Data Centers and Cloud Computing

CS377 Special Topic

Data Centers and Cloud Computing

- Intro. to Data centers
- Virtualization Basics
- Intro. to Cloud Computing
  - Case Study: Amazon EC2

Data Centers

- Large server and storage farms
  - 1000s of servers
  - Many TBs or PBs of data

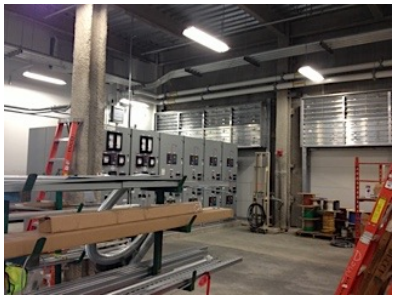
- Used by
  - Enterprises for server applications
  - Internet companies
    - Some of the biggest DCs are owned by Google, Facebook, etc
- Used for
  - Data processing
  - Web sites
  - Business apps

Google NETFLIX iCloud amazon.com

Inside a Data Center

- Giant warehouse filled with:
  - Racks of servers
  - Storage arrays
  - Network switches
- Cooling infrastructure
- Power converters
- Backup generators

MGHPCC Data Center


- Data center in Holyoke

Modular Data Center

- Use shipping containers
- Each container filled with thousands of servers
- Can easily add new containers
  - "Plug and play"
  - Just add electricity
- Allows data center to be easily expanded
- Pre-assembled, cheaper

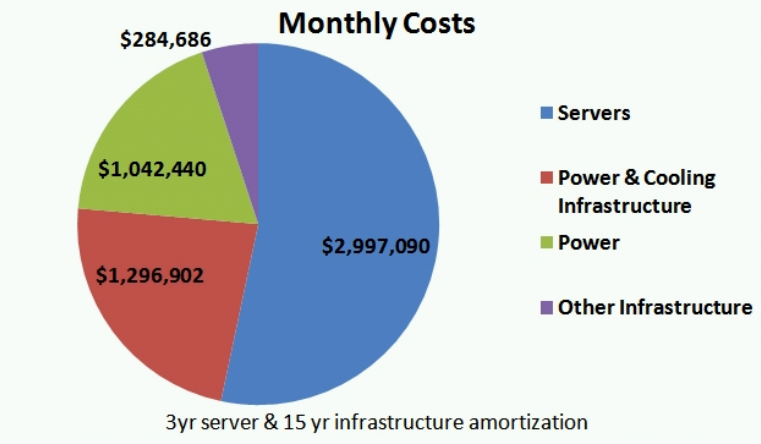

Data Center Challenges

- Resource management
  - How to efficiently use server and storage resources?
  - Many apps have variable, unpredictable workloads
  - Want high performance **and** low cost
  - Automated resource management
  - Performance profiling and prediction

- Energy Efficiency
  - Servers consume huge amounts of energy
  - Want to be "green"

Data Center Costs

- Running a data center is expensive



http://perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx

Economy of Scale

- Larger data centers can be cheaper to build and run than smaller ones
  - Lower prices for buying equipment in bulk
  - Cheaper energy rates
- Automation allows small number of sys admins to manage thousands of servers
- General trend is towards larger mega data centers
  - 100,000s of servers

Virtualization

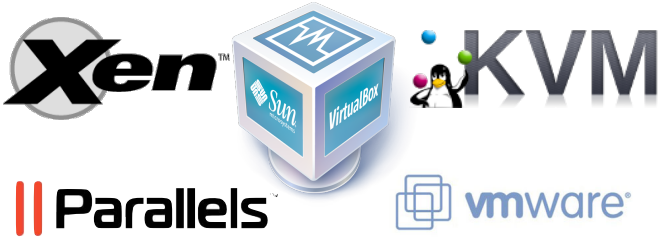
- Separation of a service request from the underlying physical delivery of that service.
- Achieving virtual machine virtualization
  - CPU virtualization
  - Memory virtualization
  - Device and I/O virtualization

