

Cloud Computing Trends

What is cloud computing?




- Cloud computing refers to the apps and services delivered over the internet.
- Software delivered over the internet is usually referred as Software as a Service (SaaS)
 - Salesforce.com
 - Google Calendar
- Cloud usually refers to data center infrastructure that runs those services
- Public cloud is shared by multiple organizations
 - Usually pay-as-you-go based access
 - Example: Amazon Web Services
- Private cloud is generally managed and owned by single organization
- Storage as a Service
 - Amazon S3
- Platform as a Service (PaaS)
 - Microsoft Azure

What is new in cloud computing?

- **Computing=utility?**
 - Analogy to power utility
 - You do not built a power generator in your home
 - No need hire someone to take care of your in-house power generator
 - Pay as much as you use
- Three new aspects
 - The illusion of infinite computing resources
 - No up-front costs to use public clouds
 - Pay-as-you-go models

What is new in cloud computing?

FIG. 2: CLOUD OPPORTUNITY

		Technology	Economic	Business Model
Mainframe		Centralized compute and storage Thin clients	Optimized for efficiency because of the high cost	High up-front costs for hardware and software
Client/Server		PCs and servers for distributed compute, storage, and so on	Optimized for agility because of the low cost	Perpetual license for OS and application software
Cloud		Large DCs, ability to scale, commodity hardware, devices	Efficiency and agility an order of magnitude better	Ability to pay as you go, and only for what you use

Source: Microsoft.

Trends supporting cloud computing

- Mobile interactive applications
 - Respond to information provided by user and sensors in real time
- Rise of analytics and big data
- Parallel batch processing
 - Hadoop, Map-reduce
- New business models
 - Pay-as-you-go

Emergence of Big Data

Big data can generate significant financial value across sectors



SOURCE: McKinsey Global Institute analysis

MGI Big data report

- “Big data has now reached every sector in the global economy. Like other essential factors of production such as hard assets and human capital, much of modern economic activity simply couldn’t take place without it.”
- Big data creates value
 - Creating transparency
 - Enabling experimentation to discover needs, expose variability, and improve performance
 - Segmenting populations to customize actions
 - Replacing/supporting human decision making with automated algorithms
 - Innovating new business models, products, and services
- Big data will create different opportunities in different industries
- To scale to big data, cloud computing technologies will be critical

Cloud computing Infrastructure Variants: Computation Model

	Amazon Web Services	Microsoft Azure	Google AppEngine
Computation model (VM)	<ul style="list-style-type: none">• x86 Instruction Set Architecture (ISA) via Xen VM• Computation elasticity allows scalability, but developer must build the machinery, or third party VAR such as RightScale must provide it	<ul style="list-style-type: none">• Microsoft Common Language Runtime (CLR) VM; common intermediate form executed in managed environment• Machines are provisioned based on declarative descriptions (e.g. which “roles” can be replicated); automatic load balancing	<ul style="list-style-type: none">• Predefined application structure and framework; programmer-provided “handlers” written in Python, all persistent state stored in MegaStore (outside Python code)• Automatic scaling up and down of computation and storage; network and server failover; all consistent with 3-tier Web app structure

Taken from U.C. Berkley Technical Report

Cloud computing Infrastructure Variants: Storage Model

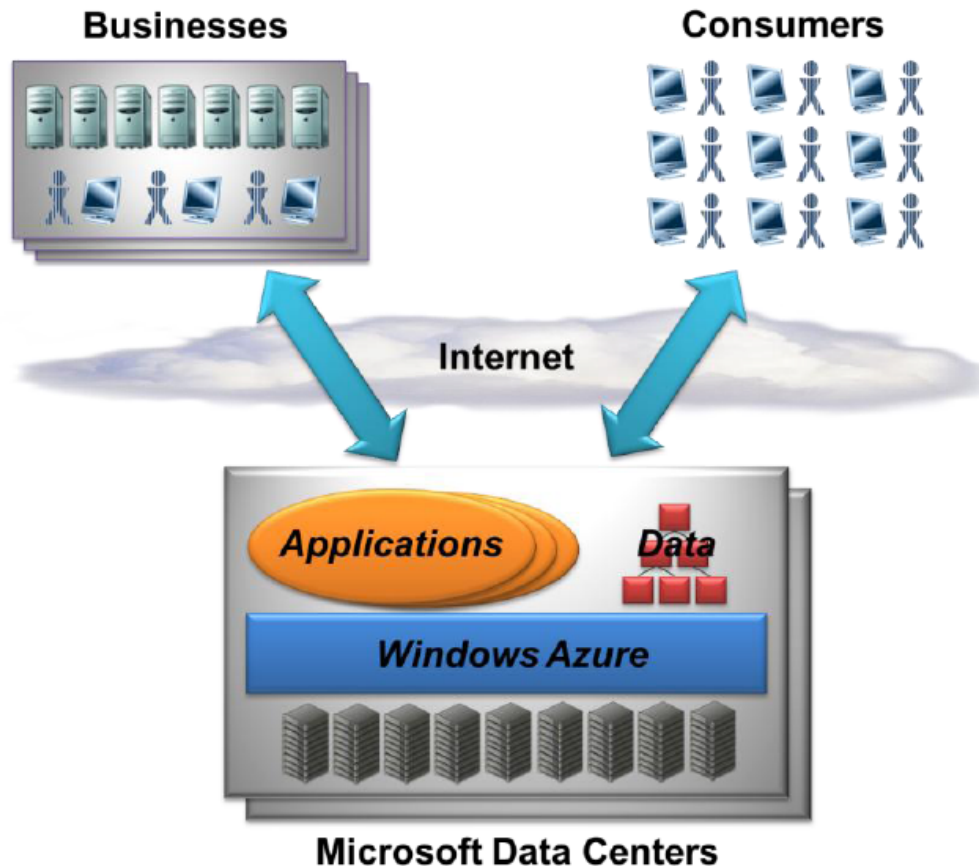
	Amazon Web Services	Microsoft Azure	Google AppEngine
Storage model	<ul style="list-style-type: none">• Range of models from block store (EBS) to augmented key/blob store (SimpleDB)• Automatic scaling varies from no scaling or sharing (EBS) to fully automatic (SimpleDB, S3), depending on which model used• Consistency guarantees vary widely depending on which model used• APIs vary from standardized (EBS) to proprietary	<ul style="list-style-type: none">• SQL Data Services (restricted view of SQL Server)• Azure storage service	<ul style="list-style-type: none">• MegaStore/BigTable

