Data Centers and Cloud Computing

- Data Centers
- Virtualization
- Cloud Computing

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

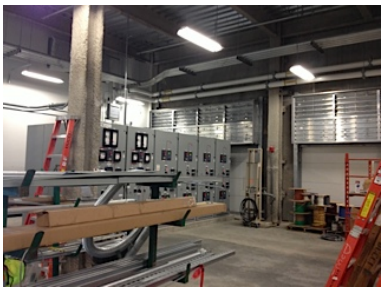
Inside a Data Center

- Giant hardware warehouse

- Racks of servers
- Storage arrays
- Network switches

- Cooling infrastructure
- Power converters
- Backup generators

MGHPCC Data Center

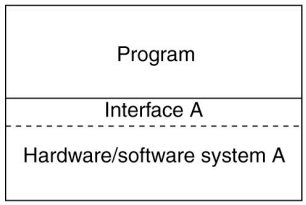


- Data center in Holyoke

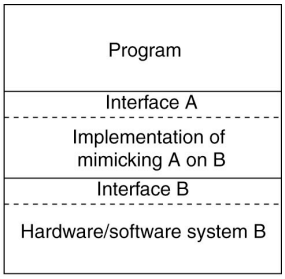
Modular Data Centers

- ...or 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

Virtualization



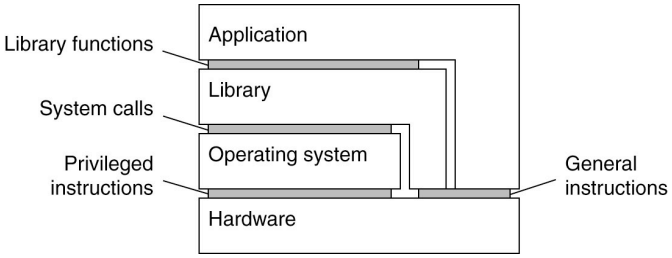
(a)



(b)

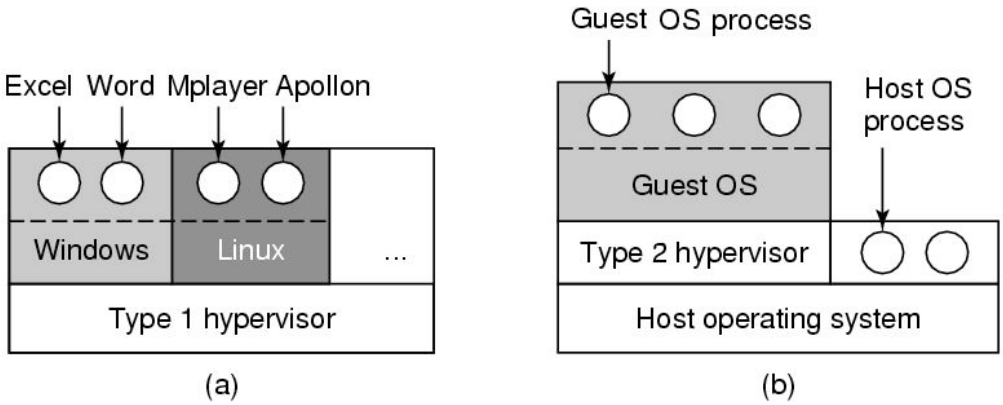
- Virtualization: extend or replace an existing interface to mimic the behavior of another system.
  - Introduced in 1970s: run legacy software on newer mainframe hardware
- Handle platform diversity by running apps in virtual machines (VMs)
  - Portability and flexibility

Types of Interfaces



- Different types of interfaces
  - Assembly instructions
  - System calls
  - APIs
- Depending on what is replaced/mimicked, we obtain different forms of virtualization
- Emulation (Bochs), OS level, application level (Java, Rosetta, Wine)

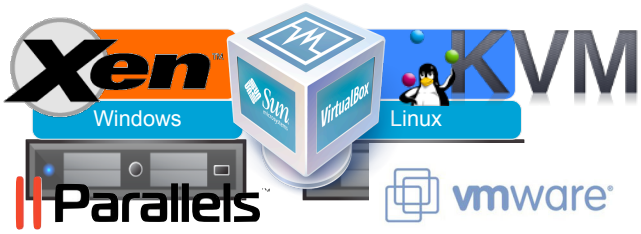
Types of OS-level Virtualization

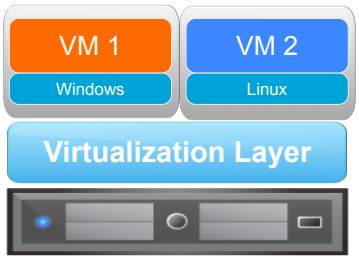


- Type 1: hypervisor runs on "bare metal"
- Type 2: hypervisor runs on a host OS
  - Guest OS runs inside hypervisor
- Both VM types act like real hardware

Server Virtualization

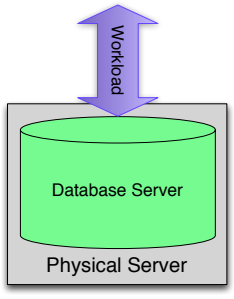
- Allows a server to be "sliced" into **Virtual Machines** (VMs)
- VM has own OS/applications
- Rapidly adjust resource allocation

