Microsoft® Research

Faculty Summit

10

YEAR ANNIVERSARY

Microsoft Cloud Computing Platform

Roger Barga Dennis Gannon Wei Lu

External Research, MSR
Cloud Computing Futures, MSR

Outline

- What is Cloud Computing?
- Data Center Design Issues
- Cloud Software Infrastructure
- A Brief Introduction to Azure
  – Computation in the Cloud
  – Storage Fabric
- Looking to the Future

What is Cloud Computing?

A Definition:

- Cloud Computing means using a remote data center to manage scalable, reliable, on-demand access to applications
- Scalable means
  - Possibly millions of simultaneous users of the app
  - Exploiting thousand-fold parallelism in the app
- Reliable means on-demand means 5 “nines” available right now
- Applications span the continuum from client to the cloud.

Quincy
Chicago
San Antonio

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

The Current Cloud Challenge

The current driver: how do you

- Support email for 375 million users?
- Store and index 6.75 trillion photos?
- Support 10 billion web search queries/month?

And

- deliver deliver a quality response in 0.15 seconds to millions of simultaneous users?
- never go down

The future applications of the cloud go well beyond web search

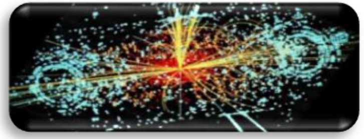
- The data explosion

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

The Future: an Explosion of Data

Experiments

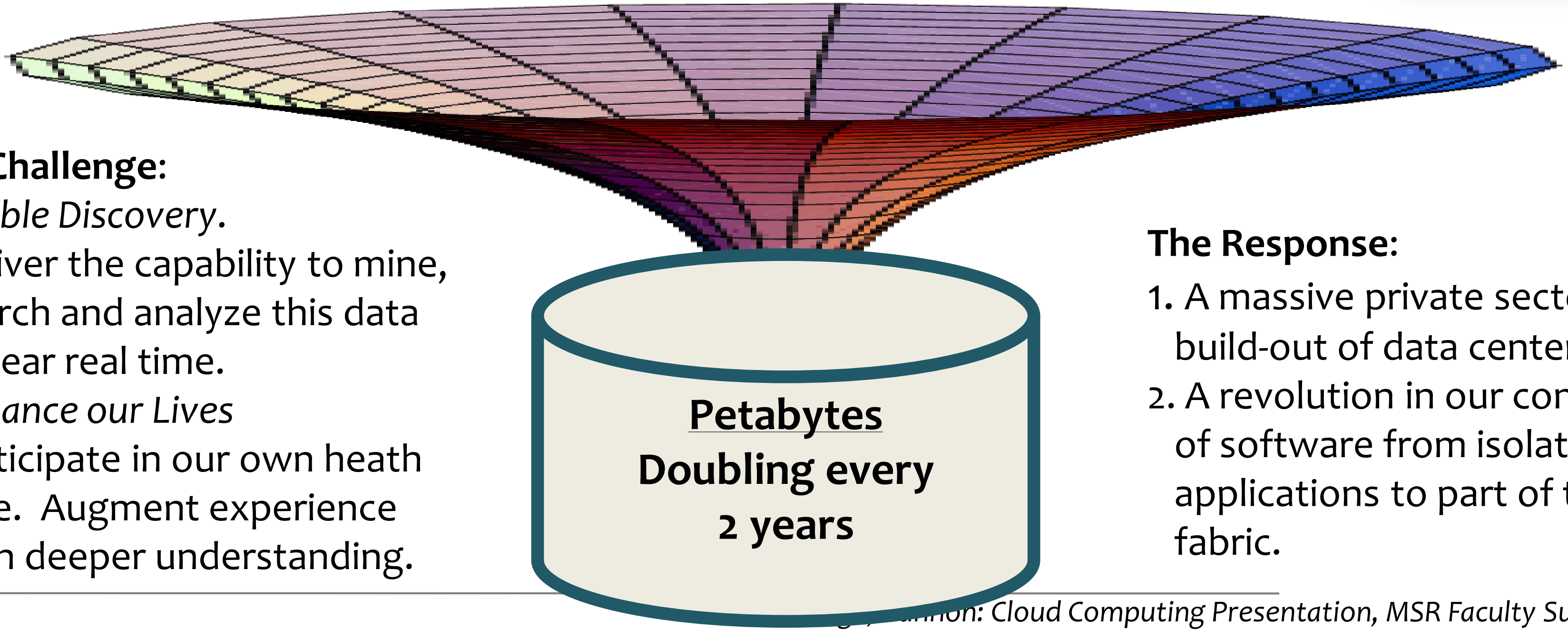
Simulations



Archives

Literature

Consumer



Petabytes

Doubling every 2 years

The Challenge:

Enable Discovery.
Deliver the capability to mine, search and analyze this data in near real time.
Enhance our Lives
Participate in our own heath care. Augment experience with deeper understanding.

