COMP5349 – Cloud Computing

Week 1: Introduction to Cloud Computing

Uwe Roehm
School of Information Technologies

Cloud Computing

Newsroom \ Announcements \ Gartner Says Public Cloud Services In Mature Asia Pacific and Japan...

Press Release Share: Like 16 Tweet 35 Share 79 +1

Singapore, January 15, 2015 View All Press Releases

Gartner Says Public Cloud Services In Mature Asia Pacific and Japan Forecast To Reach $7.4 Billion in 2015

The public cloud services market in the mature Asia/Pacific and Japan (APJ) region is on pace to grow 14.2 percent in 2015 to total $7.4 billion, up from $6.5 billion in 2014, according to Gartner, Inc. Cloud management and security services, the fastest growth segment of the cloud services market in mature APJ region, is expected to grow 29.9 percent in 2015 to $264.5 million.

Gartner predicts that in 2018, total public cloud services spending in the mature APJ region will rise to $11.5 billion.

"Many countries in the mature Asia/Pacific Japan region have solid reliable telecommunications infrastructure and relatively advanced technology usage profiles. Despite challenges in the global economy, we expect consistent and stable growth to continue through to 2018," said Ed Anderson, research vice president at Gartner.

By 2018, business process as a service (BPaaS) cloud services will make up 9.2 percent of the overall public cloud services market in APJ, platform as a service (PaaS) will represent 3 percent of the market, SaaS will be at 21.5 percent, cloud management/security services at 4 percent, infrastructure as a service (IaaS) will be at 9.8 percent, and the remaining 52.5 percent will come from cloud advertising.

Cloud management and storage and SaaS will be among the faster growing public cloud services during

"Cloud management and storage and SaaS will be among the faster growing public cloud services during through 2018 as more enterprise and government users jump onto cloud services.."

Outline

- **What is 'The Cloud"?**
- Cloud Computing Service Models
- Provisioning, Elasticity and Pay-As-You-Go
- Cloud Applications

Cloud Computing?

Source: 'Jamie's Cloud Based Web Server' on an Amazon EC2…

Google's First 'Data Center'

http://www.shoppingdict.com/2010/08/illustrate-google-data-center.html

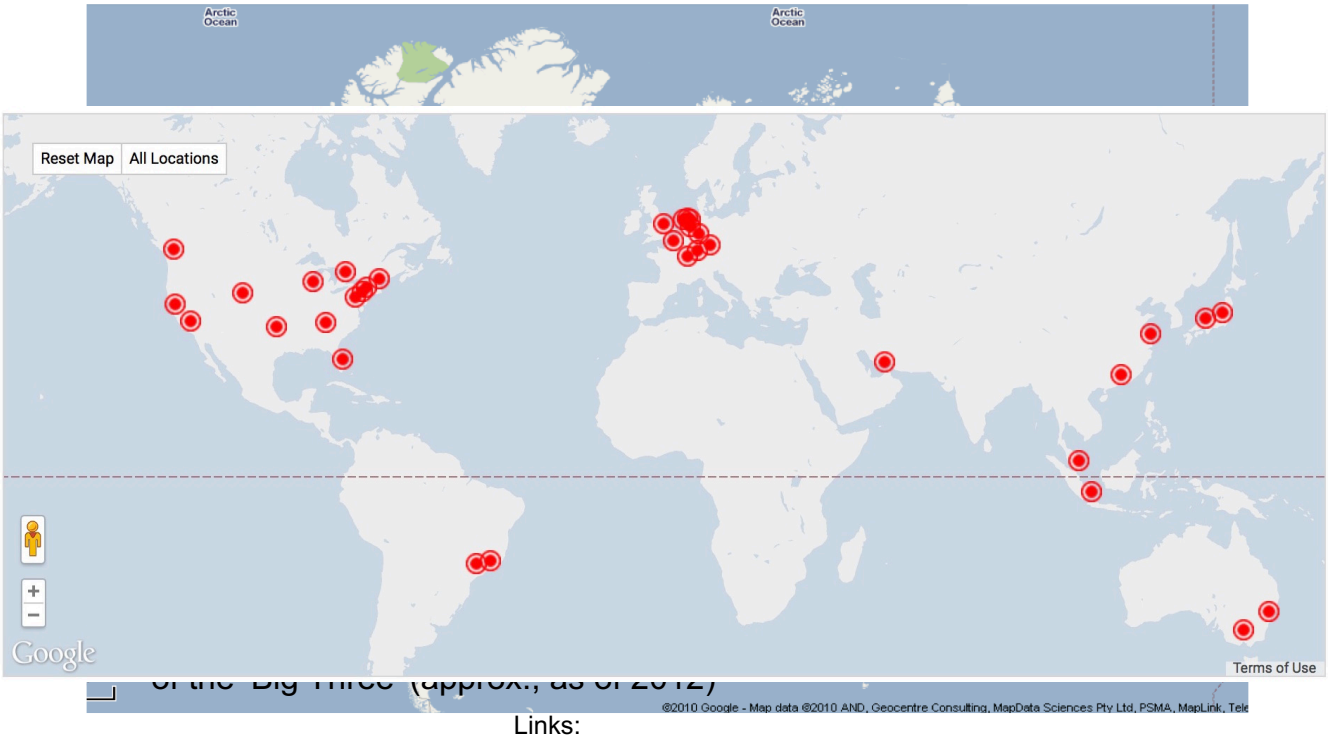
Typical Setting Nowadays

Data Centers

- **Data center site selection criteria**
  1. a large number of low-cost electricity;
  2. green energy, more focus on renewable energy;
  3. close to rivers or lakes; (equipment cooling needs of a large number of water)
  4. a broad land; (privacy and security)
  5. and other data center distance; (fast links between data centers)
  6. tax incentives.

Cf. James Hamilton (perspectives.mvdirona.com).