Taken from U.C. Berkley Technical Report

Cloud computing Infrastructure Variants: Networking Model

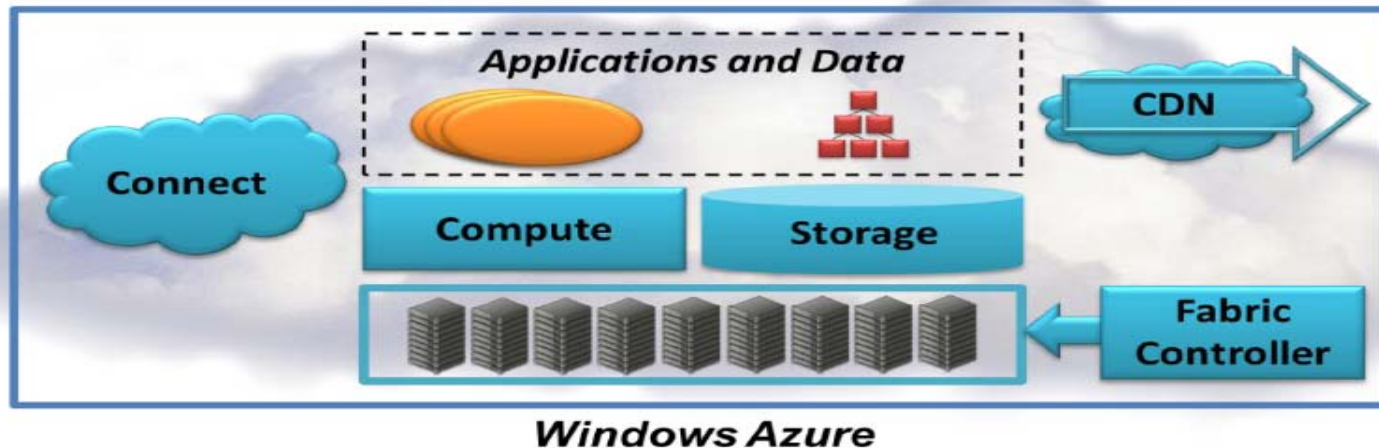
	Amazon Web Services	Microsoft Azure	Google AppEngine
Networking model	<ul style="list-style-type: none">• Declarative specification of IP-level topology; internal placement details concealed• Security Groups enable restricting which nodes may communicate• Availability zones provide abstraction of independent network failure• Elastic IP addresses provide persistently routable network name	<ul style="list-style-type: none">• Automatic based on programmer's declarative descriptions of app components (roles)	<ul style="list-style-type: none">• Fixed topology to accommodate 3-tier Web app structure• Scaling up and down is automatic and programmer-invisible

Example: Microsoft Azure



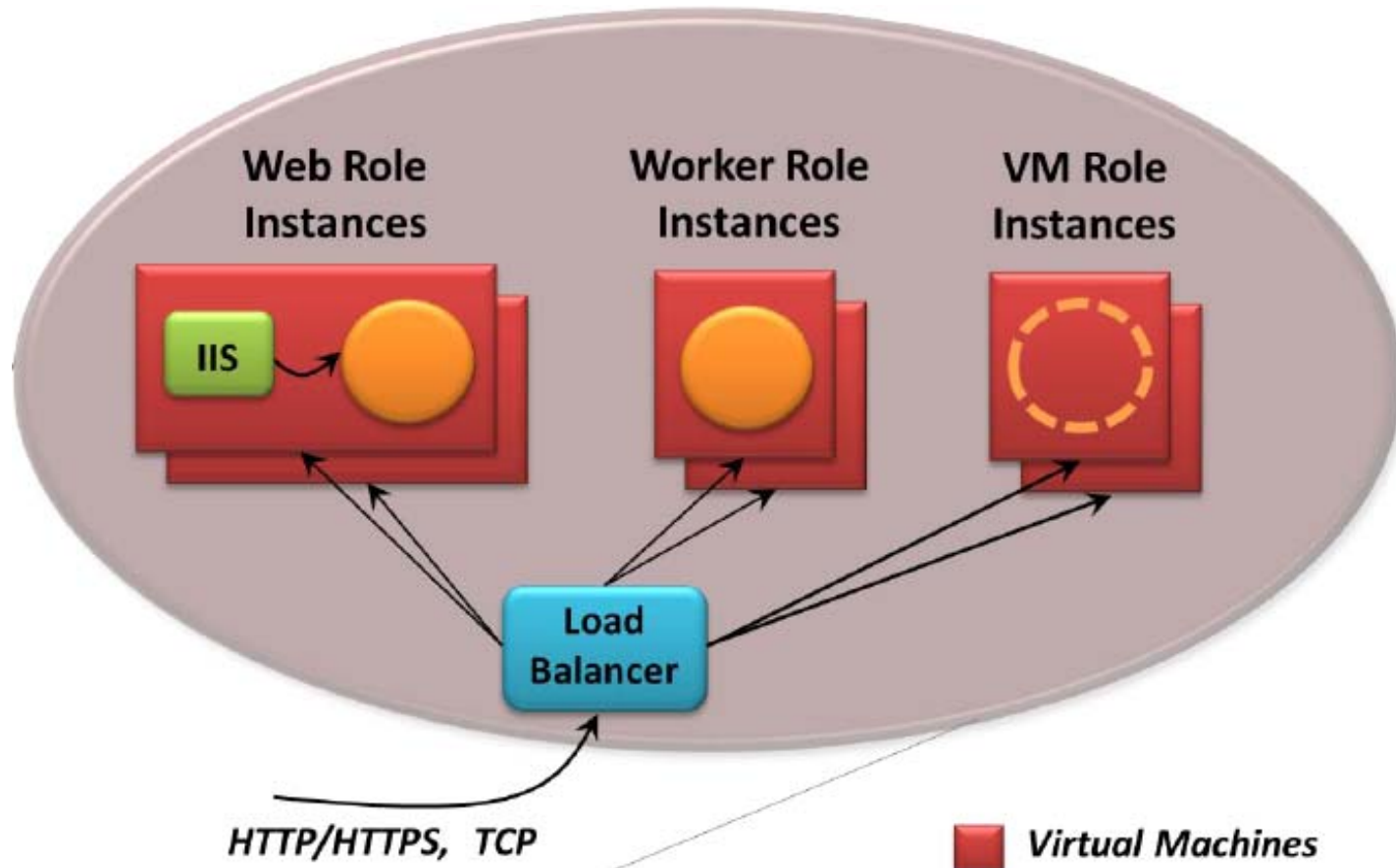
- Platform as a service
- Provides tools for building SaaS
- Software sits on Microsoft Data Centers

MS AZURE Architecture Overview



- **Compute:** runs applications in the cloud. Those applications largely see a Windows Server environment, although the Windows Azure programming model isn't exactly the same as the on-premises Windows Server model.
- **Storage:** stores binary and structured data in the cloud.
- **Fabric Controller:** deploys, manages, and monitors applications.
- **Content Delivery Network (CDN):** speeds up global access to binary data in Windows Azure storage by maintaining cached copies of that data around the world.
- **Connect:** allows creating IP-level connections between on-premises computers and Windows Azure applications.