The Response:

1. A massive private sector build-out of data centers.
2. A revolution in our concept of software from isolated applications to part of the fabric.

...on: Cloud Computing Presentation, MSR Faculty Summit 2009

The Data Center Landscape

Range in size from “edge” facilities to megascale.

Economies of scale

Approximate costs for a small size center (1K servers) and a larger, 50K server center.

Technology	Cost in small-sized Data Center	Cost in Large Data Center	Ratio
Network	$95 per Mbps/ month	$13 per Mbps/ month	7.1
Storage	$2.20 per GB/ month	$0.40 per GB/ month	5.7
Administration	~140 servers/ Administrator	>1000 Servers/ Administrator	7.1

Each data center is **11.5 times** the size of a football field

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Advances in DC deployment

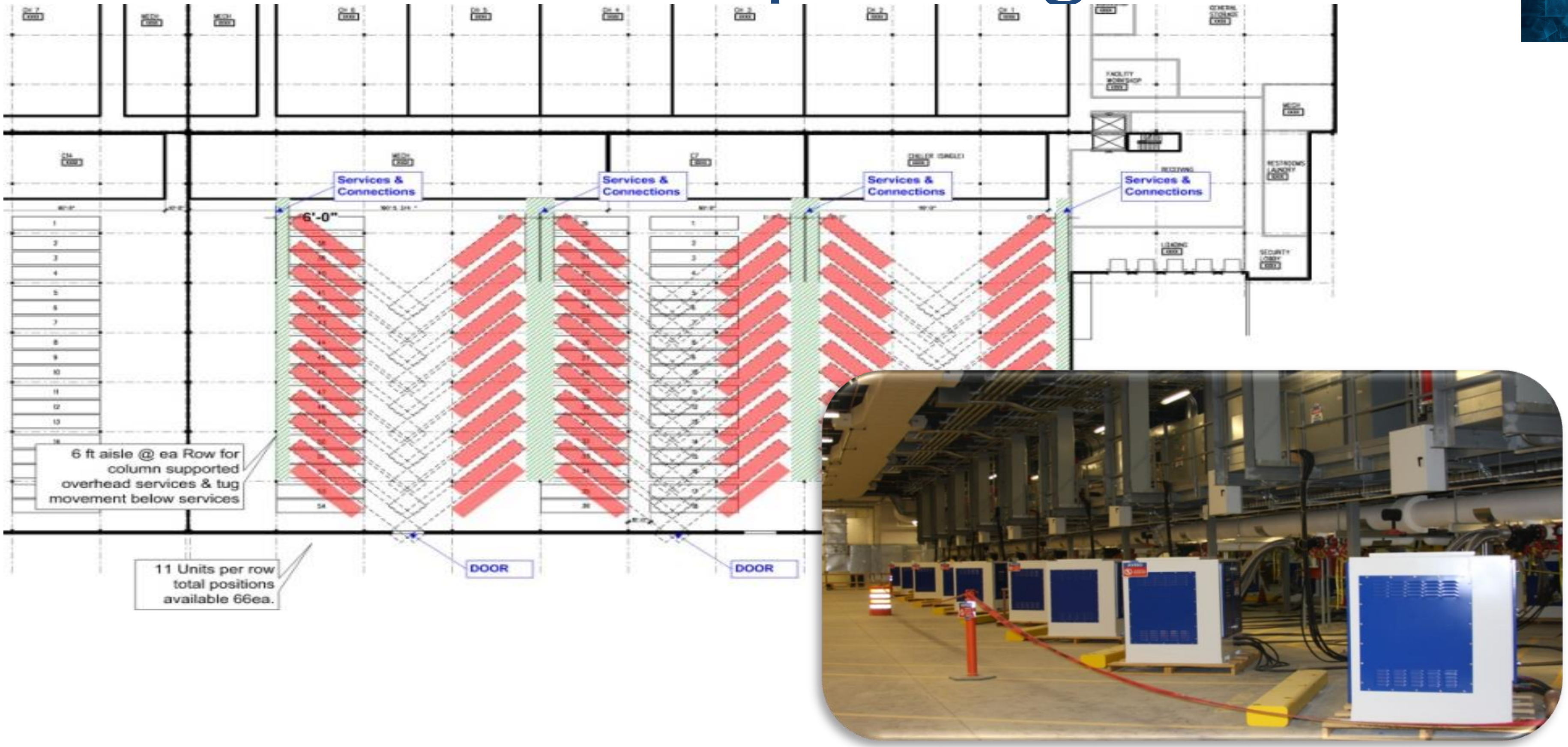
Conquering complexity.