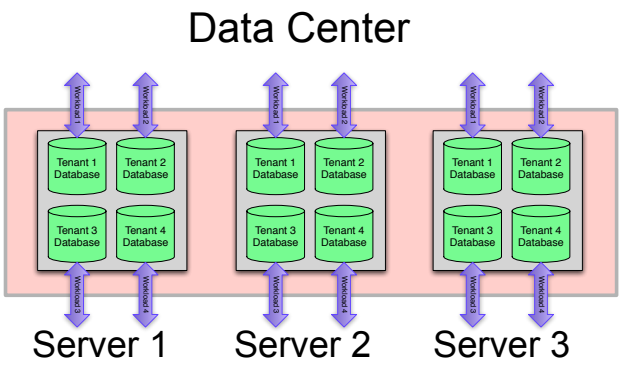




Example: Virtualized Database Servers

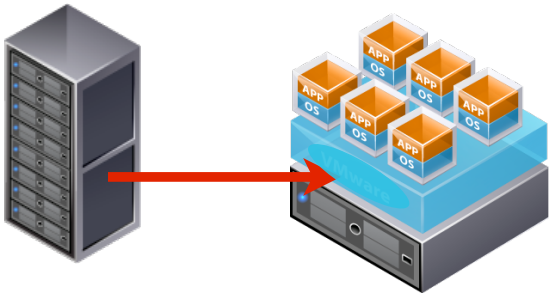
- Conventional: one physical server, one database server
- Data center: multiple physical servers, multiple database servers per (virtualized) physical server



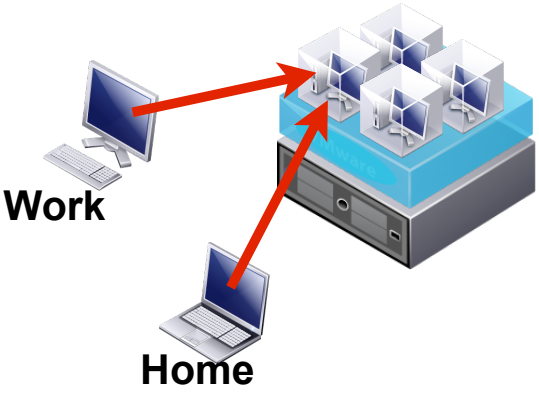


Virtualization in Data Centers

- 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

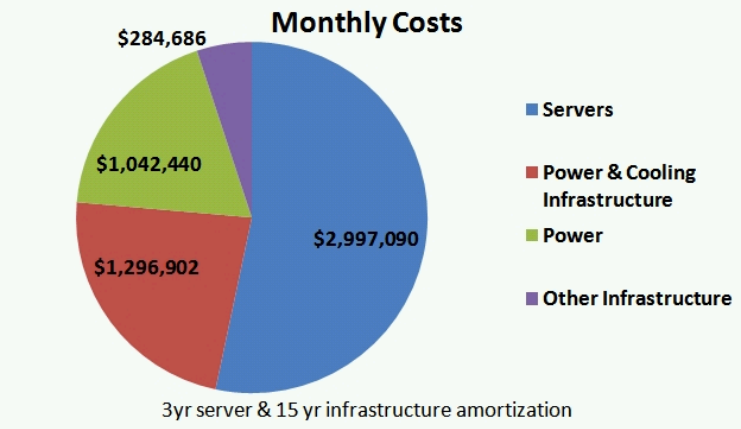


Data Center Challenges

- 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"
  - Want to save money

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 buy 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
- Has helped grow the popularity of **cloud computing**

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

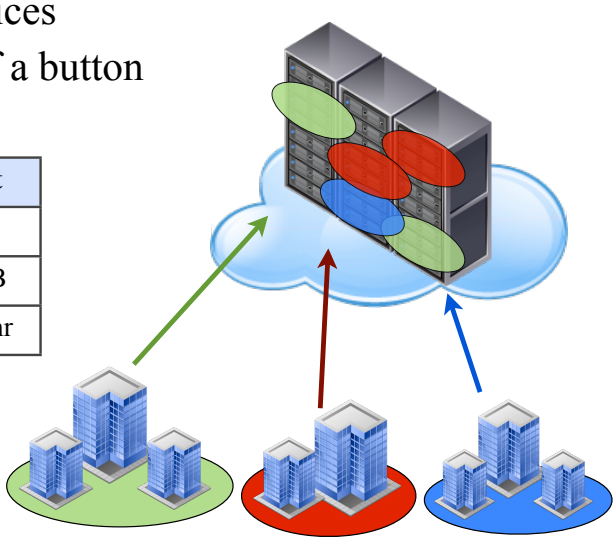
Raw infrastructure
Can do whatever you want with it

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

	Smallest	Medium	Largest
VCPUs	1	5	33.5
RAM	613MB	1.7GB	68.4GB
Price	$0.02/hr	$0.17/hr	$2.10/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 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

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

Programming Models

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

Cloud Challenges

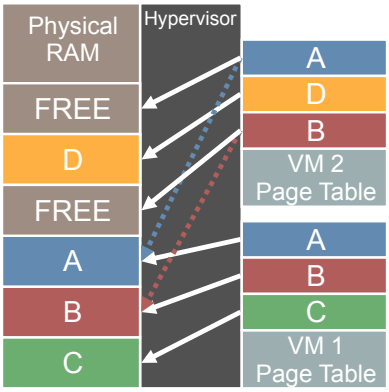
- Privacy / Security
  - How to guarantee isolation between client resources?
- Extreme Scalability
  - How to efficiently manage 1,000,000 servers?
- Programming models
  - How to effectively use 1,000,000 servers?