Virtualization Timeline

- The pioneer project
  - 1964, Started in IBM Cambridge Science Center, CP-40
- VMWare IA-32 virtual platform
  - 1999, with the company founded in the previous year
- First open source x86 hypervisor
  - 2003, Computer Laboratory, University of Cambridge, Xen
- First professional open source virtualization software
  - 2007, VirtualBox

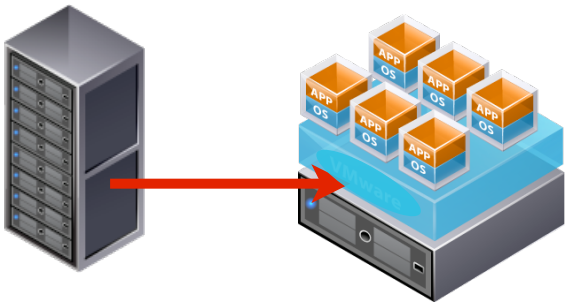
Server Virtualization

- Allows a server to be "sliced" into Virtual Machines
- VM has own OS/applications
- Rapidly adjust resource allocation
- VM migration within a LAN

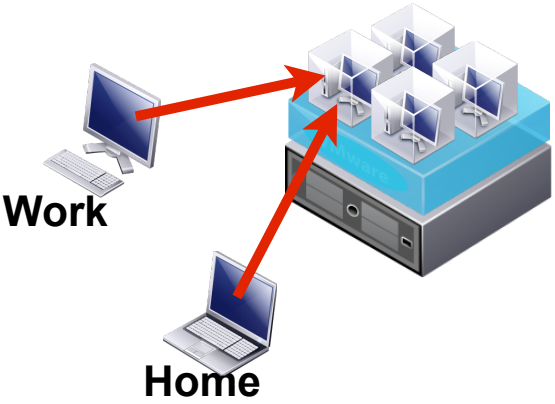


Virtualization in Data Centers

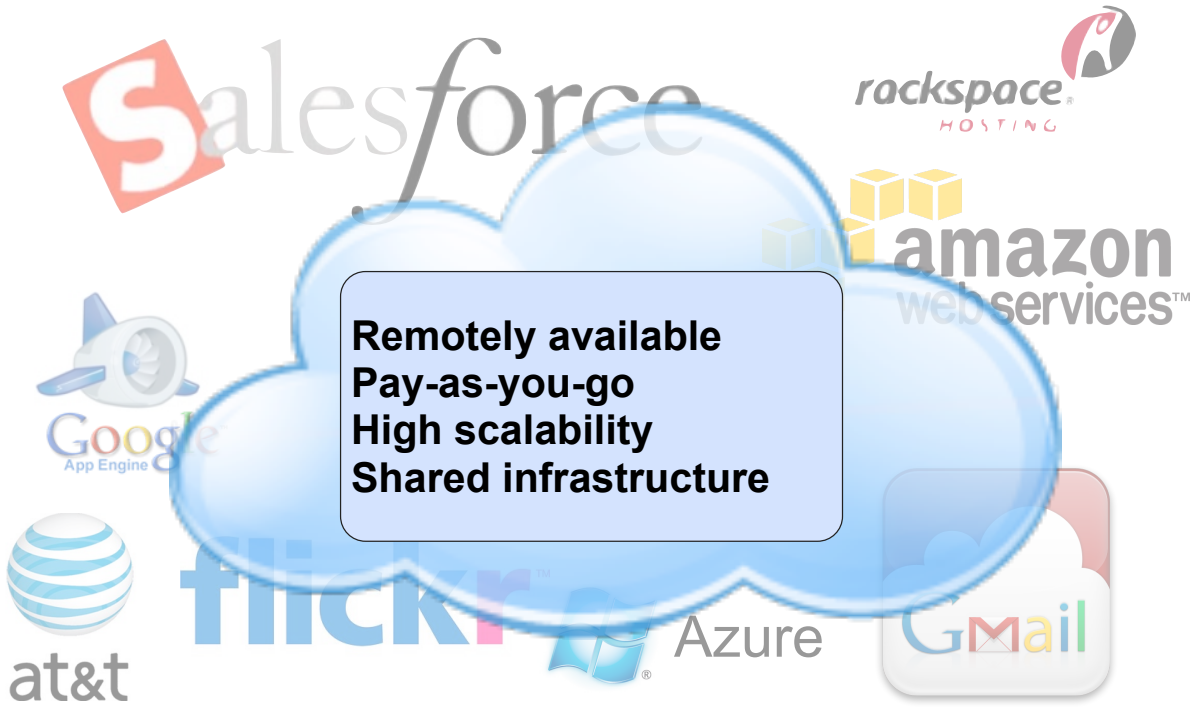
- Virtual Servers
  - Consolidate servers
  - Faster deployment
  - Easier maintenance