- Building racks of servers and complex cooling systems all separately is not efficient.
- Package and deploy into bigger units:

Generation 4 data center video

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Containers: Separating Concers



6 ft aisle @ ea Row for column supported overhead services & tug movement below services

11 Units per row total positions available 66ea.

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Why is This not Good Ol' Supercomputing

- A Supercomputer is designed to scale a single application for a single user.
  - Optimized for peak performance of hardware.
  - Batch operation is not "on-demand".
  - Reliability is secondary
    - If MPI fails, app crashes. Build checkpointing into app.
  - Most data center apps run continuously (as services)
- Yet, "in many ways, supercomputers and data centers are like twins separated at birth."*

* Dan Reed

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Making Clouds Greener

- EPA released a report saying:
  - In 2006 data centers used 61 *Tera*watt-hours of power
  - Total power bill: $4.5 billion
  - 7 GW peak load (15 power plants)
  - 44.4 million mt CO_2 (0.8% emissions)
  - This was 1.5 % of all US electrical energy use
  - Expected to double by 2011
- A new challenge and a green initiative
- A deeper look and a few ideas

Data Center Design Issues

Where are the costs?

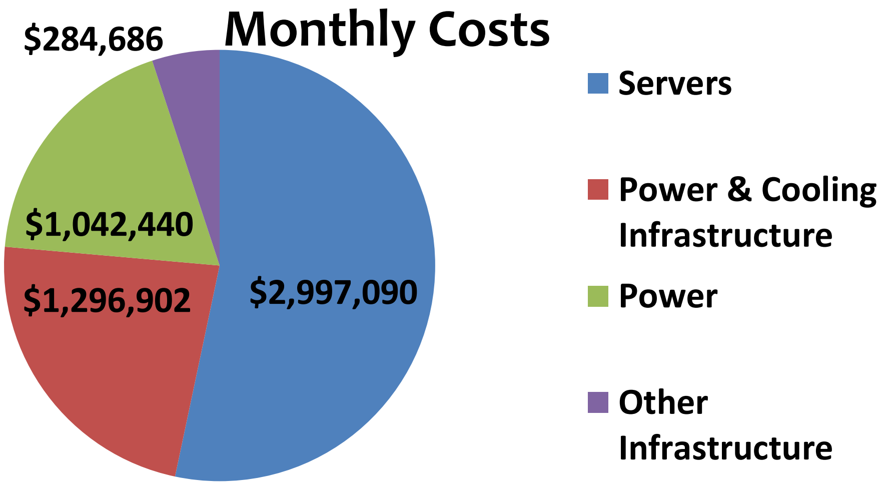
Mid-sized facility (20 containers)

- Cost of power ($/kwh): $0.07
- Cost of facility: $200,000,000 (amortize 15 years)
- Number of Servers: 50,000 (3 year life) @$2K each
- Power critical load 15MW
- Power Usage Effectiveness (PUE) 1.7

Observe:

- Fully burdened cost of power = power consumed + cost of cooling and power distribution infrastructure

As cost of servers drops and power costs rise, power will dominate all other costs.



3yr server and 15 yr infrastructure amortization

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

What can we Do About Power Costs?

Data Centers use 1.5% of US electricity

- $4.5 billion annually
- 7 GW peak load (15 power plants)
- 44.4 million mt CO_2 (0.8% emissions)

Rethink Environmentals

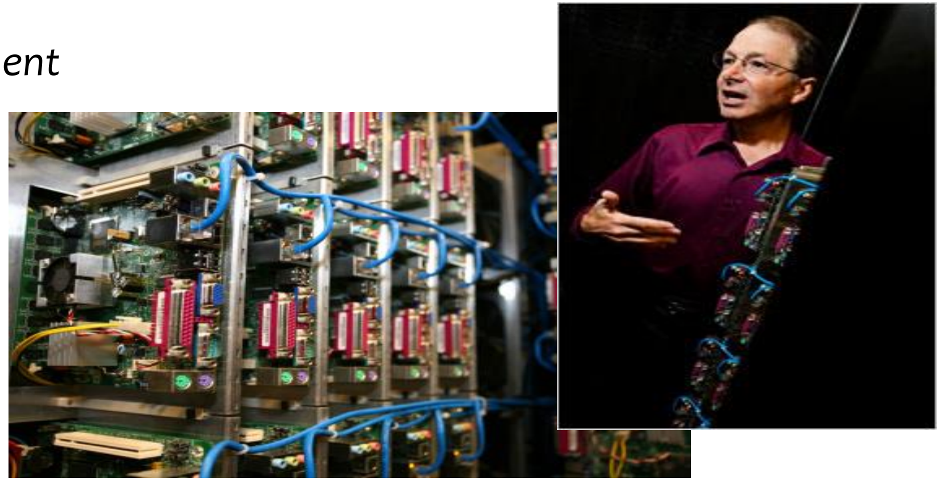
- Run them in a wider rage of conditions
  - *Christian Belady's "In Tent" data center experiment*

