



Data Centers and 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



Traditional vs “Modern”

- Data Center architecture and uses have been changing
- Traditional - static
 - Applications run on physical servers
 - System administrators monitor and manually manage servers
 - Use Storage Array Networks (SAN) or Network Attached Storage (NAS) to hold data
- Modern - dynamic, larger scale
 - Run applications inside virtual machines
 - Flexible mapping from virtual to physical resources
 - Increased automation allows larger scale



Inside a Data Center

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



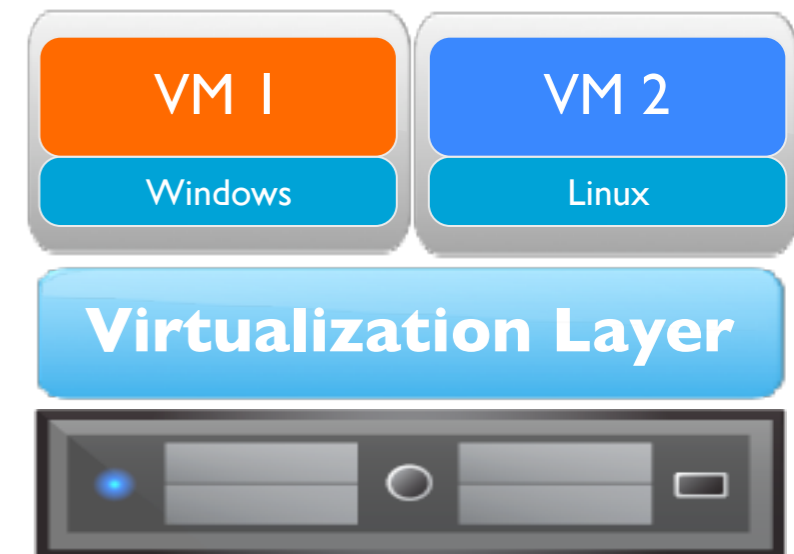
Modular Data Center

- ...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



Server Virtualization

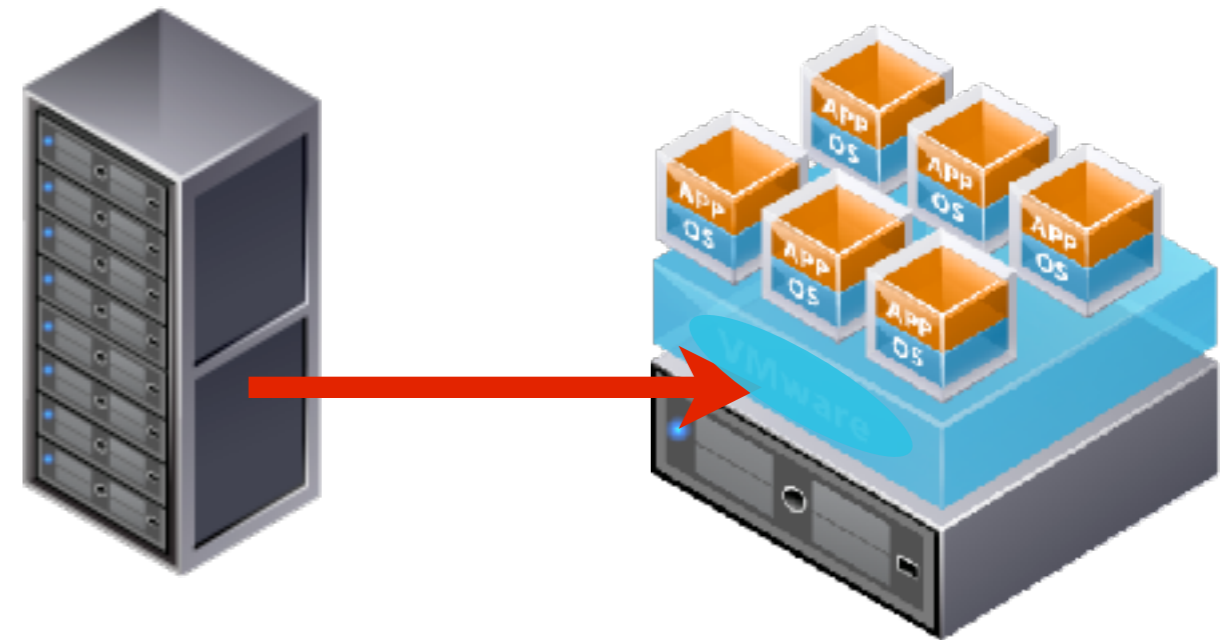
- Allows a server to be “sliced” into Virtual Machines
- VM has own OS/applications
- Rapidly adjust resource allocations
- VM migration within a LAN