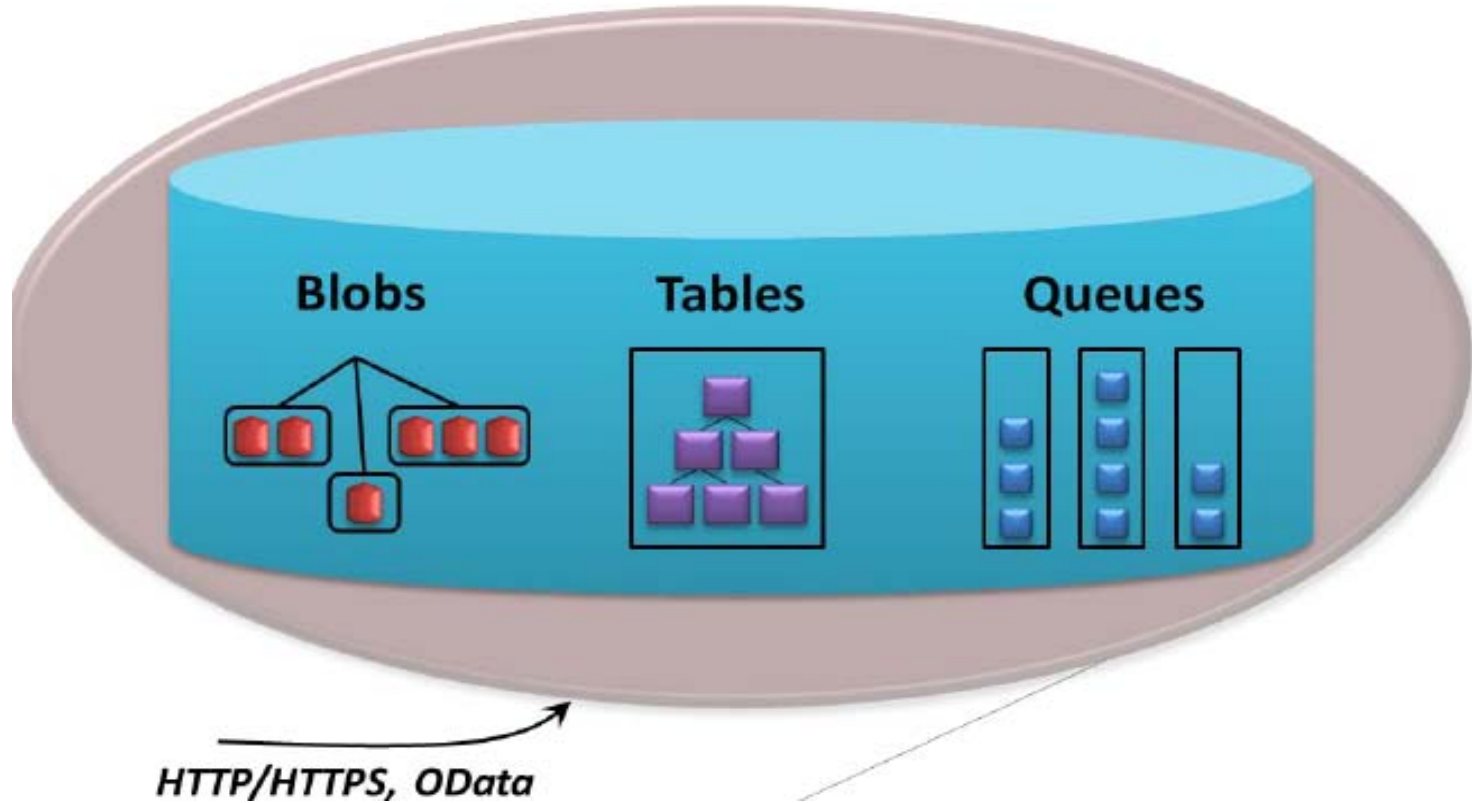
MS AZURE: Compute



MS AZURE: Compute

- Developer needs to submit configuration to define various roles.
- Each role will be assigned a VM by Fabric controller
- Various internet protocols could be used for accessing roles
- Load balancer could assign requests to roles arbitrarily
 - State information needs to be maintained through database
- Number of role instances could be dynamically increased or reduced.
- Monitoring and debugging services are provided by Azure

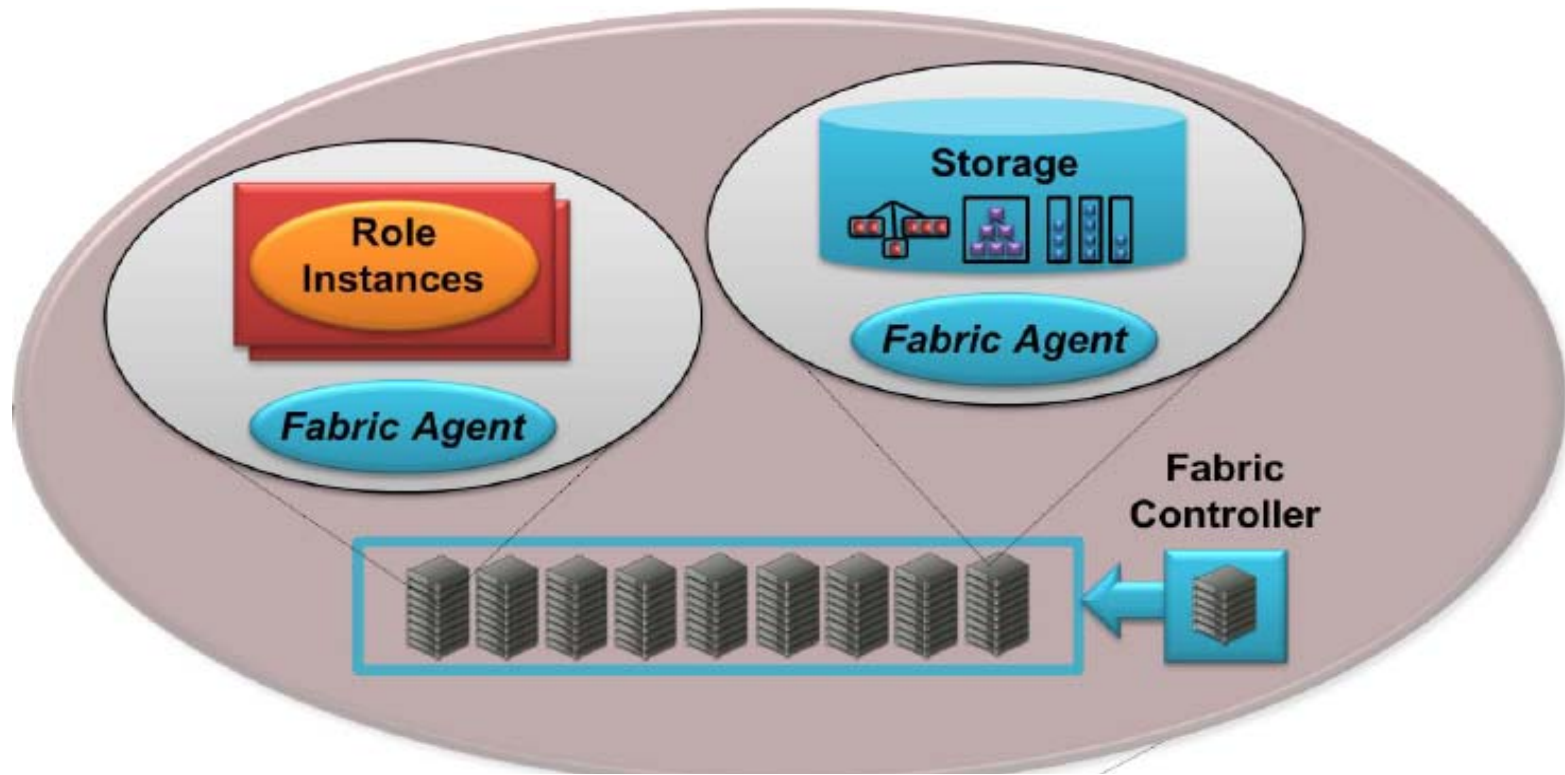
MS Azure: Storage



MS Azure: Storage

- Blobs are used to store binary objects such as videos, images.
 - Similar to Blobs in databases
 - Containers can contain multiple blobs
 - `http://<StorageAccount>.blob.core.windows.net/<Container>/<BlobName>`
 - `<StorageAccount>` is a unique identifier assigned when a new storage account is created, while `<Container>` and `<BlobName>` are the names of a specific container and a blob within that container.
 - It could be public or private
 - For private, you need to authenticate by signing requests
- Tables
 - Similar to Google Bigtable, (many entities, each entity has properties, each property is (name,type, value))
 - Need to use Odata to query
 - Distributed storage to scale to big data
- Queues
 - Allows Web roles to communicate with Worker roles
 - Web role can issue a computation request
 - Worker role can return results using other queues
- Replication is done automatically
 - Three replicas similar to Hadoop
 - Back up copy of all data is kept in another DC
- Provides restful interface to access data
 - Simple Http calls are enough to access data