Rethink UPS

- Google's battery per server

Rethink Architecture

- Intel Atom and power states
- Marlowe Project



Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Marlowe and the Big Sleep

Adaptive Resource Management

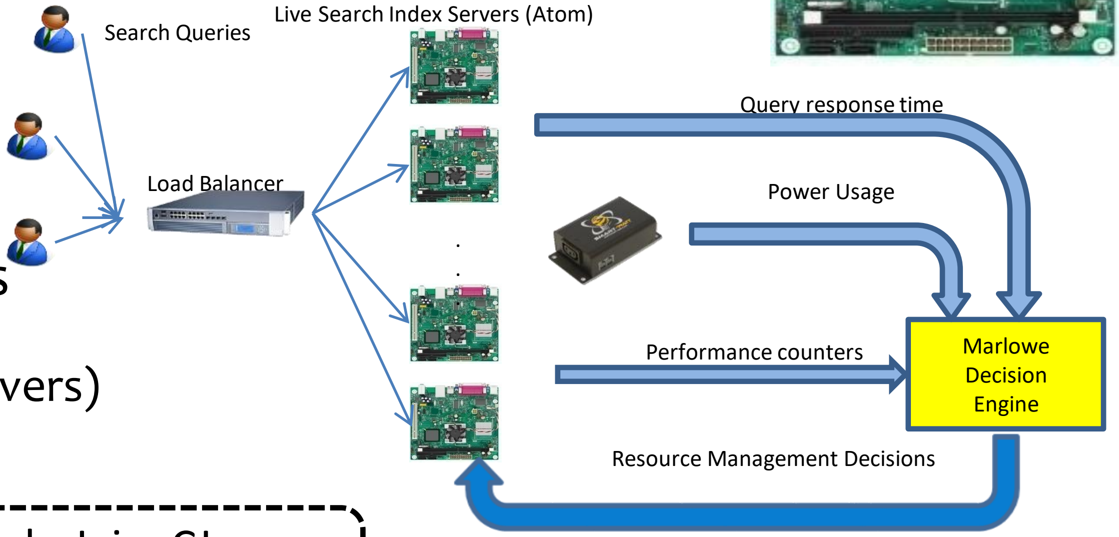
- Monitor the data center and its apps
- Use rules engine and fuzzy logic to control resources

~90% idle CPU for most current workloads

- spare capacity needed for peaks and growth

Reduce energy to idle machines

- sleep/hibernate 3 – 4 watts (vs. 28 – 36 watts for Atom servers)
- 5 – 45 sec. to reactivate server



Created by Navendu Jain, CJ Williams, Dan Reed and Jim Larus

Programming the Cloud

- Cloud Apps connect ***people*** to
  - Insight from Information
  - Experience
  - Discovery
- Most Cloud Apps are immediate, scalable and persistent
- The Cloud is also a platform for massive data analysis
  - Not a replacement for leading edge supercomputers
- The Programming model must support scalability in two dimensions
  - Thousands of simultaneous users of the same app
  - Apps that require thousands of cores for each use

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

The Challenge of Large-Scale Data Analysis

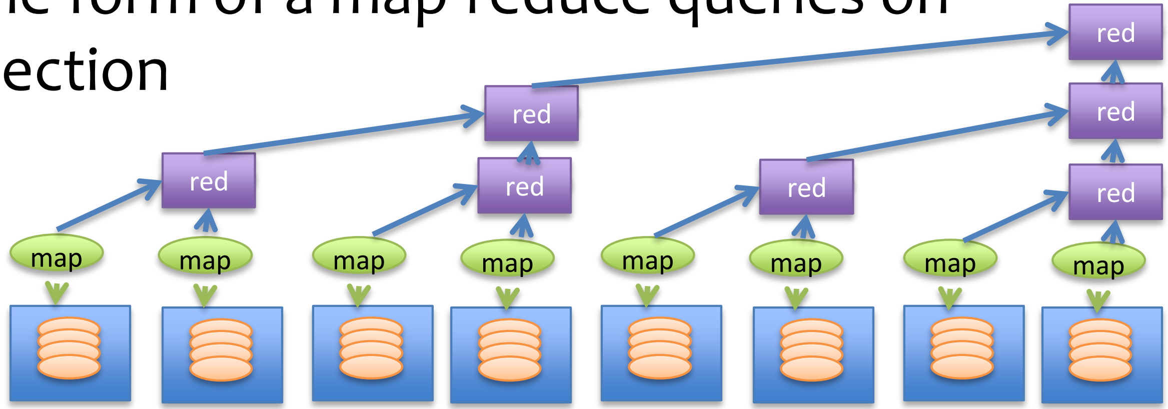
Consider web search and index creation

- Over 100.1 million websites operated as of March 2008*
- As of March 2009, the indexable web contains at least 25.21 billion pages*
  Assume each page is 100KB. Total size of web is 2.5PB
- It would take 2,500 servers with 1 TB disk to store the web
- Disks fail constantly so assume 10K servers to store web with 4-way redundancy

Data analysis often takes the form of a map-reduce queries on this distributed data collection