Virtualization in Data Centers

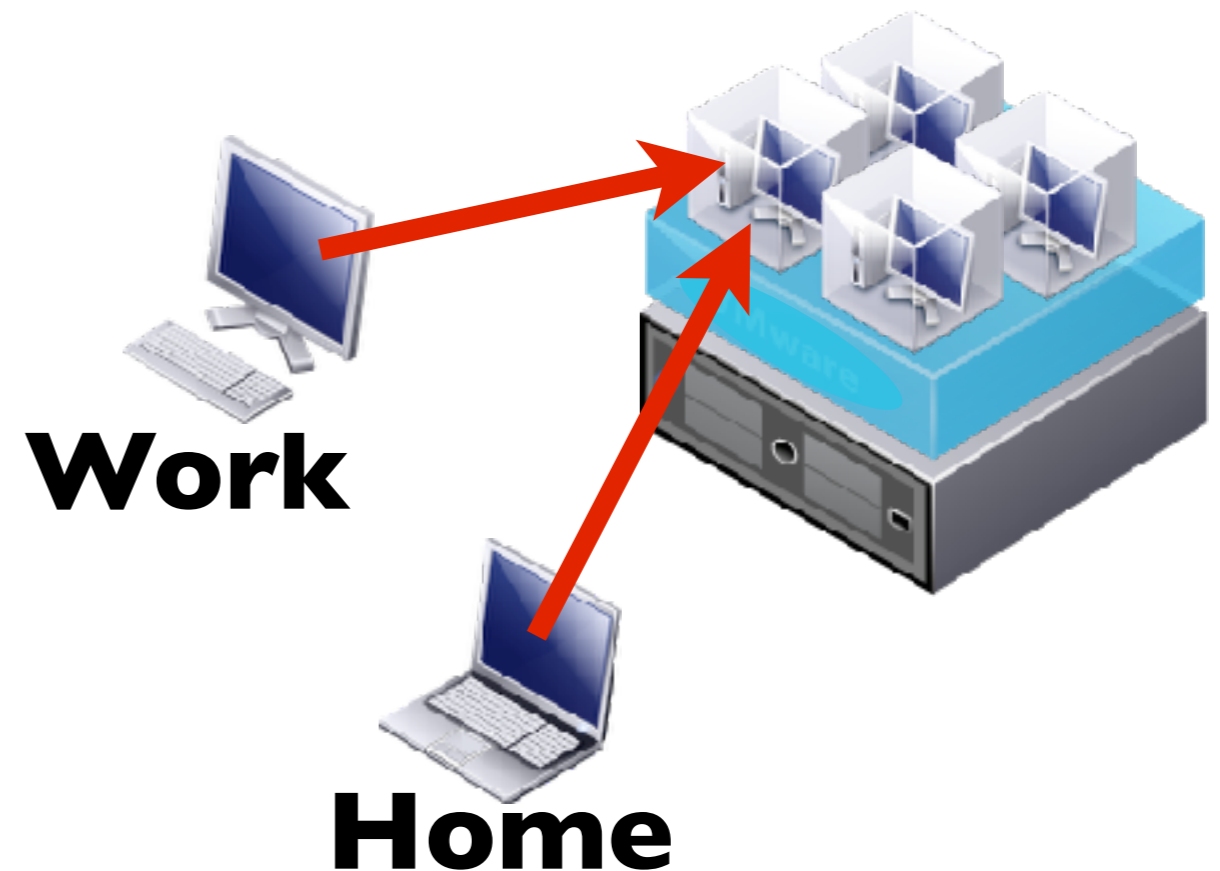
- Virtual Servers

- 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



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



Reliability Challenges

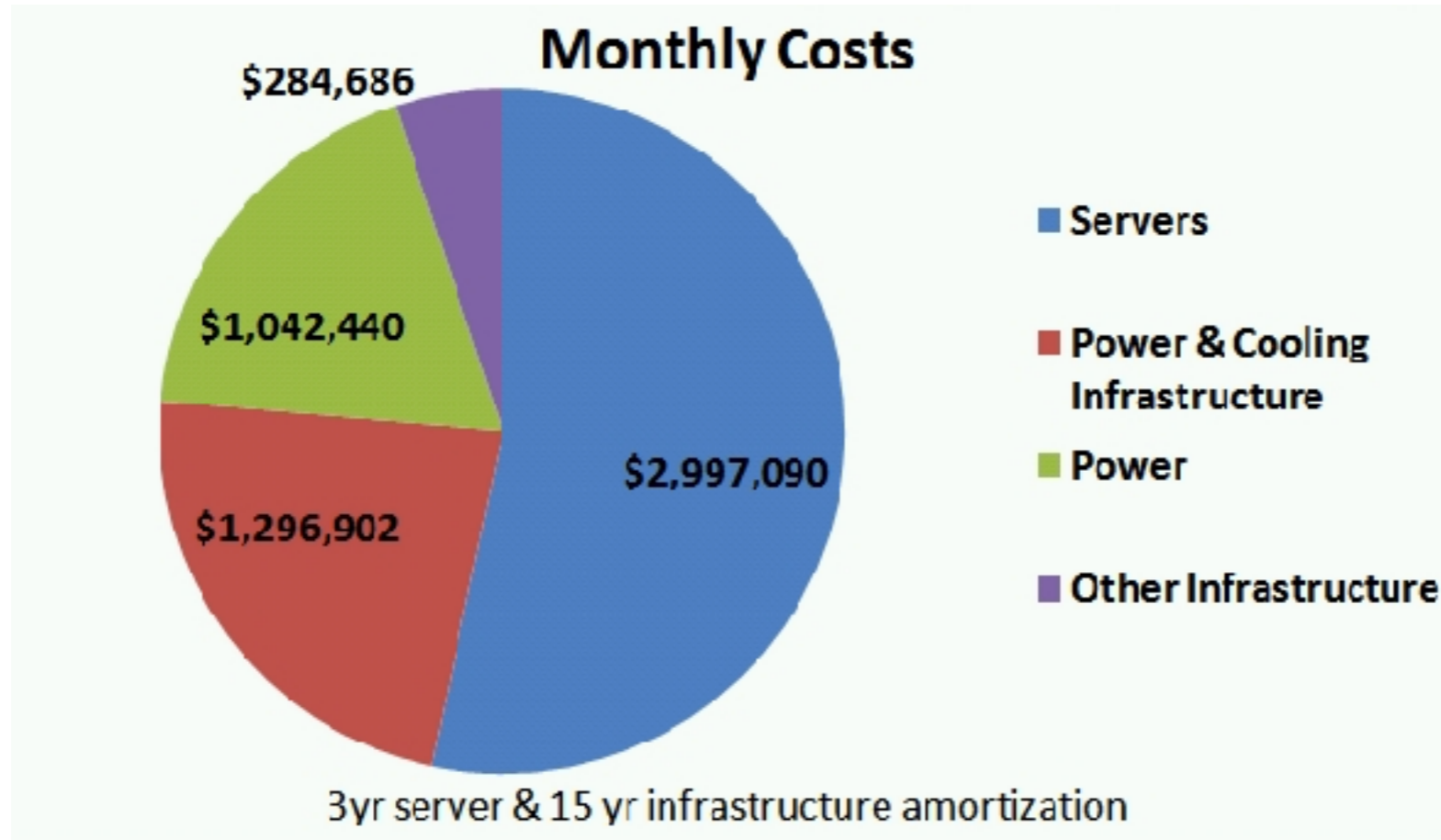
- Typical failures in first year of a google data center:
 - 0.5% overheat (power down most machines in under five minutes, expect 1-2 days to recover)
 - 1 PDU (Power Distribution Unit) failure (about 500-1000 machines suddenly disappear, budget 6 hours to come back)
 - 1 rack-move (You have plenty of warning: 500-1000 machines powered down, about 6 hours)
 - 1 network rewiring (rolling 5% of machines down over 2-day span)
 - 20 rack failures (40-80 machines instantly disappear, 1-6 hours to get back) 5 racks go wonky (40-80 machines see 50% packet loss)
 - 8 network maintenances (4 might cause ~30-minute random connectivity losses)
 - 12 router reloads (takes out DNS and external virtual IP address (VIPS) for a couple minutes)
 - 3 router failures (have to immediately pull traffic for an hour)
 - dozens of minor 30-second blips for DNS
 - 1000 individual machine failures
 - thousands of hard drive failures

http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/people/jeff/stanford-295-talk.pdf



Data Center Costs

- Running a data center is expensive
- Efficiency captured as PUE (Power Usage Effectiveness)
 - Ratio of IT Power / Total Power (typical: 1.7, Google PUE ~ 1.1)



<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?