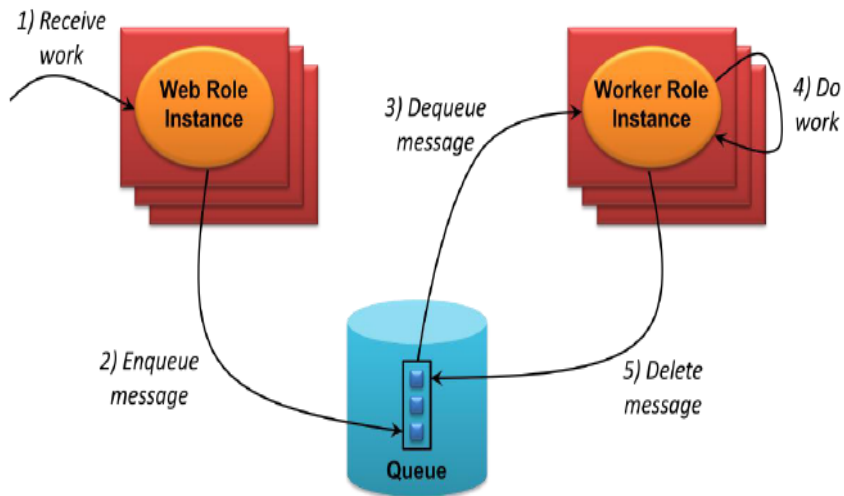
MS Azure: Fabric Controller



MS Azure: Fabric Controller

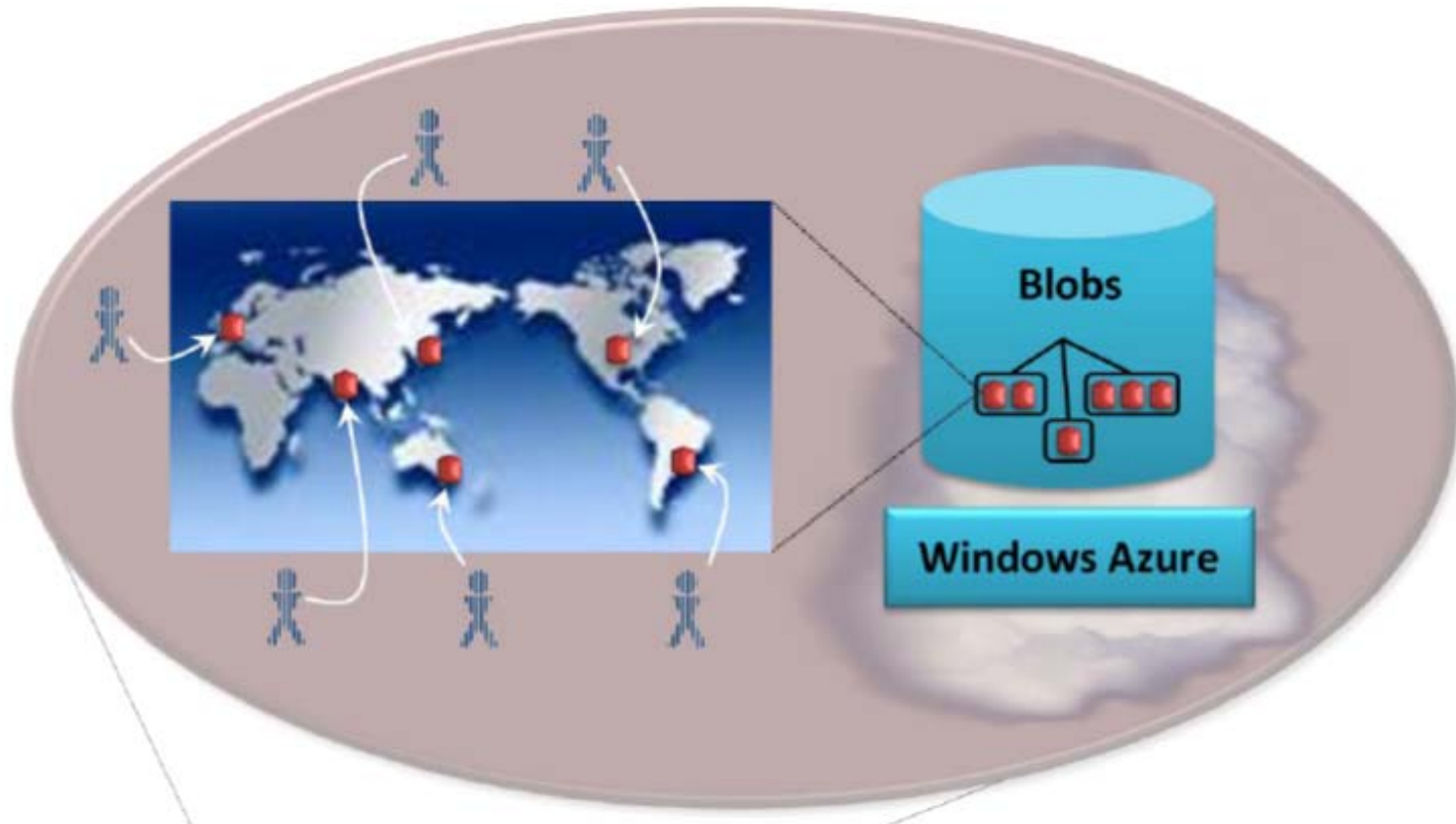
- Controls all resources including load balancers, computer, switches etc.
- Each computer runs a fabric agent
- Handles load balancing and recovery based on the XML configuration provided by the developer
 - Moves role instances to different VMs for load balancing
 - Add new role instances if some of them dies
 - Allocates role instances to different size machines (e.g., extra small to, extra large)
 - Instances are grouped to prevent single point of failure
- Updates and patches for VMs running web and worker role instances are handled automatically
 - Assumes at least two instances are running for each role
 - Stops one for upgrades while the other instance is running
- Updates and patches for applications running on Azure, update domains are created.
 - Fabric controller stops machines in one update domain, updates the app and moves on to next domain.

MS Azure: Queues



- Web roles instance sends a message to worker role using queues.
- Work role instance deletes the message after it is done.
 - Why?
- Not a typical queue.
 - Each message can appear multiple times
 - No first in/ first out semantics
 - No guarantees on ordering of messages