Some Data Centre Locations

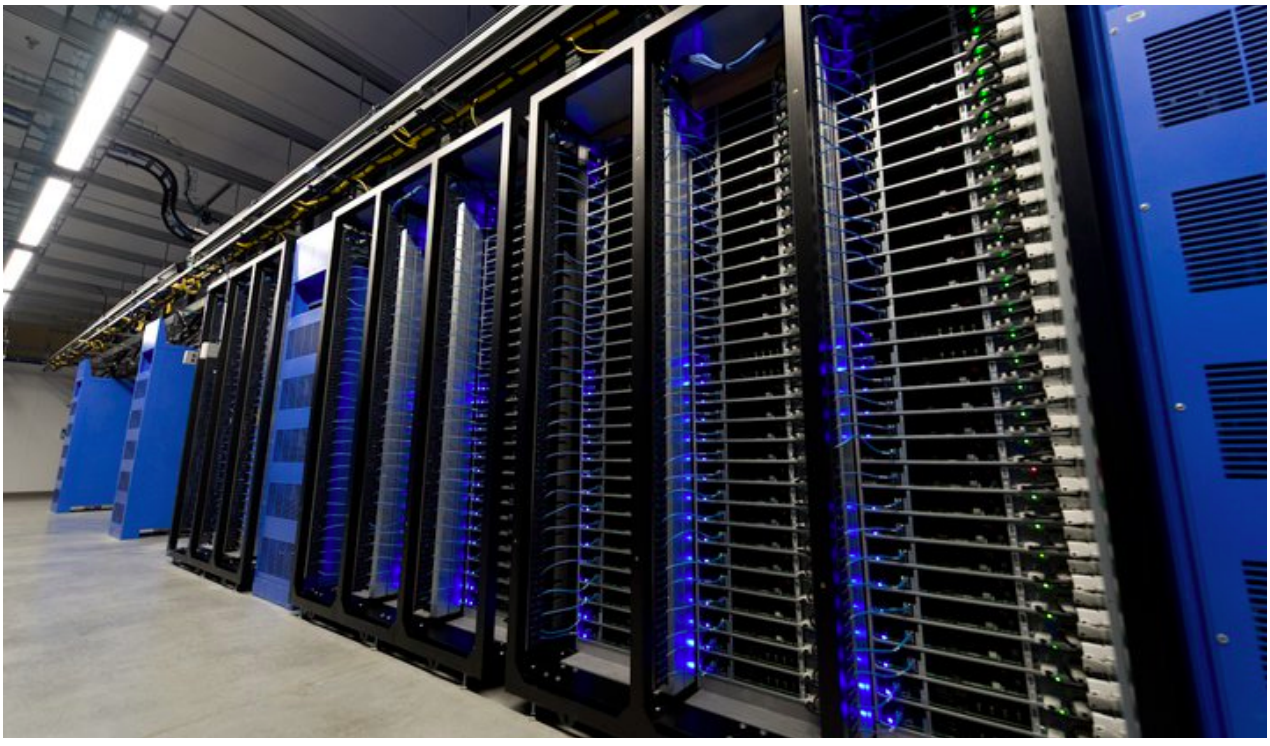


of the "Big Three" (approx., as of 2012)

Links:
http://www.equinix.com.au/locations/asia-colocation/asia-data-centers/
http://www.datacentermap.com/cloud.html

FaceBook Prineville Data Center

http://seattletimes.nwsource.com/
http://www.facebook.com/note.php?note_id=10150150581753133
http://www.datacenterknowledge.com/the-facebook-data-center-faq/



http://www.facebook.com/photo.php?fbid=10150163542757694&set=pu.193287527693&type=1&theater

http://cdn.techsnapr.com/wp-content/uploads/2011/04/FB_SERVER_040_x900.jpg

Apple's 'iCloud' Data Center (N.C.)

Source: http://www.idownloadblog.com

Observation One

- The ‘Cloud’ is very physical…
  …. consisting of globally distributed datacenters

- Not all clouds are the same ;)

- Also: Network latency is a cost
  - Wide-area to get into ‘the cloud’ (aka: your nearest data center)
  - Several layers inside too

What’s New?

- Data Centers are nothing new
  - around for years now to support typically one specific enterprise
  - or to co-locate the servers of several companies
  - Goals:
    - consolidate HW and SW services
    - Provide redundancy and 99.x% availability
  - But still: human factor big cost factor
    - Rule-of-thumb: 1 sys admin for 100 machines
    - Does not scale ;)

Cf. James Hamilton (perspectives.mvdirona.com).