Hadoop does MapReduce

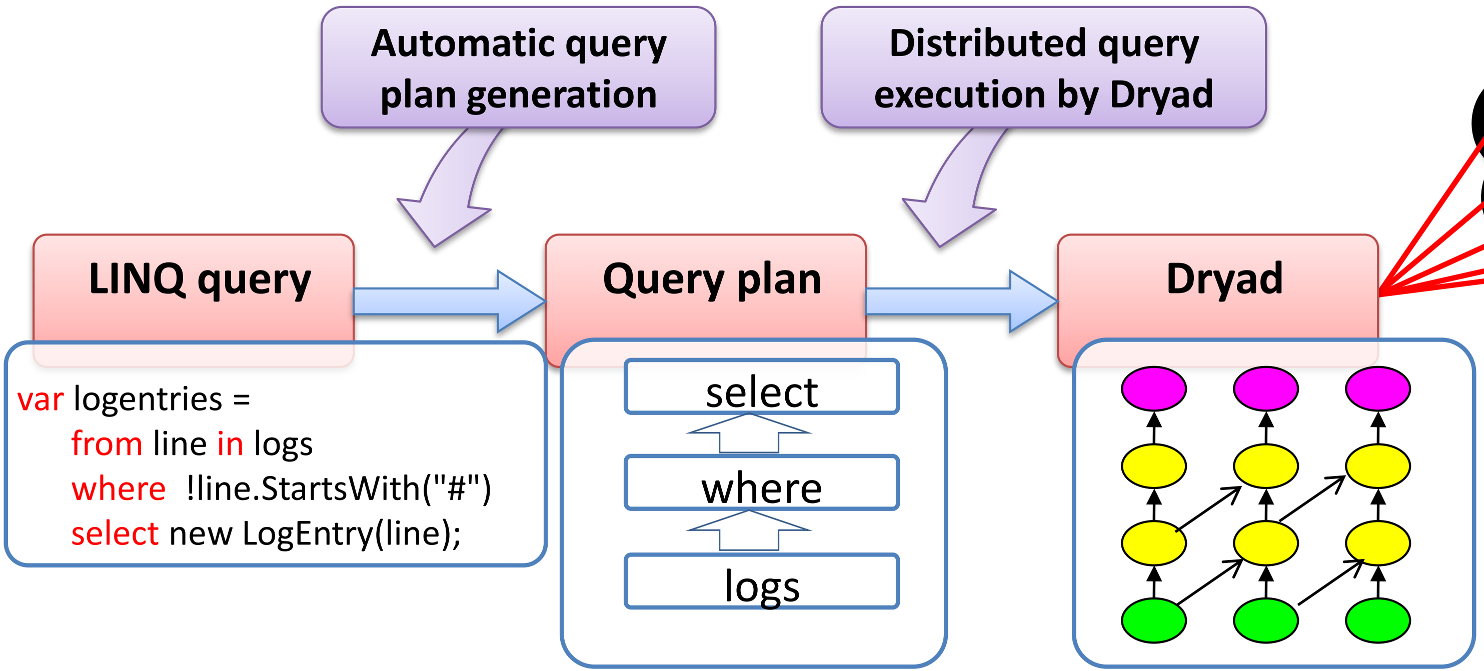
Dryad does **much more**



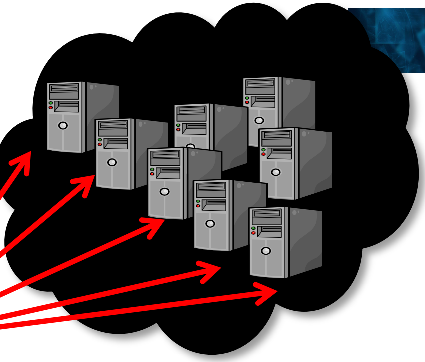
*according to wikipedia, so who knows if it is true?

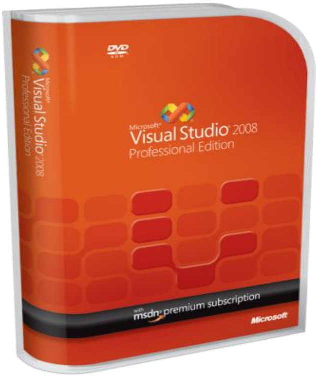
Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Generalizing MapReduce: DryadLINQ


Automatic query plan generation
Distributed query execution by Dryad
LINQ query
Query plan
Dryad
select
where
logs

```
var logentries =
    from line in logs
    where  !line.StartsWith("#")
    select new LogEntry(line);
```





LINQ: .NET Language Integrated Query

- Declarative SQL-like programming with C# and Visual Studio
- Easy expression of data parallelism
- Elegant and unified data model

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Cloud Models

Infrastructure as a Service

- Provide a way to host virtual machines on demand
  - Amazon ec2 and S3 – you configure your VM, load and go