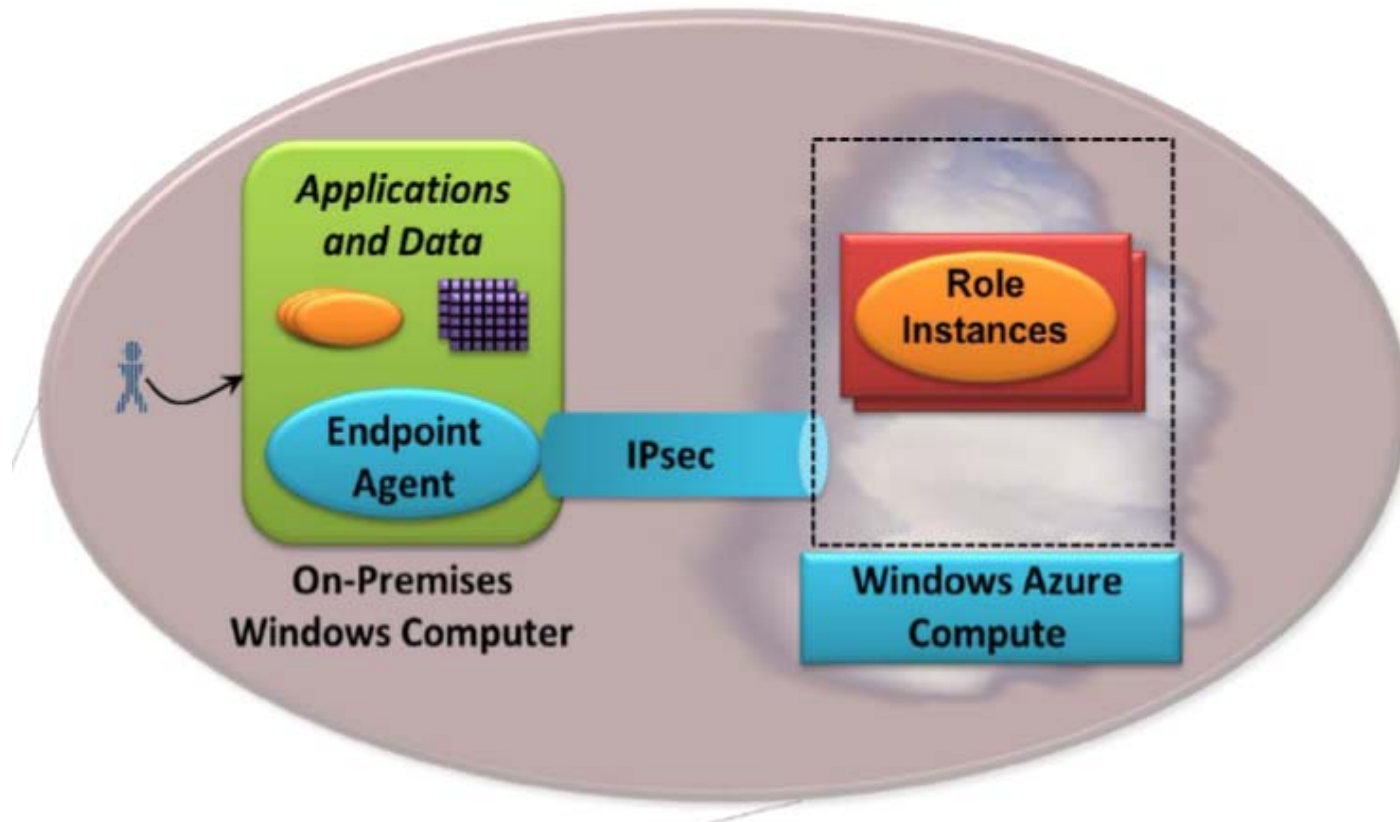
MS Azure: CDN



MS Azure: CDN

- Provides capabilities similar to traditional CDN
- Caches blob content closer to actual users
- Useful for delivering multimedia files
- Enables better experience for end-users

MS Azure: Connect

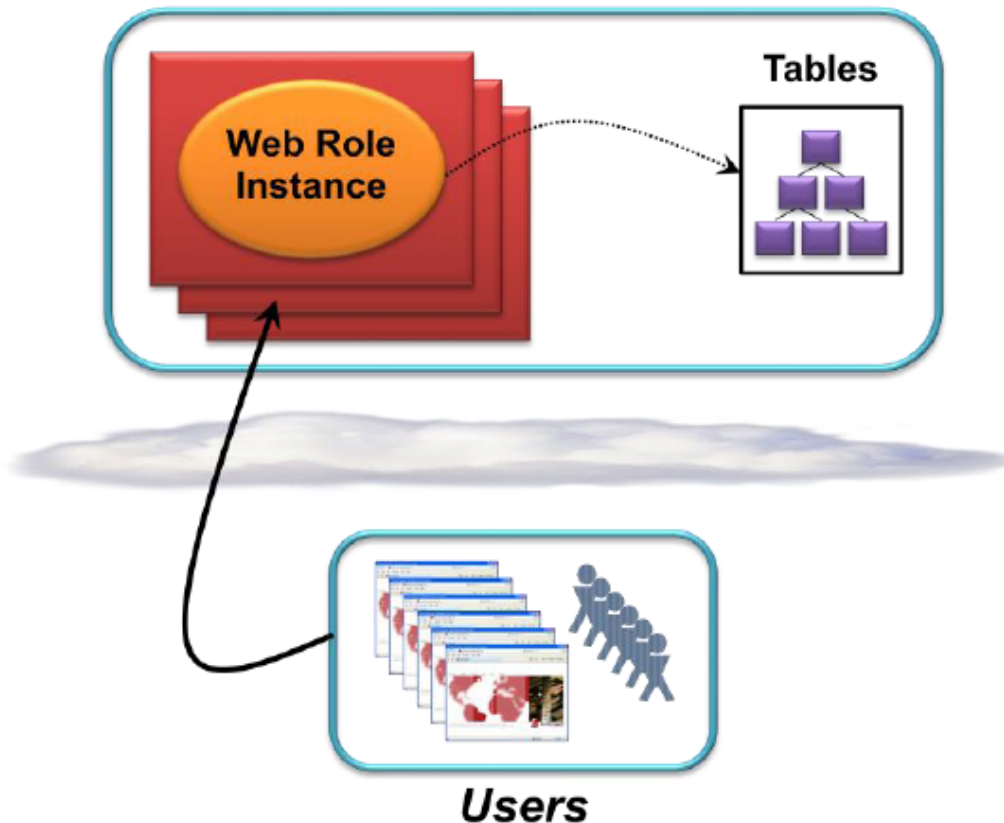


MS Azure: Connect

- Allows connecting local resources with Azure
- Need to run Azure endpoint agent on Windows
- Connections is done through IP V6
- Once connected Azure application appears to run on the same IP network
 - Useful for connecting in-house databases with Azure roles

MS Azure: Scenarios

Scalable Web Application



Allows a typical web app to scale up and down as needed

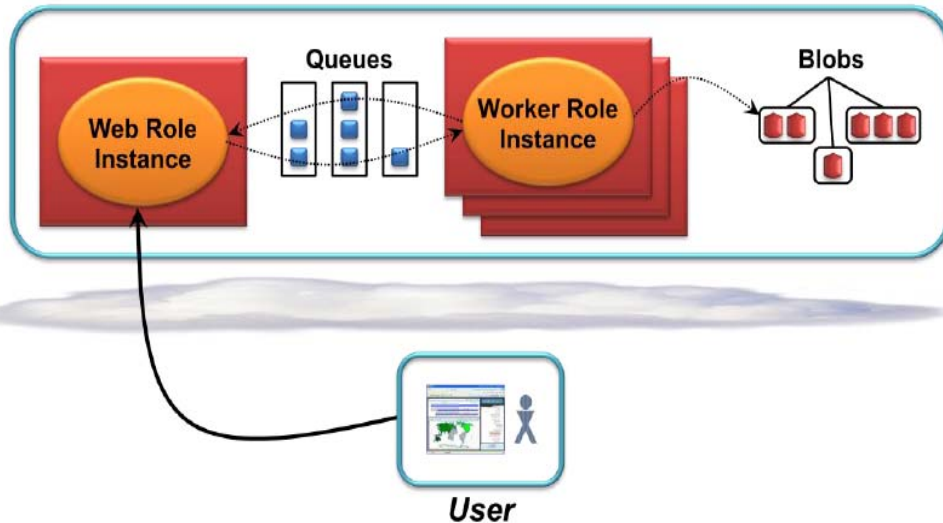
- Increase/decrease the web roles as needed