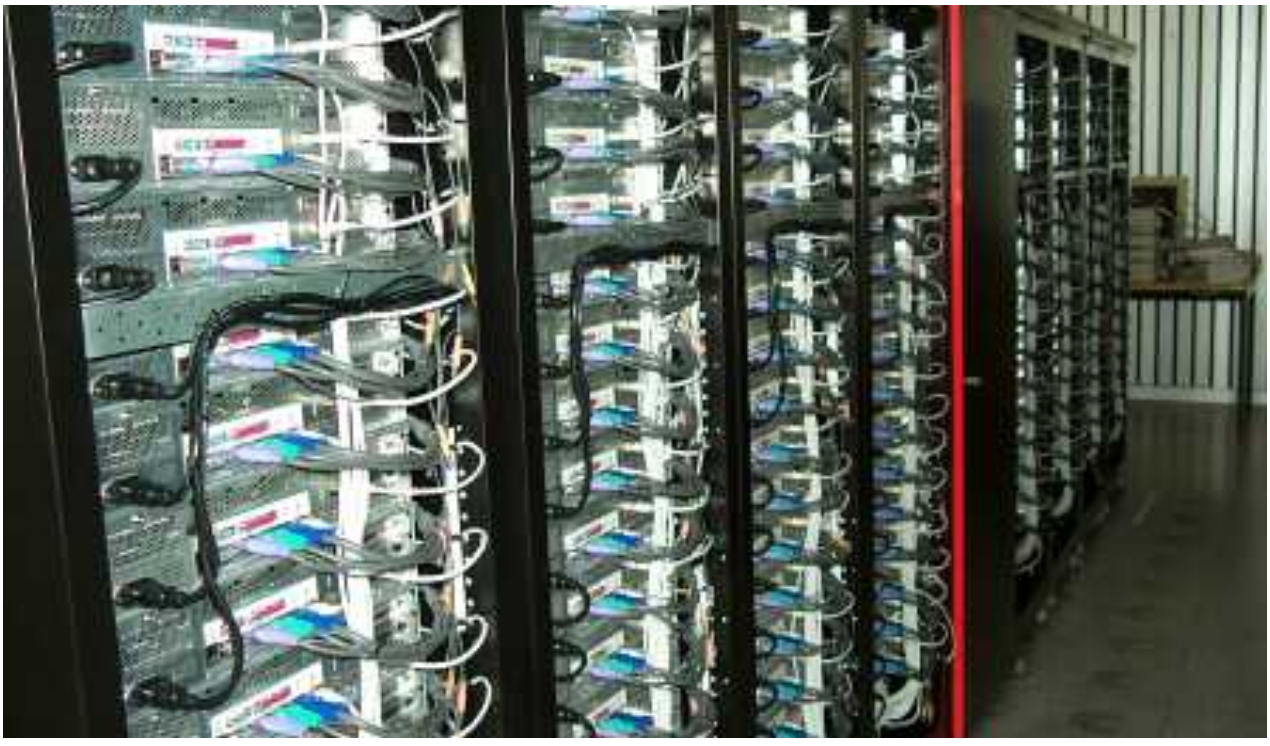
Google start: Hand-Crafted ‘Servers’

Pentium II CPU, 1GB memory and ‘easily swappable’ IDE drives…

Cf *Joel Hruska:* “**The Beast unveiled: inside a Google server”**
-> careful, Google ‘releases’ such details traditionally always on a 1 April…

See also the first Google Rack from 1999 in the Computer History Museum:
http://www.youtube.com/watch?v=z19-6tvGSq4

What is inside?



Case Study: Xibalba Cluster

- Joined Project between ETH Zurich and Microsoft Research
  - 128 nodes
  - Commodity PCs in racks (8 units per rack)
  - Pentium III, 1GHz, 256 MB
  - 4 x 18 GB HDDs
  - 2 Fast Ethernet networks (control and data)
  - Multi-boot Linux or Win2000 (no virtualization!)

  - Those were the days…. Back in 2000

What you can buy today

- **Some Vendor 2U Rack Server (HPC market)**
  - up to 2 x 18-core Intel® Xeon® processors
  - up to 1.5 TB DDR4 ECC RAM
  - embedded hyper-visor (for efficient virtualization)
  - up-to 16 TB internal storage (16 hot-swappable 2.5'' drive bays)
  - Integrated RAID controller
  - 4 x Gigabit-Ethernet NICs; optional 2x10GigE or Infiniband
  - Two redundant 2000W hot-plug power supplies

- Or see also:
  http://www.sgi.com/products/servers/rackable/

“Scale-Up” versus “Scale-Out”

“MPC”

Cluster

Data Center Building Blocks

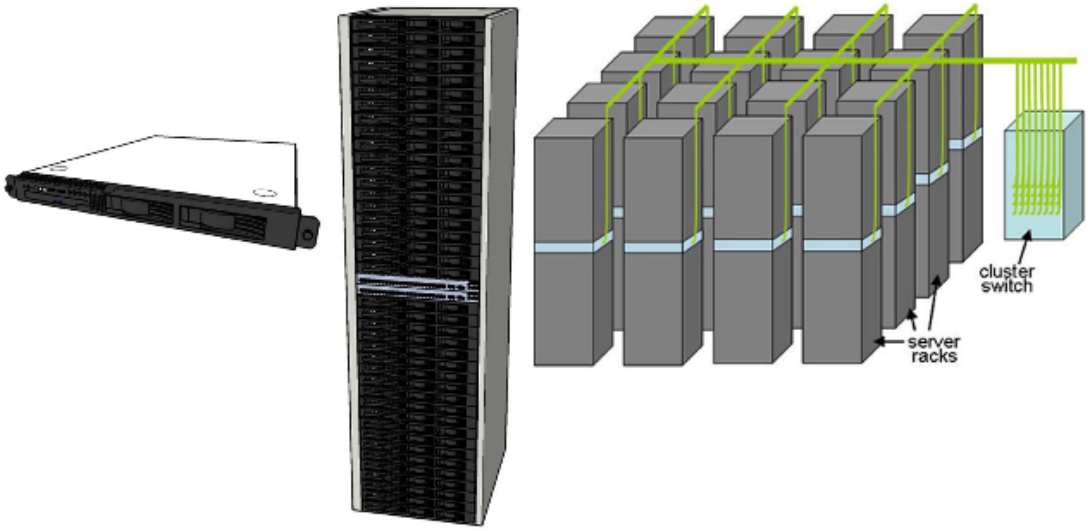


FIGURE 1.1: Typical elements in warehouse-scale systems: 1U server (left), 7′ rack with Ethernet switch (middle), and diagram of a small cluster with a cluster-level Ethernet switch/router (right).

Infrastructure of Scale

- Typical 'cloud-size' data center will have about 40-50k servers
  - 'commodity' design with multiple cores
  - virtualization to host multiple services on same shared harware

  => Plunging cost of computing

- Multiple datacenters
  - At scale multiple datacenters can be used
    - Close to customer
    - Cross data center data redundancy
    - Address international markets efficiently

- Avoid massive upfront data cost & years to fully utilize – Scale supports pervasive automation investment

Why Now?