Platform as a Service

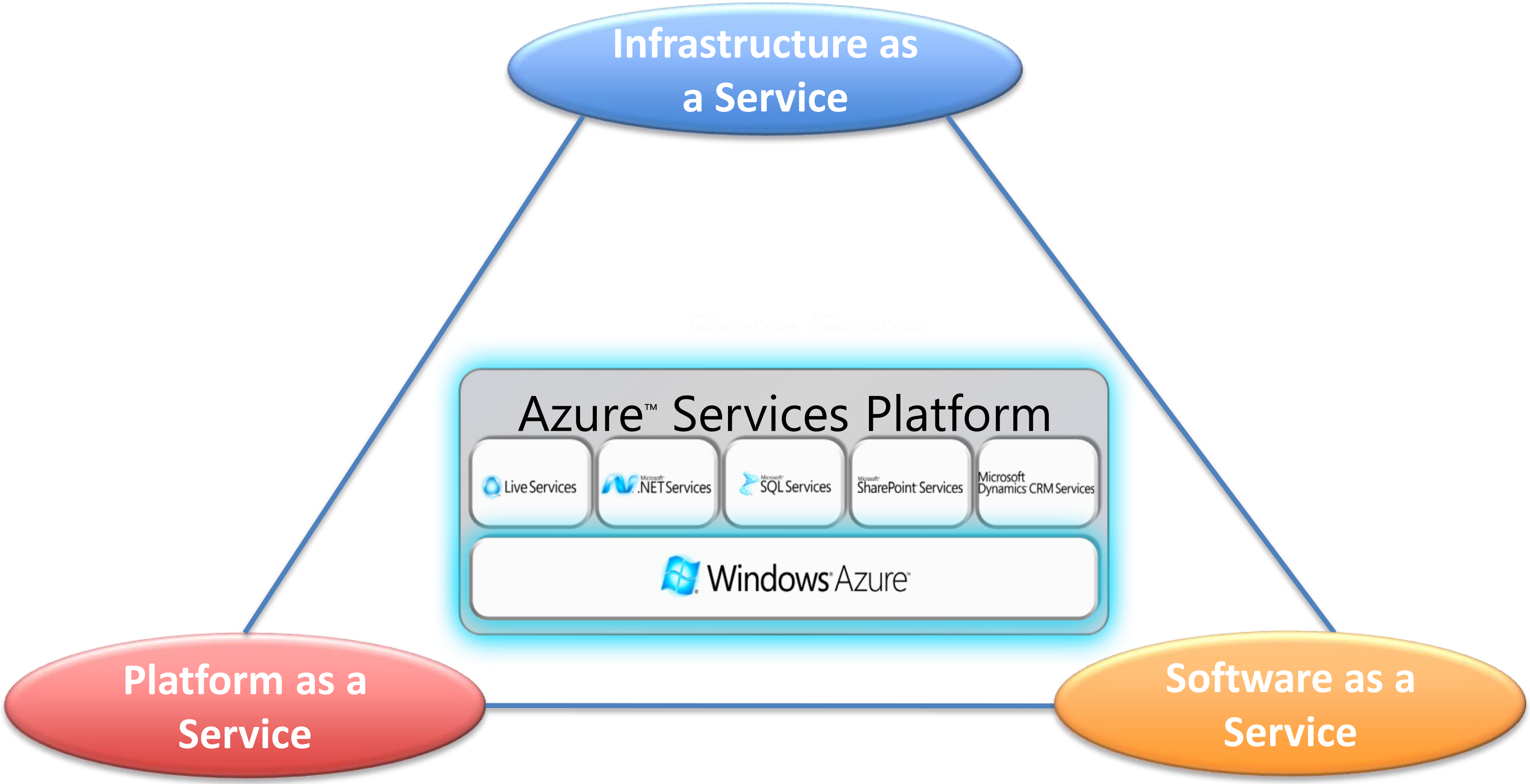
- You write an App to cloud APIs and release it. The platform manages and scales it for you.
- Google App engine:
  - Write a python program to access Big Table. Upload it and run it in a python cloud.
  - Hadoop and Dryad are application frameworks for data parallel analysis

Software as a Service

- Delivery of software to the desktop from the cloud
  - Stand-alone applications (Word, Excel, etc)
  - Cloud hosted capability
    - doc lives in the cloud
    - Collaborative document creation

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Cloud Application Frameworks



Infrastructure as a Service

Azure™ Services Platform

Live Services

Microsoft .NET Services

Microsoft SQL Services

Microsoft SharePoint Services

Microsoft Dynamics CRM Services

Windows Azure

Platform as a Service

Software as a Service

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

What is Cloud Computing?