Easy to scale to large demands

All load balancing and management done by Fabric controller

MS Azure: Scenarios

Parallel Processing Application



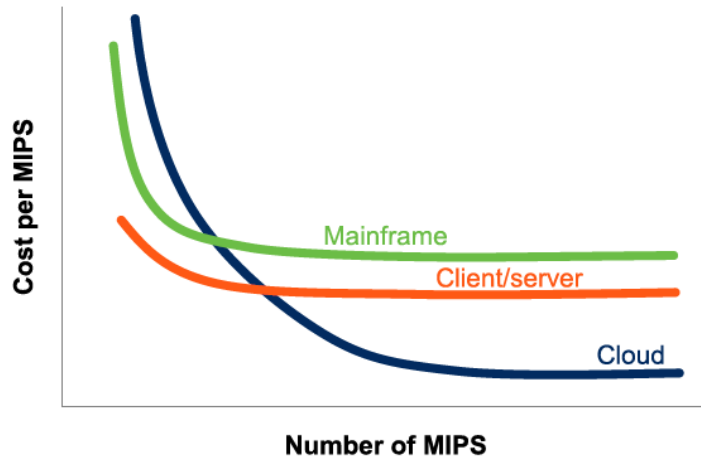
- Parallel processing can be done easily by creating multiple worker roles
 - Note that running 10000 machines one hour, costs similar to running one machine 1000 hours
- Queues are used to communicate messages

Economics of Cloud Computing

- Provides economies of scale in three areas:
 - Supply-side savings
 - Large scale data centers cost much less
 - Demand-side aggregation
 - Smooths the variability
 - Multi-tenancy efficiency
 - Maintenance cost is divided over multiple users

Supply-side Economies of Scale

FIG. 4: ECONOMIES OF SCALE (ILLUSTRATIVE)



Source: Microsoft.

- Cost of power
 - Usually DCs are located in low power cost areas.
 - 1 KWH= 3.6 cents in Idaho versus 1KWH=18.0 in Hawaii
- Infrastructure labor costs
 - An admin can service thousand of computers easily
- Buying Power
 - Google pays less than you do for an Intel CPU 😊

Company	Location	Cost (\$ in millions)	Size (in sq. feet)
Internet Villages JUL 2009	Annandale, Scotland	1,600	3,000,000
National Security Admin. JUL 2009	Camp Williams, Utah	2,000	1,000,000
Lockerbie Data Centers DEC 2009	Lockerbie, Scotland	1,500	N/A
Microsoft SEP 2009	Chicago, Illinois	500	700,000
I/O Data Centers JUN 2009	Phoenix, Arizona	N/A	538,000
Apple MAY 2009	Maiden, North Carolina	1,000	500,000
Microsoft JUN 2010	Dublin, Ireland	500	N/A
U.S. Social Security Admin. FEB 2009	Baltimore, Maryland	400	N/A
Facebook FEB 2010	Prineville, Oregon	N/A	307,000
Next Generation Data MAR 2010	Cardiff, Wales	301	N/A

Sources: Press releases.

Supply Side Economics

Table 2: Economies of scale in 2006 for medium-sized datacenter (≈ 1000 servers) vs. very large datacenter ($\approx 50,000$ servers). [24]

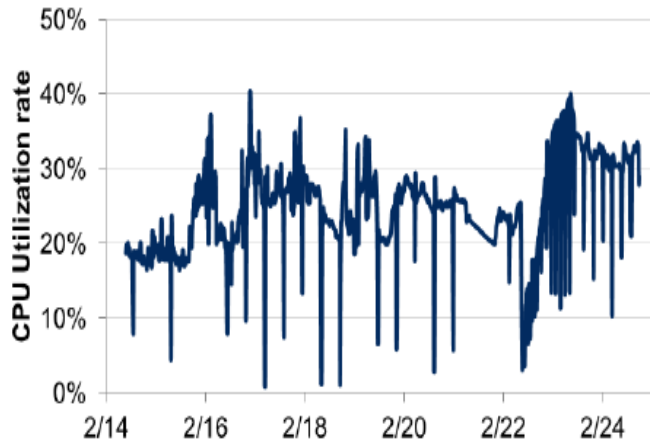
Technology	Cost in Medium-sized DC	Cost in Very Large DC	Ratio
Network	\$95 per Mbit/sec/month	\$13 per Mbit/sec/month	7.1
Storage	\$2.20 per GByte / month	\$0.40 per GByte / month	5.7
Administration	≈ 140 Servers / Administrator	> 1000 Servers / Administrator	7.1

Demand Side Economies of Scale

- Utilization is critical for efficiency
 - In non-virtualized world, typically each app runs on its dedicated server
 - Typically utilization is low
- Low utilization reasons
 - Randomness (people check their Facebook pages at different times)
 - Time of the day pattern (people watch Netflix in the evening more often)
 - Industry Specific Variability
 - Uncertain Growth Patterns

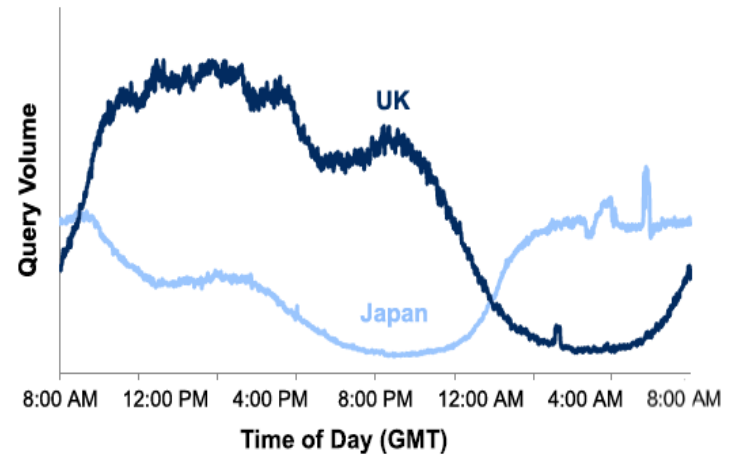
Demand side of economies of scale

FIG. 6: RANDOM VARIABILITY (EXCHANGE SERVER)



Source: Microsoft.

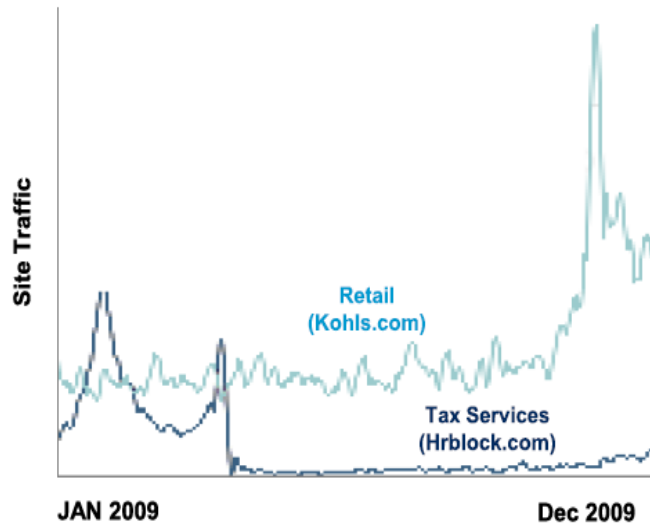
FIG. 7: TIME-OF-DAY PATTERNS FOR SEARCH



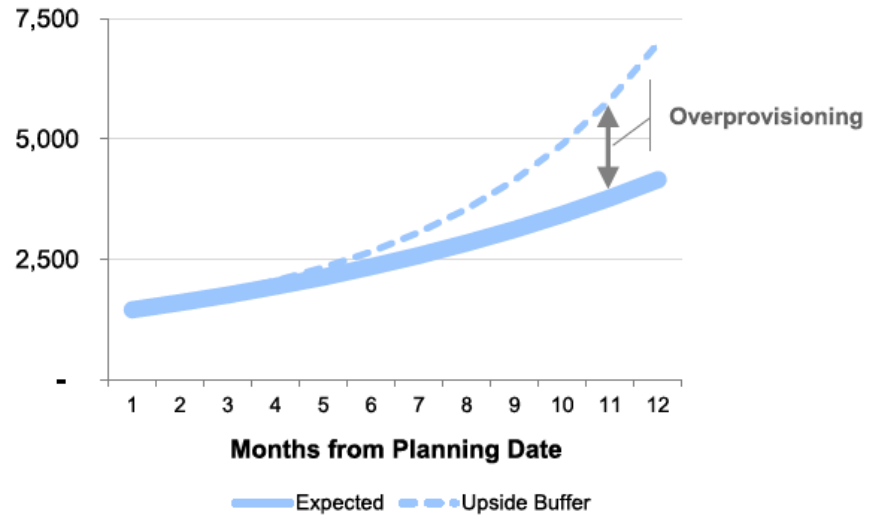
Source: Bing Search volume over 24-hour period.