- Virtual Desktops
  - Host employee desktops in VMs
  - Remote access with thin clients
  - Desktop is available anywhere
  - Easier to manage and maintain



What is the cloud?



The Cloud Stack

Software as a Service

Hosted applications
Managed by provider

Platform as a Service

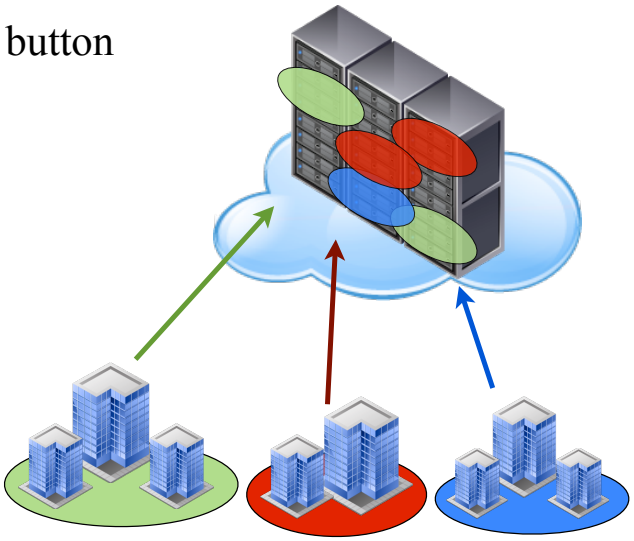
Platform to let you run your own apps
Provider handles scalability

Infrastructure as a Service

Raw infrastructure
Can do whatever you want with it

IaaS: Amazon EC2

- Rents servers and storage to customers
  - Uses virtualization to share each server for multiple customers
  - Economy of scale lowers prices
  - Can create VM with push of a button



Amazon Pricing

- EC2 Instances
  - Different instance types provides different CPU, RAM, storage and networking capacity.
  - On-demand, reserved and spot instances

	t1.micro	r3.4xlarge	r3.8xlarge
VCPUs	1	16	32
RAM	1GB	122GB	244GB
On-demand	$0.013/hr	$1.400/hr	$2.800/hr
Spot	$0.0031/hr	$0.128/hr	$0.256/hr

Storage	$0.10/GB per month

Bandwidth	$0.10 per GB

PaaS: Google App Engine

- Provides highly scalable execution platform
  - Must write application to meet App Engine API
  - App Engine will autoscale your application
  - Strict requirements on application state
    - "Stateless" applications are much easier to scale

- Not based on virtualization
  - Multiple users' threads running in same OS
  - Allows google to quickly increase number of "worker threads" running each client's application

- Simple scalability, but limited control
  - Only supports Java and Python
  - Now also supports PHP and Go

Public or Private

- Not all enterprises are comfortable with using **public cloud** services
  - Don't want to share CPU cycles or disks with competitors
  - Privacy and regulatory concerns
- Private Cloud
  - Use cloud computing concepts in a private data center
    - Automate VM management and deployment
    - Provides same convenience as public cloud
    - May have higher cost
- Hybrid Model
  - Move resources between private and public depending on load

Cloud Challenges

- Privacy / Security
  - How to guarantee isolation between client resources?
- Extreme Scalability
  - How to efficiently manage 1,000,000 servers?
- Programming models
  - How to effectively use 1,000,000 servers?

Programming Models

- Client/Server
  - Web servers, databases, CDNs, etc
- Batch processing
  - Business processing apps, payroll, etc
- Map Reduce
  - Data intensive computing
  - Scalability concepts built into programming model