Cloud Computing means using a remote data center to manage scalable, reliable, on-demand access to applications and data

- *Scalable means*
  - Possibly millions of simultaneous users of the application
  - Exploiting massive parallelism in the applications
- *Reliable, elastic and on-demand*
  - 5 “nines” available right now

Classic cloud applications

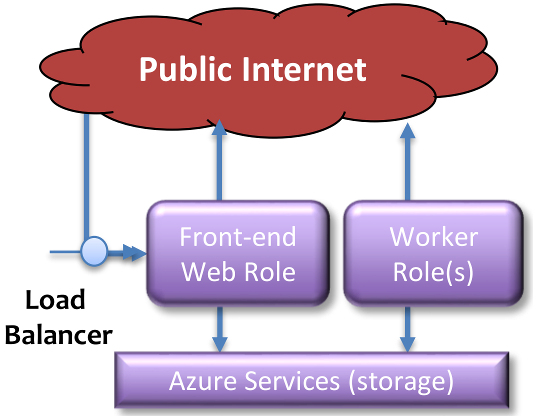
- *Web search, social media, “cloud” storage, e-mail*

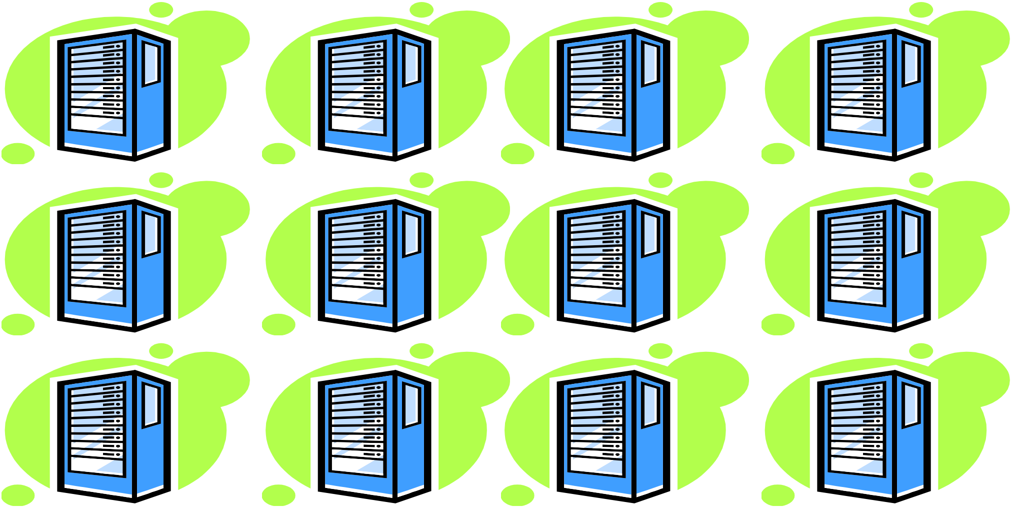
Second generation cloud applications include