- Experience with large scale data centers is here
- Unprecedented economies of scale due to Moore's Law
- Plus
  - Pervasive broadband Internet access (the customers are there)
  - Fast CPU virtualization
  - Inexpensive storage
  - Pay-as-you-go billing modes
  - Robust standard software stack
  - Client computers powerful enough for significant client-side computations

- James Hamilton:
  "Every day, Amazon Web Services adds enough new capacity to support all of Amazon.com's global infrastructure through the company's first 5 years, when it was a $2.76B annual revenue enterprise"
  [Amazon Open House slides, 07 June 2011]

Outline

- What is 'The Cloud"?
- **Cloud Computing Service Models**
- Provisioning, Elasticity and Pay-As-You-Go
- Cloud Applications

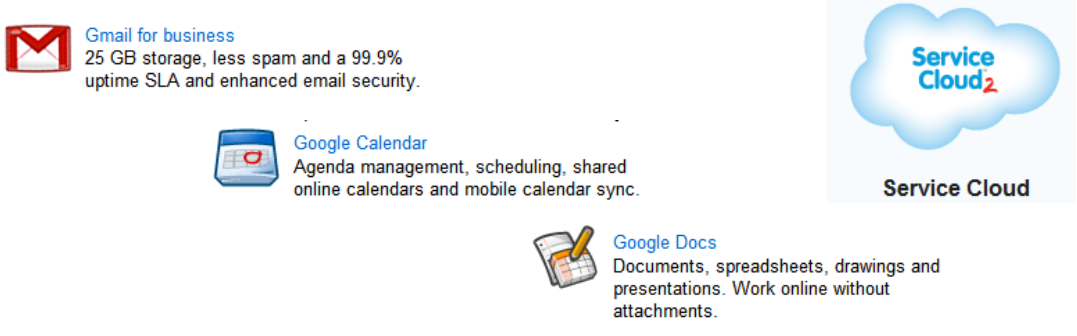
Cloud Computing– a Broad Definition

- A definition by US Governments' National Institute of Standard and Technology
  - "Cloud computing is a model for enabling ubiquitous, convenient, on-demand **network access** to a **shared pool** of **configurable computing resources** (e.g., *networks, servers, storage, applications*, *and services*)"

Slides 3-7 are based on *Cloud Computing Use Cases Whitepaper Version 4.0*

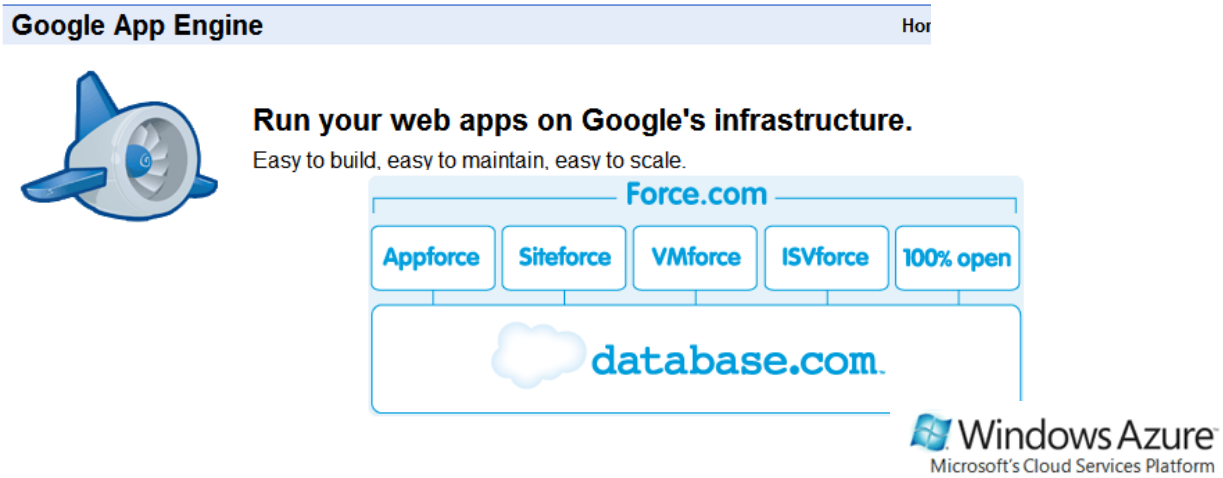
Cloud Computing Service Models

- In this definition, cloud computing has three delivery models:
  - **Software as a Service (SaaS):** The consumer uses an **application**, but does not control the operating system, hardware or network infrastructure on which it's running.
    - *Applications* are restricted to business applications or applications that may normally installed in a business network or personal computer
    - Examples
      - Business applications: CRM solutions from *salesforce*.com
      - Business/Personal applications: Gmail, Google Doc, etc.



Cloud Computing Service Models (II)

- **Platform as a Service (PaaS):** The consumer uses a **hosting environment** for their applications. The consumer controls the applications that run in the environment (and possibly has some control over the hosting environment), but does not control the operating system, hardware or network infrastructure on which they are running. The platform is typically an **application framework**.



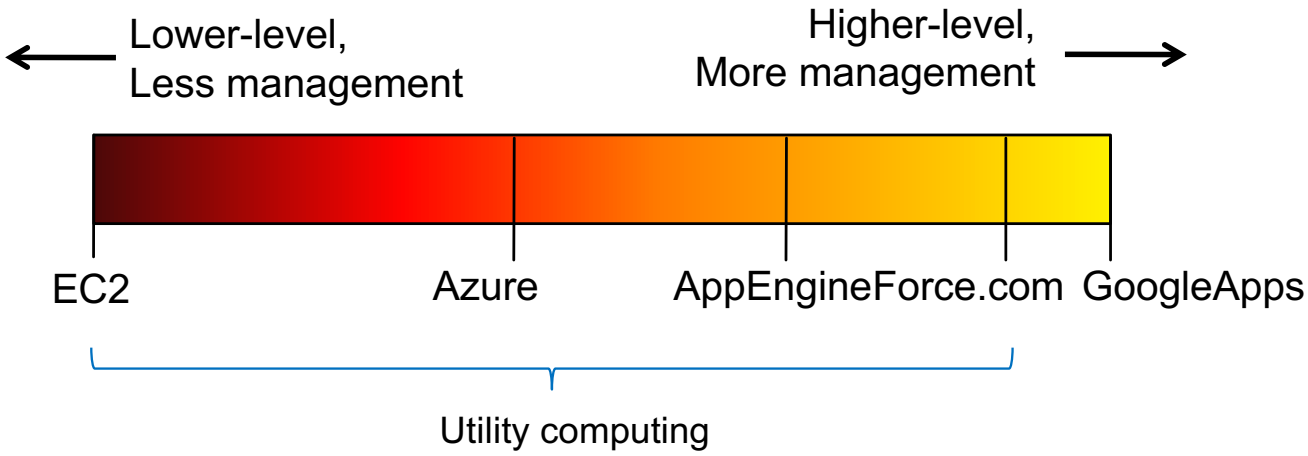
Cloud Computing Service Models (III)