Demand side economies of scale

FIG. 8: INDUSTRY-SPECIFIC VARIABILITY



Source: Alexa Internet.



Source: Microsoft.

Demand Side Economies of Scale

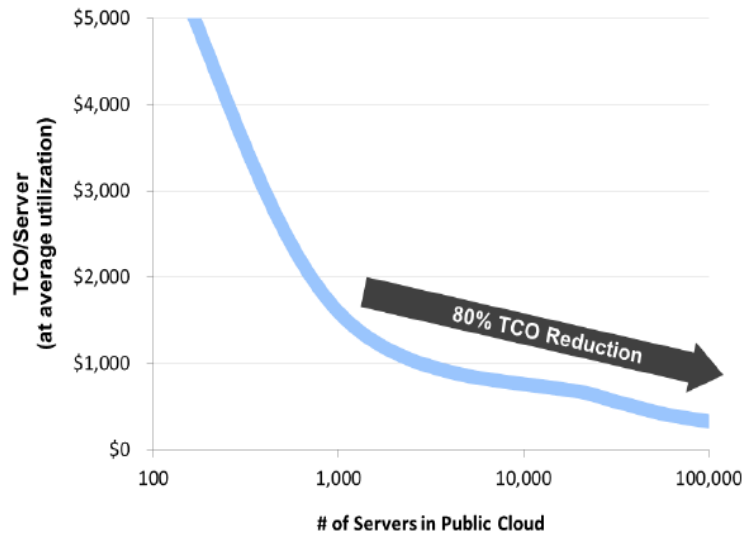
- Aggregating over multiple organization, applications, and industries may
 - reduce the variability in demand
 - Increase resource utilization
 - Prevent overprovisioning
 - For example, peak demand/average traffic approximately 4 in general retail
- Flexible use of money and resources
 - 1000 machines x one hour usage= 1 machine x 1000 hour usage
- Shifting the risk in terms of resource estimation to cloud
- No up front investment
- Pay-as-you-go models

Multi-tenancy Economies of Scale

- Fixed application labor amortized over a large number of customers
 - Cost is shared by multiple organizations
- Variability in demand could be further decreased by Multi-tenancy if the tenants are from different industries

Overall Impact

FIG. 15: ECONOMIES OF SCALE IN THE CLOUD



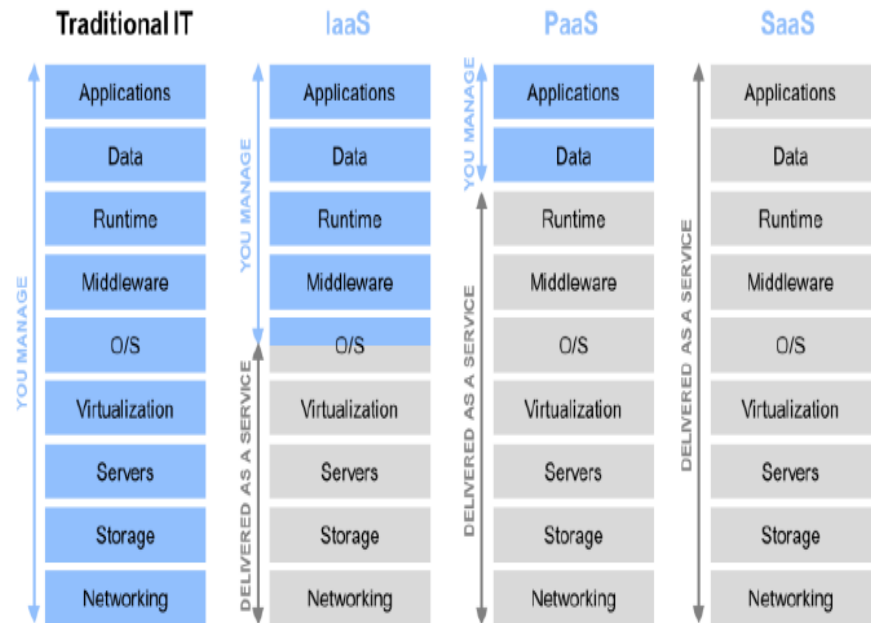
Source: Microsoft.

- Due to economies of scale, large DC may have up to %80 reduction in Total cost of ownership.
- For some apps, moving to cloud can create significant cost savings.

Moving to Cloud?

- As we have seen in the Azure case, creating apps that can leverage cloud could be challenging
- SaaS does not require such development
 - MS Office 365-S

FIG. 17: CAPTURING CLOUD BENEFITS



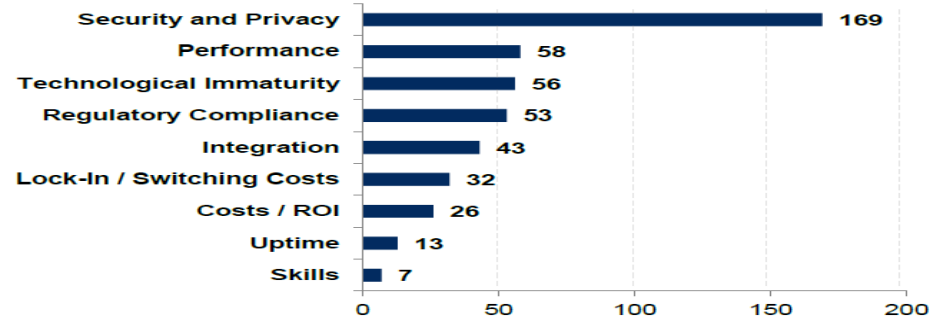
Source: Microsoft.

Possibilities

- Elasticity could be important for some apps
 - Running an multiple instances at the same time
 - New massively parallel applications
- Elimination of capital expenditure
 - Critical for new start-ups
 - No need to buy infrastructure to create your next idea.
- Reduction of Complexity
 - You do not have to manage your infrastructure
 - You do not need to manage your on web server etc.