Challenge: Memory Efficiency

- May be running multiple virtual machines on a single server that have a lot of data in common
- For example, ten copies of Linux in separate VMs
  - Many customers running an Apache webserver
- Can we eliminate duplicated memory?
  - Fit more virtual machines with the same physical resources

Content Based Page Sharing

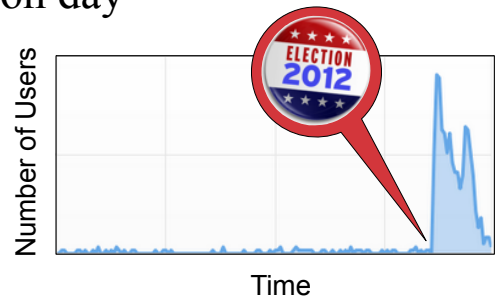
- Approach: eliminate **identical pages** of memory across multiple VMs
- Virtual VM pages mapped to physical pages
- Hypervisor detects duplicates
- Replaced with copy-on-write references



Challenge: Dynamic Workloads

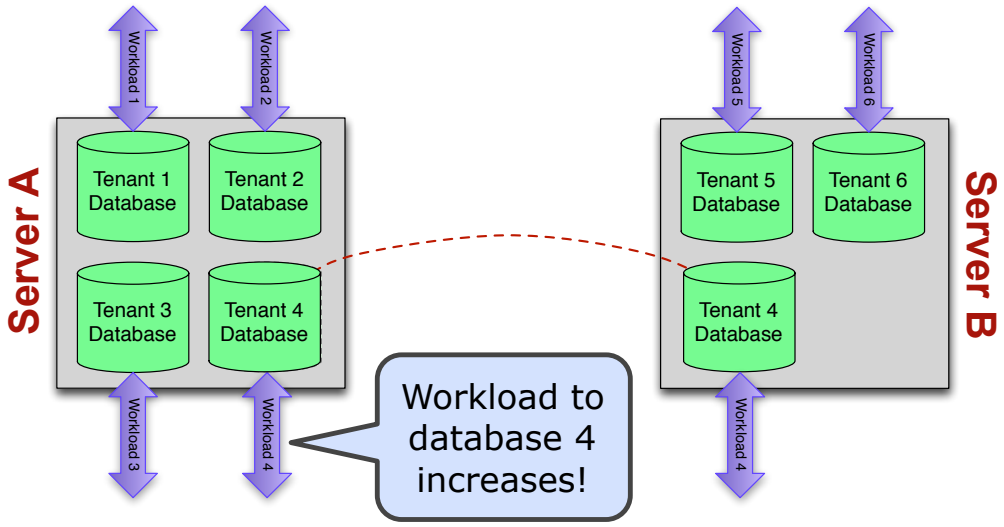
- Server workloads change over time
- Time of day variations
- Flash crowds
- Example: social media on election day

- Workload changes may require more resources!



Virtual Machine Migration

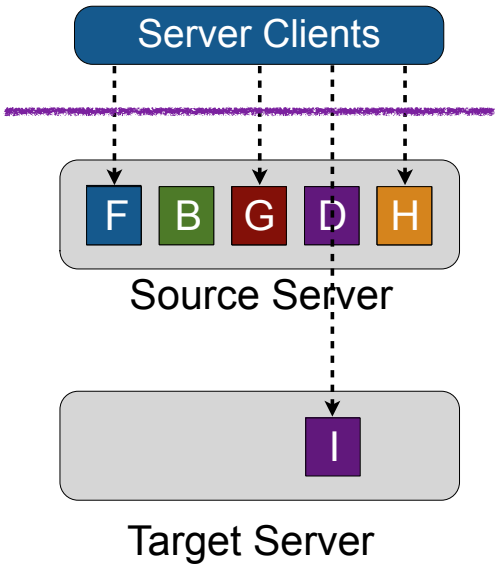
- Approach: move (migrate) a virtual machine from one physical server to another (with more available resources)



- Nice, but incurs downtime!

Live Migration

- Migrate without stopping
  - (1) Copy pages of memory
    - Continue handling workload
  - (2) Update changed pages
    - Multiple rounds
  - (3) Switch workload to target
    - Brief downtime



Summary

- Many services moving to the cloud
  - Remotely available
  - Pay-as-you-go
  - High scalability
- Operating in large, shared data centers
- Data centers use virtualization to increase utilization and decrease costs
- Many challenges in resource management using virtualized data centers