- **Infrastructure as a Service (IaaS):** The consumer uses "**fundamental computing resources**" such as processing power, storage, networking components or middleware. The consumer can control the operating system, storage, deployed applications and possibly networking components such as firewalls and load balancers, but not the cloud infrastructure beneath them.

amazon web services
orionVM
rackspace HOSTING
. . .

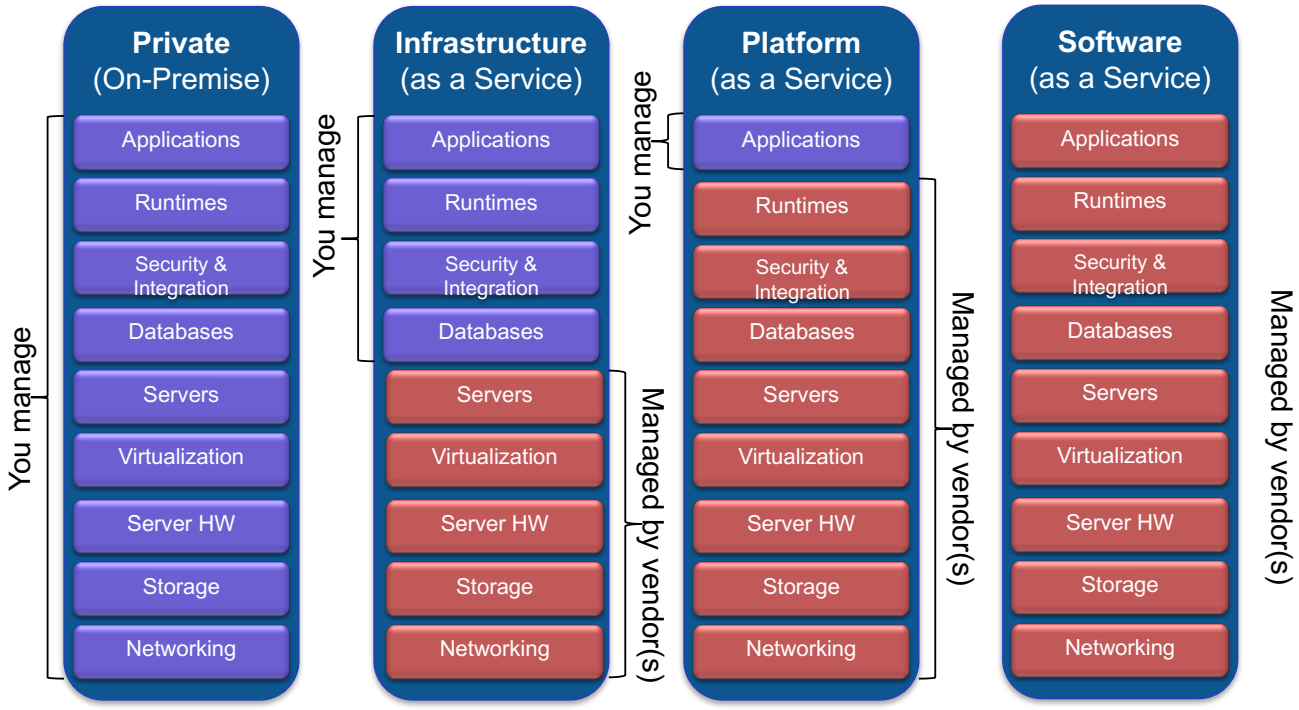
Cloud Server and Data Center Map:
http://www.datacentermap.com/cloud.html

Spectrum of Cloud Services



From Berkeley Cloud presentation: http://berkeleyclouds.blogspot.com/

Clouds Servicing Models



Incentives for Cloud Providers

- Improve **utilization** through **virtualization** or **multitenancy,** and **isolation**
  - **IaaS/PaaS** is built on *virtualization* technology (relatively mature hypervisor technology)
  - **PaaS/SaaS** is built on *multitenancy* technology (very hard!)
  - They all need to provide *isolation*!
- Economy of Scale

Resource	Cost in Medium DC	Cost in Very Large DC	Ratio
Network	$95 / Mbps / month	$13 / Mbps / month	7.1x
Storage	$2.20 / GB / month	$0.40 / GB / month	5.7x
Administration	≈140 servers/admin	>1000 servers/admin	7.1x

Incentives for Cloud Users

- Cloud user
  - Better **provisioning** through **elasticity** and **pay-as-you-go** and other fine-grained pricing models
  - Lift the burden of operational management
  - CapEx vs. OpEx tradeoff
- Example
  - **Netflix**: world's leading Internet subscription service for movies and TV shows
  - Netflix migrated from its own data centers to AWS in 2010
    - Capacity growth rate is accelerating, unpredictable
      - Year on year customer growth is 52%, year on year customers using streaming is up 145% (from ~4M to ~11M).
    - Product lunch spikes– iPhone, WII, PS3, XBox
    - Datacenter is large inflexible capital commitment

Netflix Example: Reasons for Moving to the Cloud

- "We needed to re-architect, which allowed us to question everything, including whether to keep building out our own datacenter solution."
- "Letting Amazon focus on datacenter infrastructure allows our engineers to focus on building and improving our business."
- "We're not very good at predicting customer growth or device engagement."
- "We think cloud computing is the future."

http://techblog.netflix.com/2010/12/four-reasons-we-choose-amazons-cloud-as.html

Q & A with Cloud Architect, at Netflix.