Salesforce

rockspace
HOSTING

amazon
web services™

Remotely available
Pay-as-you-go
High scalability
Shared infrastructure

Google
App Engine

at&t

flickr™
Azure

GMail

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



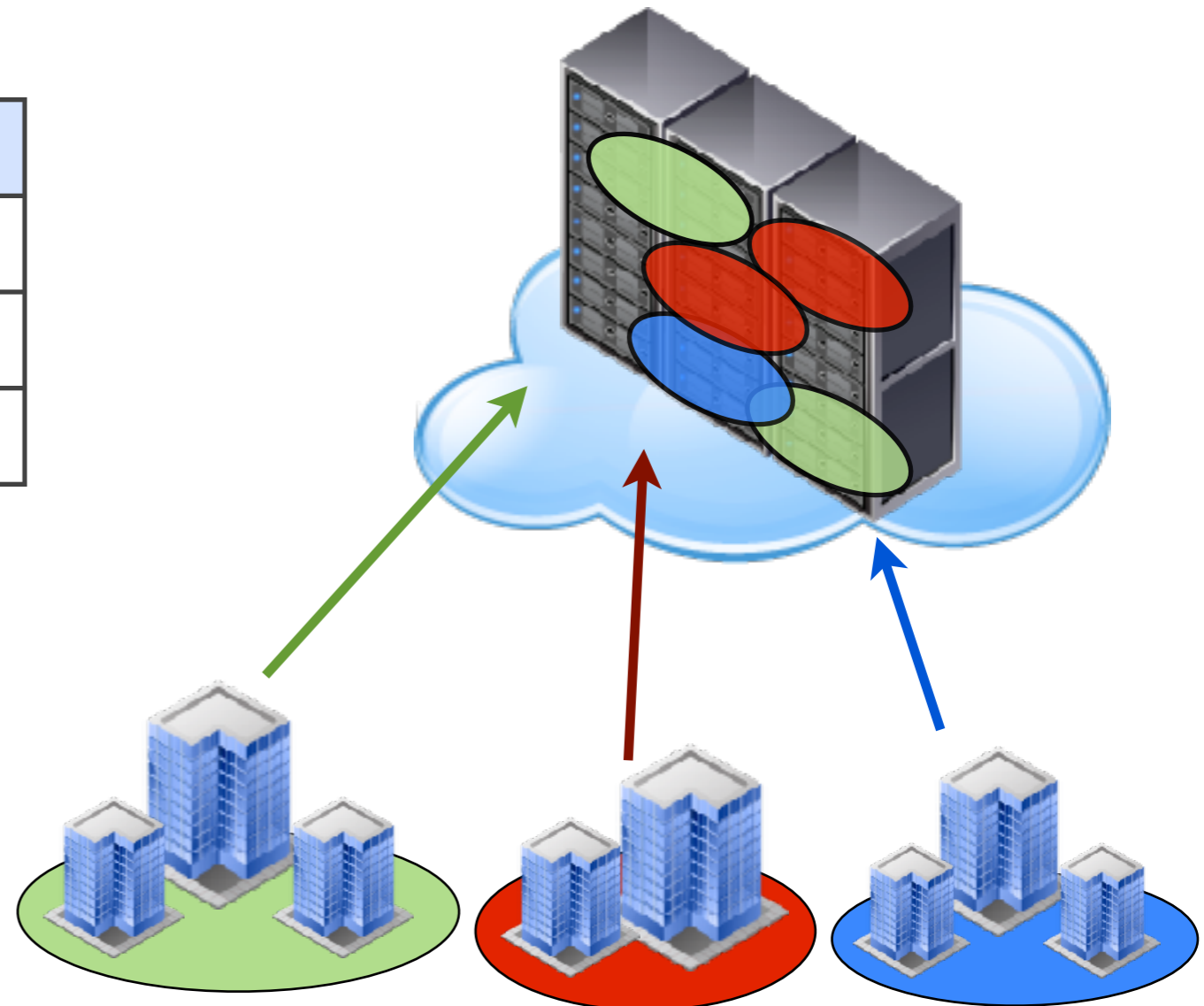
Raw infrastructure
Can do whatever you
want with it

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
 - Calculate here: <https://aws.amazon.com/tco-calculator/>

	Smallest	Largest
VCPUs	1	448
RAM	613MB	12 TB
Price/hr	\$0.02/hr	\$30.539

Storage	SSD vs HDD
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, Go, PHP, 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
 - Cloud Bursting



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

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?



Further Resources

- “Above the Clouds” - cloud computing survey paper from Berkeley
- Workshops & Conferences
 - Hot Topics in Cloud Computing (HotCloud)
 - Symposium on Cloud Computing (SOCC)
 - lots of other small workshops
 - most recent systems conferences (NSDI, USENIX ATC, OSDI, SOSP)
- Other
 - Google App Engine / Amazon EC2 blogs
 - James Hamilton’s Perspectives: <http://perspectives.mvdirona.com/>