Obstacles

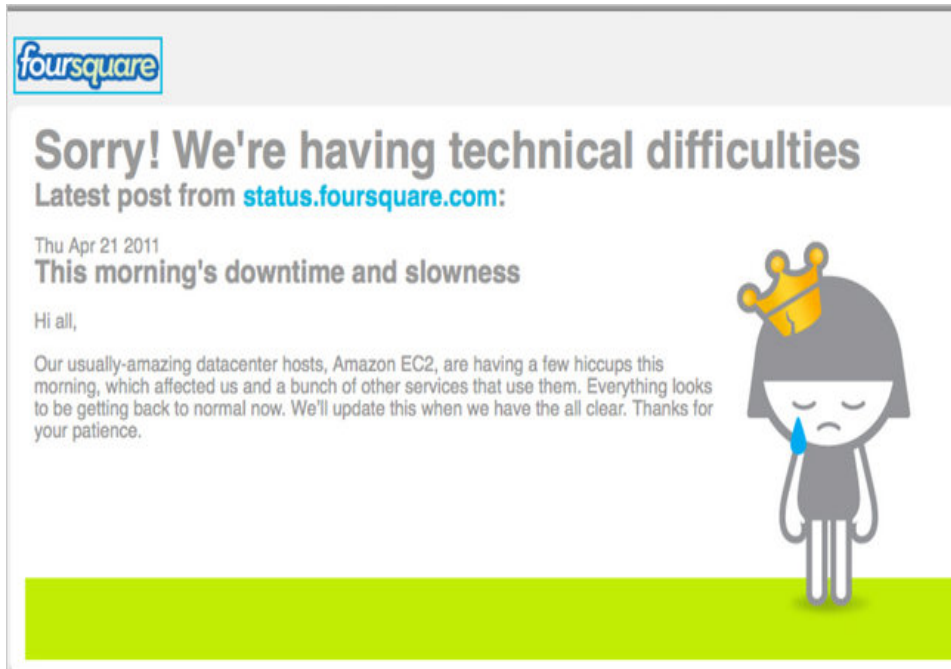
FIG. 19: PUBLIC CLOUD CONCERNS



Source: Gartner CIO survey

- There are many obstacles in cloud adoption
- Possible perception issues related to cloud

Obstacle: Availability of Service



- Too big to fail in Cloud computing
- Need to run your apps on multiple providers to prevent single point of failure
- **Opportunity:** Cloud elasticity can make DDS attacks more costly
 - Suppose EC2 can handle 500 bots
 - Attack generates 1GB/sec traffic using 500,000 bots costing 15000\$
 - At AWS attack will cost 360\$ per hour in bandwidth and \$100 per hour in computation need 1000 EC2 instances)
 - After 32 hours, cost of the defense is larger than attack
 - Attacker needs to sustain attacks longer.

Obstacle: Data Lock-In

- Your data may die with the company
 - Linkup shut down after losing 45% of customer data
 - It turns out Linkup used another service called Nirvanix.
- Once you lock into a cloud provider, they can increase the prices
- Possible solutions
 - Standard APIs and tools for data and app migration
 - Support of Hybrid Models

Obstacle: Security, Privacy and Compliance

- Possible security and privacy issues related to data that is pushed to public cloud
 - This will be our main focus during the rest of the class.
- Will Amazon fight to protect your data against Government subpoenas?
- Compliance issues (e.g., HIPAA)
- Possible solutions
 - Encryption (we will spend four weeks on encryption related solutions
 - New cloud auditing solutions
 - VPNs in the cloud
 - Application and VM firewalls
 - Location aware data storage
 - Keep your data in European jurisdiction that provides higher privacy guarantees.

Obstacle: Data Transfer Bottlenecks

- Transferring large data (10TB) to Amazon may take long time on 20 Mbit/sec connection
 - $10 \cdot 10^{12} / ((20 \cdot 10^6) / 8) = 4,000,000$ seconds approximately 45 day!!!
 - Overnight shipping of hard disk would be much faster 😊
- Possible solutions:
 - Keep all your data in the cloud.
 - Amazon now stores some publicly available data sets.
 - Faster and cheaper WANs

Obstacle: Performance unpredictability

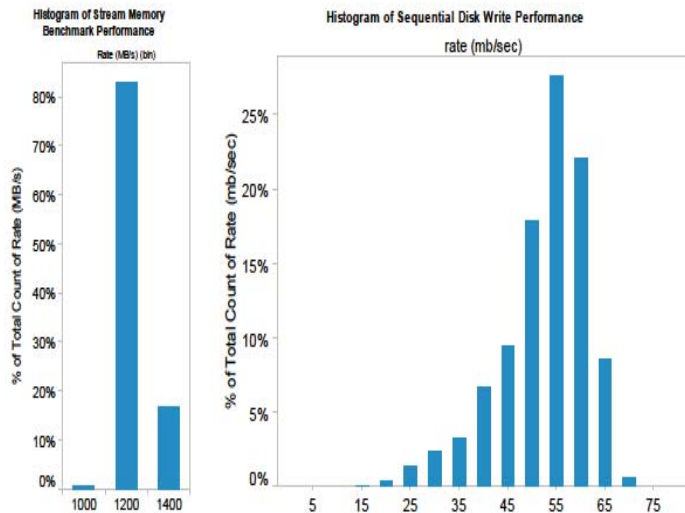


Figure 3: (a) Memory benchmark performance on 75 Virtual Machines running the STREAM benchmark on left and (b) Disk performance writing 1 GB files on 75 Virtual Machines on right.

- If multiple VMs run on the same server, I/O performance can significantly vary.
- **Possible solutions**
 - Better virtualization
 - Flash memory based storage (still somewhat expensive)

Other obstacles

- Storage systems that can easily scale up and down as the demand changes
- Development issues (i.e., bugs) and movement costs
- Scaling automatically and quickly
 - Scaling is currently manual and not trivial
- Bad guys using the cloud
 - Spams coming from EC2
 - Hosting spam web pages on EC2
- Need for new software licensing

Private Cloud Architecture?

FIG. 20: COMPARING VIRTUALIZATION, PRIVATE CLOUD, AND PUBLIC CLOUD

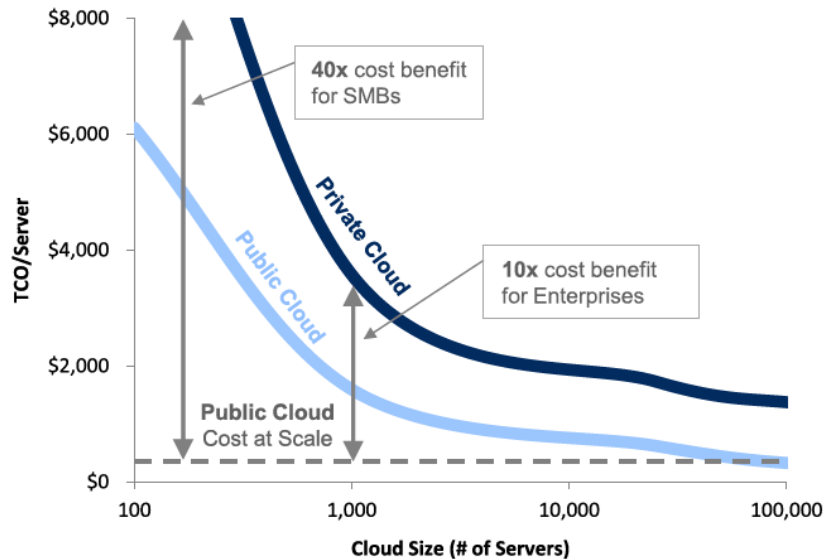
	Operator	Automated Management	Homogenous Hardware	New App Model
Public Cloud	Department, Central IT, Third-party Provider	✓	✓	✓
Private Cloud	Department, Central IT, Third-party Provider	✓	✓	✓
Virtual Server	Department, Central IT, Third-party Provider	✓	X	X
Traditional Server	Department, Central IT, Third-party Provider	X	X	X

Source: Microsoft. Shaded checks indicate an optional characteristic.

- Resources are pooled across company
- Possibility of using in-house versions of Azure

Cost Trade off public versus private

FIG. 22: COST BENEFIT OF PUBLIC CLOUD



Source: Microsoft.

- Private cloud is a cheaper option for large companies
- Still analysis on the left does not consider the security risks and costs,

Other architectures?

- Here at UT Dallas, we advocate for hybrid cloud solutions
 - Sensitive data will be mostly kept in private cloud
 - Sensitive data in public cloud will be encrypted
 - Intelligent query processing techniques will be used efficiently combine the resources of public and private clouds.