"*Many folks claim that, they can deliver a private cloud at a similar price point to AWS. I assume you ran the numbers yourself. In whatever detail you can share, what does the ROI look like for Netflix?*"

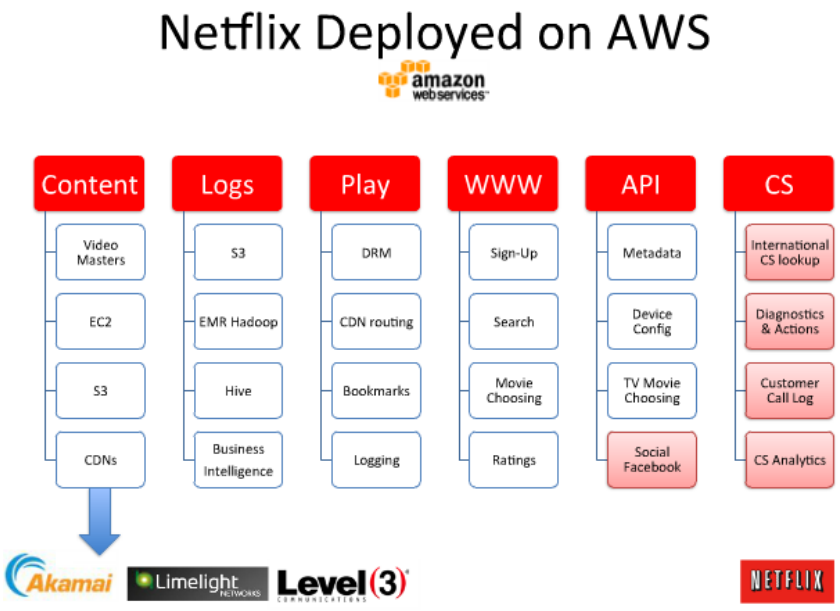
› "Oracle on IBM is very expensive, so AWS looks cheap in comparison"

› "AWS costs are fully burdened, and we could not have hired enough SAs and DBAs to build out our own datacenter this fast."

› "costs are elastic, you start paying for a resource just before it goes live, and if you stop using a resource you stop paying for it"

http://cloudscaling.com/blog/cloud-computing/cloud-innovators-netflix-strategy-reflects-google-philosophy

Netflix Example (con't)



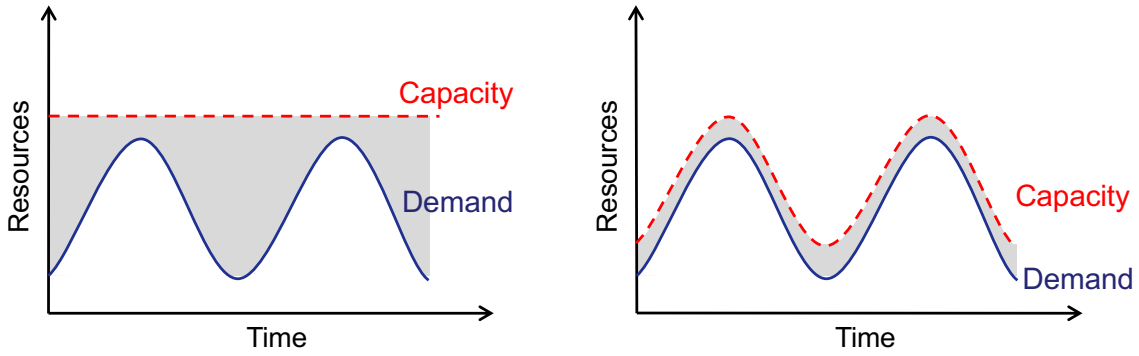
http://www.slideshare.net/adrianco/global-netflix-platform (slide 15)

Outline

- What is the 'Cloud'?
- Cloud Computing Service Models
- **Provisioning, Elasticity and Pay-As-You-Go**
- Cloud Applications

The Provisioning Problem

- It is very hard to predict usage and to provision sufficient capacities



Unused resources

From Berkeley Cloud presentation: http://berkeleyclouds.blogspot.com/

Elasticity and Pay-As-You-Go

- Elasticity
  - The cloud allows scaling up and scaling down of resource usage on an ‘as-needed’ basis. Elapsed time to increase or decrease usage is measured in seconds or minutes
- Pay-As-You-Go
  - Consumer is charged based on resourced they used (per instance or per cpu time)

Region: Asia Pacific (Sydney)

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
t2.micro	1	Variable	1	EBS Only	$0.02 per Hour
t2.small	1	Variable	2	EBS Only	$0.04 per Hour
t2.medium	2	Variable	4	EBS Only	$0.08 per Hour
t2.large	2	Variable	8	EBS Only	$0.16 per Hour

