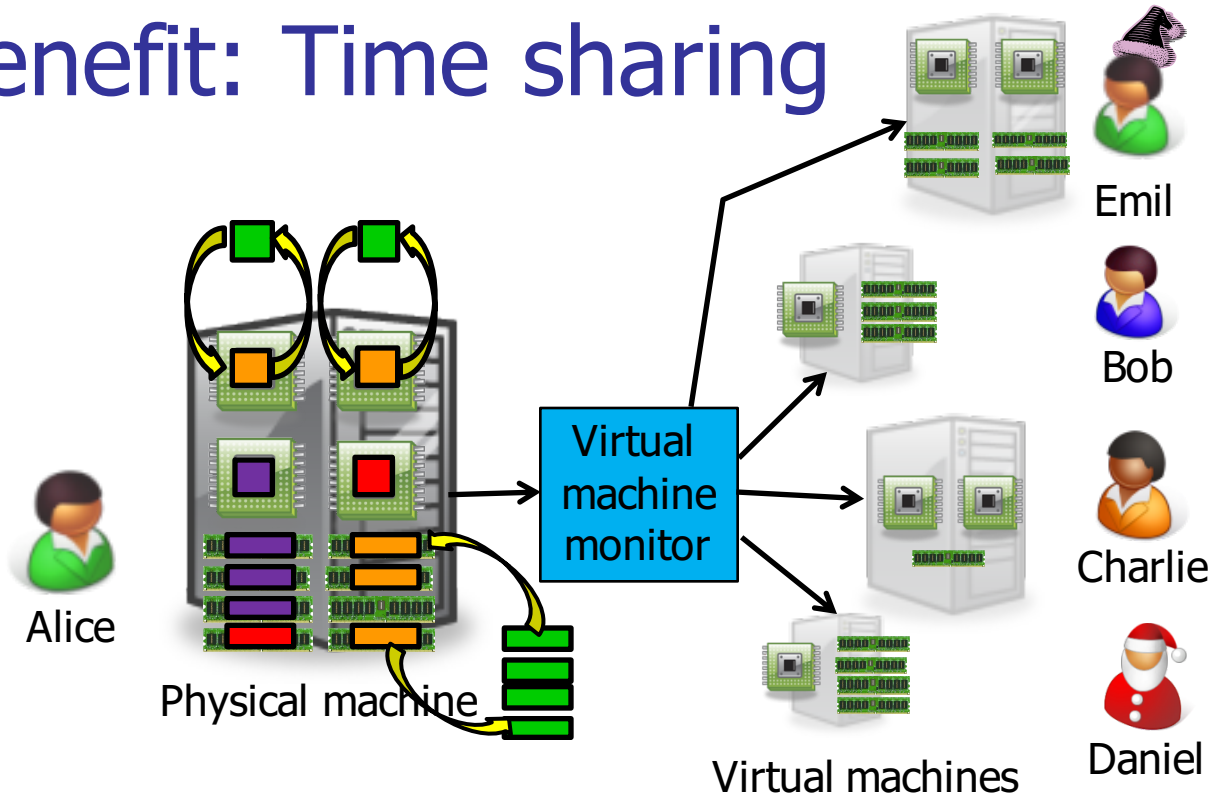
© 2015 M. Canini

Benefit: Migration



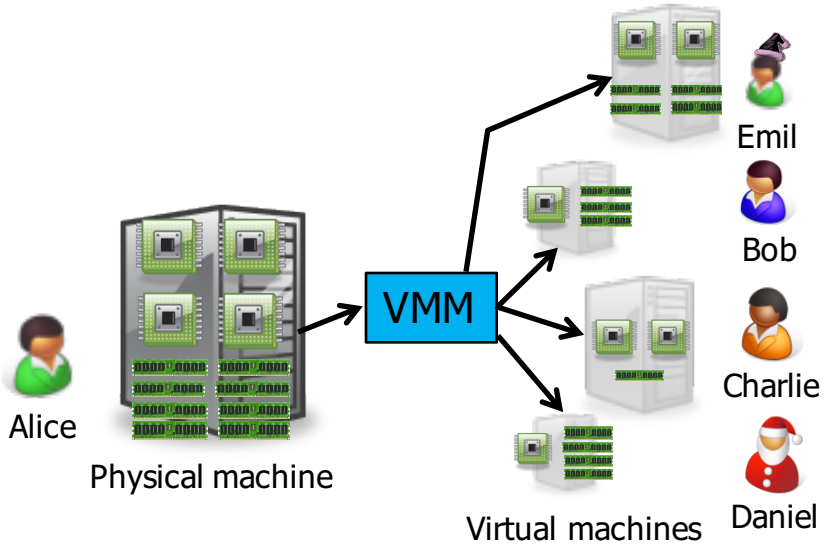
- What if the machine needs to be shut down?
  - e.g., for maintenance, consolidation, ...
  - Alice can migrate the VMs to different physical machines without any customers noticing

Benefit: Time sharing



- What if Alice gets another customer?
  - Multiple VMs can time-share the existing resources
  - Oversubscription: 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

© 2015 M. Canini

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

© 2015 M. Canini

Plan for today

- Computing at scale
  - 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?
  - Evolution of software business models
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization: How clouds work 'under the hood'
  - Some cloud computing challenges NEXT

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
ECC	~2 hrs	6/30/12

Some recent 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?

© 2015 M. Canini

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 [AF10]

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

© 2015 M. Canini

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
   - Fine-grain accounting?

© 2015 M. Canini

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, FBI raid after criminal activity
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

© 2015 M. Canini

Plan for today

- Scalable 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?
  - Evolution of software business models
  - What kinds of clouds exist today?
  - What kinds of applications run on the cloud?
  - Virtualization: How clouds work 'under the hood'
  - Some cloud computing challenges

Any questions?

Stay tuned

http://www.flickr.com/photos/10909957@N03/3135455311/

Next time you will learn about:

Design for large scale; Concurrency, consistency, fault tolerance

© 2015 M. Canini