http://aws.amazon.com/ec2/pricing/

An Example of Amazon’s Scaling Mechanism

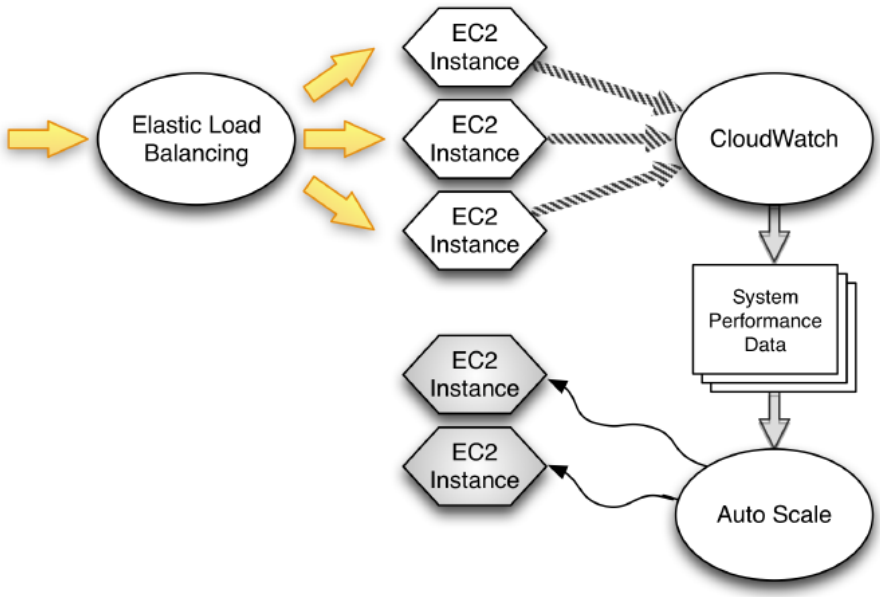
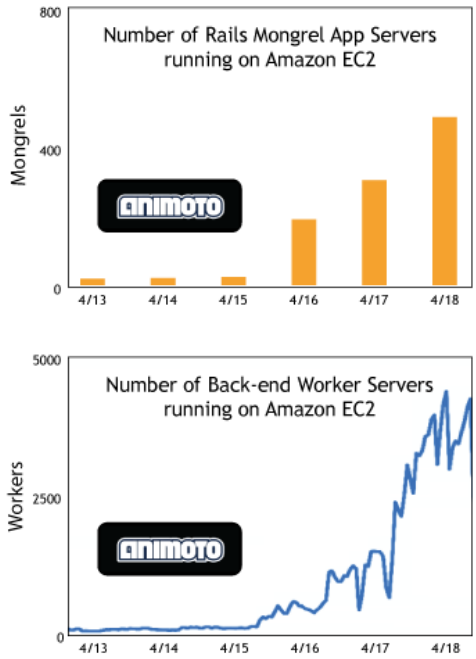


Figure 7.1. The relationship between elastic load balancing, CloudWatch, and auto scale

Jeff Barr, Host Your Web Site in the Cloud, published by Sitepoint, 2010

The famous Animoto Example

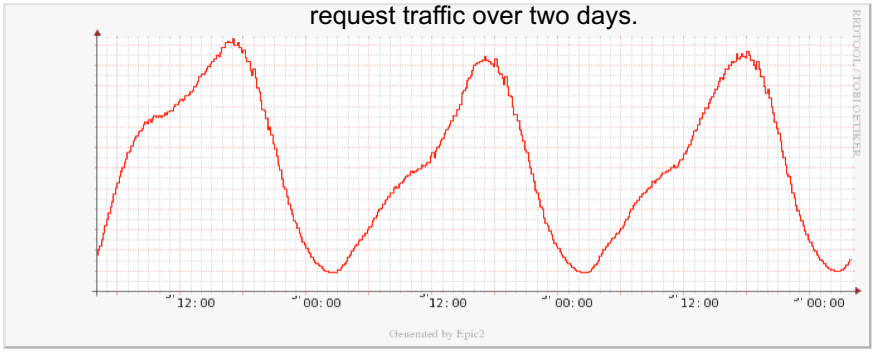


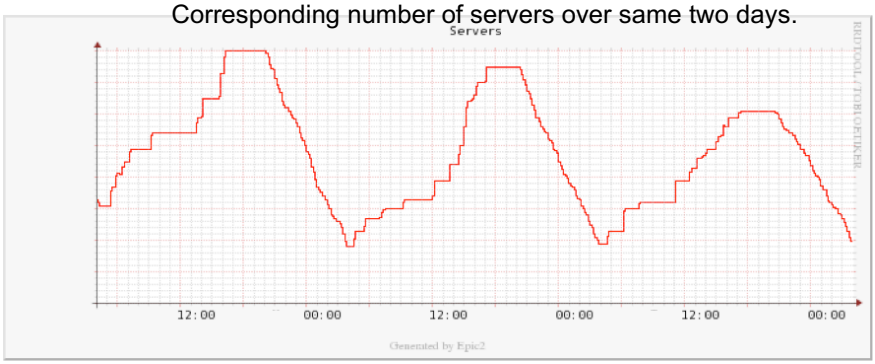
"They had 25,000 members on Monday, 50,000 on Tuesday, and 250,000 on Thursday. Their EC2 usage grew as well.

For the last month or so they had been using between 50 and 100 instances. On Tuesday their usage peaked at around 400, Wednesday it was 900, and then 3400 instances as of Friday morning."

http://aws.typepad.com/aws/2008/04/animoto--scali.html

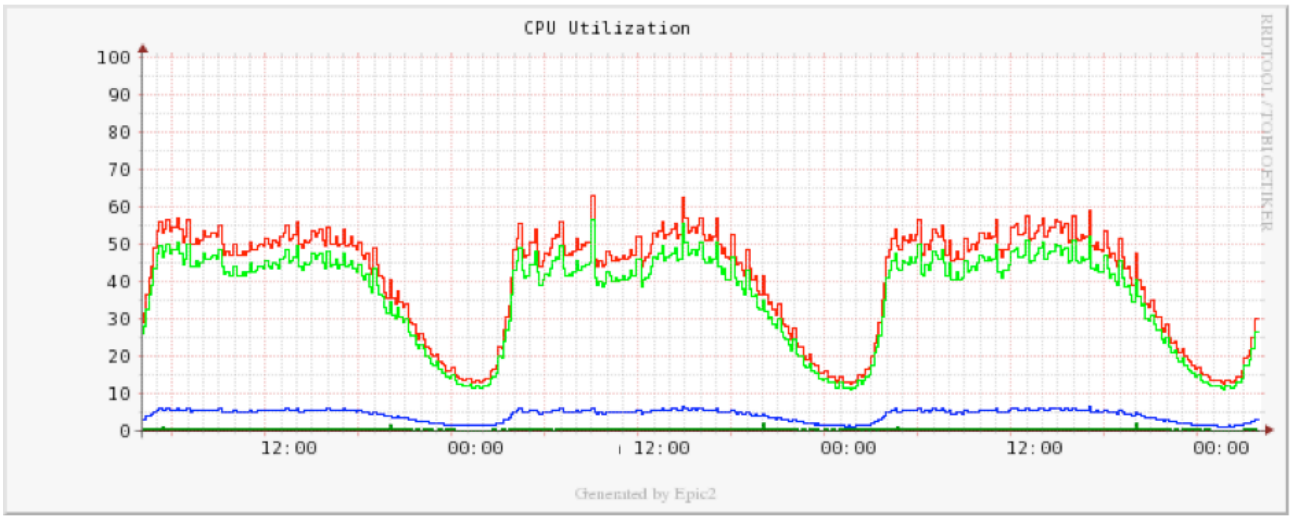
Netflix Auto-scaling Observations





The Netflix Tech Blog: Auto Scaling in Amazon cloud (Jan. 18, 2012)
http://techblog.netflix.com/search/label/autoscaling

Netflix Auto-scaling Observations (cont)



Aggregated CPU utilization during this time period:
Note that under load the aggregate CPU is essentially flat

The Netflix Tech Blog: Auto Scaling in Amazon cloud (Jan. 18, 2012)
http://techblog.netflix.com/search/label/autoscaling

Outline

- What is the 'Cloud'?
- Cloud Computing Service Models
- Utilization, Virtualization, Multitenancy and Isolation
- **Cloud Applications**

Cloud Computing 'Killer Apps'

- Online commerce and communities
  - Amazon; eBay; Facebook; Netflix; Twitter; …
- Mobile applications
  - Siri…
- Data Analytics ('Big Data')
  - Washington Post Engineer used **200** EC2 instances (1,407 server hours) to convert **17,481** pages of Hillary Clinton's travel documents into a form more friendly to use in WWW presentation
  - NY Times used **100** instances of Amazon EC2 to convert 11 million historical articles from TIFF to PDF, within **24** hours, all articles as **4TB** data were converted into **1.5 TB** of pdf.
  - NY Times builds its own Hadoop Toolkit to enable easy writing of MapReduce jobs
    - Motivated by huge volume of data log and the difficulties of running it

http://open.blogs.nytimes.com/tag/hadoop/

Summary

- **Cloud computing**
  - a model for enabling ubiquitous, convenient, on-demand network access to a **shared** pool of **configurable** computing resources
- **Cloud delivering models**
  - *X*aaS (**IaaS**, **PaaS**, **SaaS**, …)
    - SaaS based on multi-tenancy
    - IaaS and PaaS: virtualization technologies
  - Private vs. **Public Cloud**
- **Motivation for Cloud Computing**
  - Better utilization/economy of scale on the cloud provider
  - Better scalability / elasticity with Pay-as-you-go for customers

References

- Armbrust et al: "Above the Clouds: A Berkeley View of Cloud Computing", TR EECS-2009-28, UC Berkeley, 2009.
- C.D. Weismann, S. Bobrowski: "The design of the force.com multitenant internet application development platform", SIGMOD 2009.
- The Netflix Tech Blog [accessible from: http://techblog.netflix.com]
- Arik HesselDahl, *Seven Questions for Adam Selipsky, VP at Amazon Web Services*, All things Digital, March, 7, 2011 [accessible from: http://newenterprise.allthingsd.com/20110307/seven-questions-for-adam-selipsky-head-of-amazon-web-services/]
- The Global Netflix platform[http://qconsf.com/dl/qcon-sanfran-2011/slides/AdrianCockcroft_NetflixInTheGlobalCloud.pdf]
- High Scalability Blog [accessible from: http://highscalability.com/]
- "Above the Clouds: A Berkeley View of Clould Computing", 2009 [accessible from: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf]
- Werner Vogels, *Beyond Server Consolidation*. ACM Queue, Jan/Feb 2008.

Readings for Next Week:

Readings for Week 2:

- Werner Vogels: **"Beyond Server"**, ACM Queue 2008.
- Paul Barham et al: "**Xen and the Art of Virtualization**", SOSP 2003.

- readings are linked on Piazza

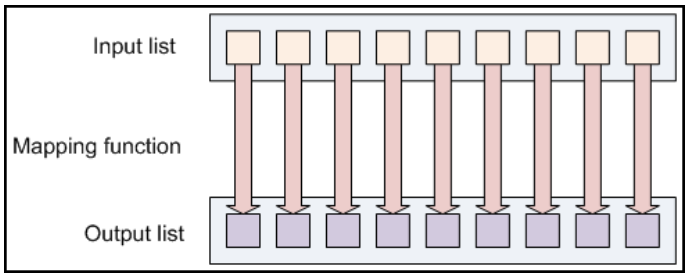
Reminder: 1st Programming Homework

- Homework for **next week (Week 2** – Thursday, 16 March):
- Given:
  - Example data set: 'MovieLens' data set about user-film ratings
  - Example Python/Java code of a file scanner in Map/Reduce fashion
- Task: Extend the given skeleton code so that it computes the average movie rating for films from a given range.

- The aim is to assess your programming skills
  - Understanding and working with a given processing framework
  - Functional thinking: Writing code for functions which solve a well-defined task in the context of a larger framework (in this case: filtering and aggregation)
  - Java and Unix skills (compiling, debugging)
- Details will be published on Piazza and eLearning

Homework Tip 1: MapReduce Model

Inspired by **map** and **fold** in FP
Input & Output: each a set of key/value pairs



Occasionally, map input key is used to associate map's input and output

Keys divide the reduce space
all of the output values are not usually reduced together. All of the values *with the same key* are presented to a single reducer together