- *Large scale data analysis and scientific collaboration*

2009

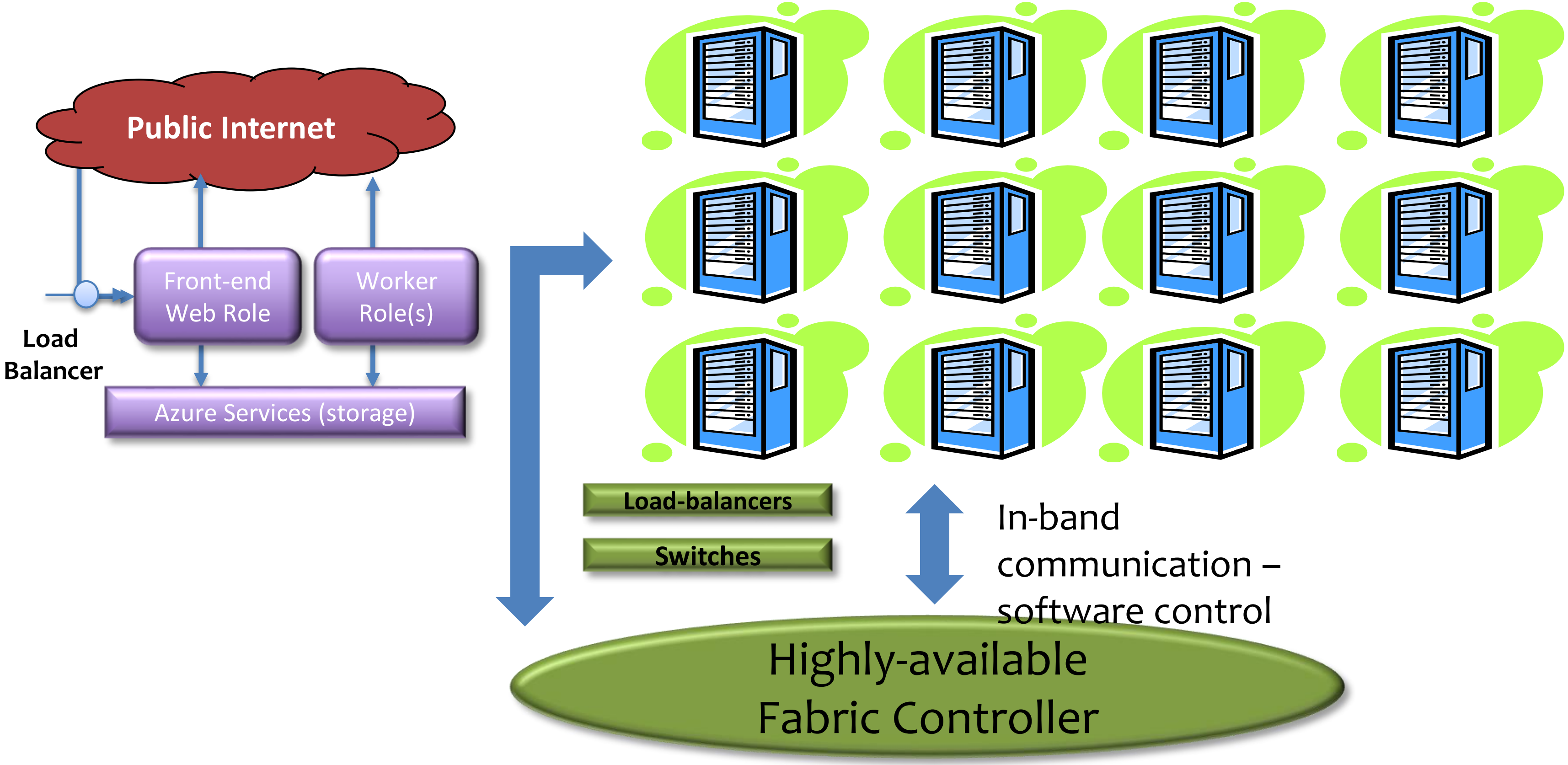
Azure Virtualization Architecture





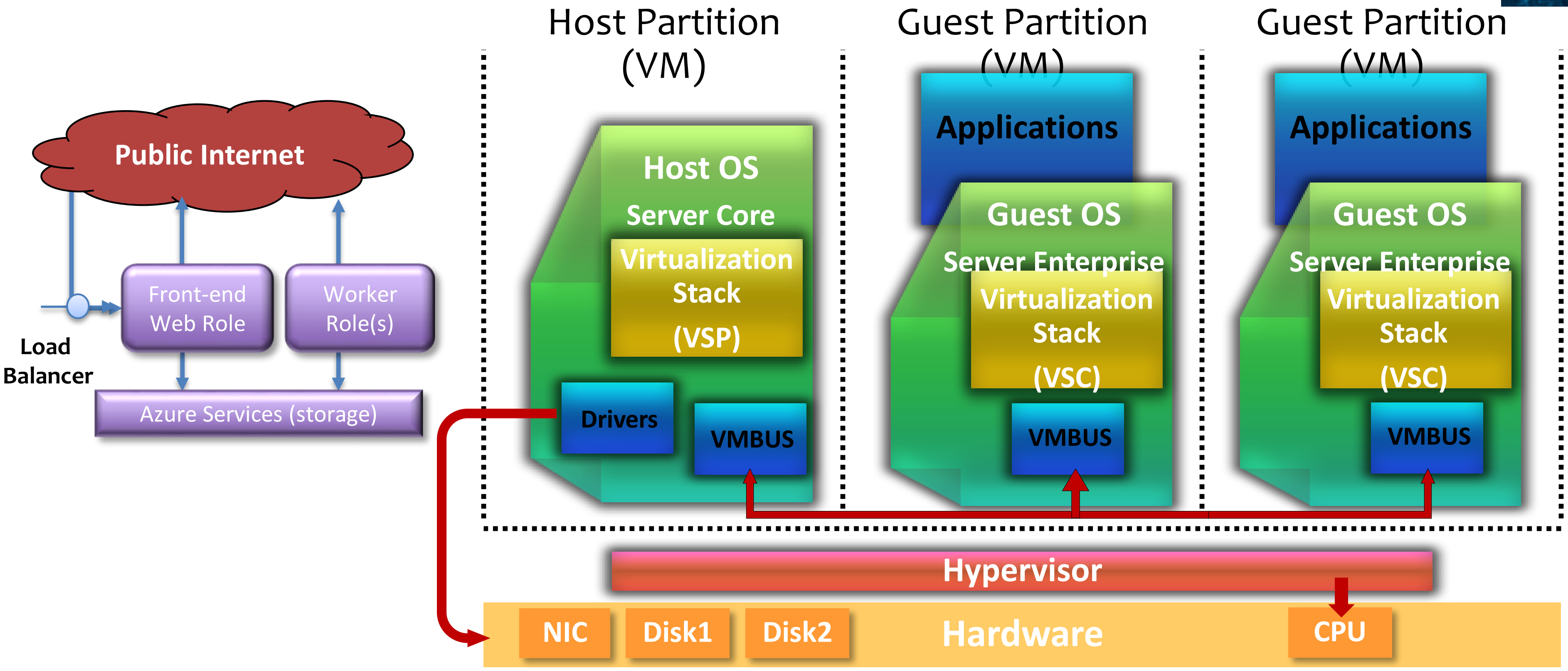
Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Azure Virtualization Architecture



Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Azure Virtualization Architecture

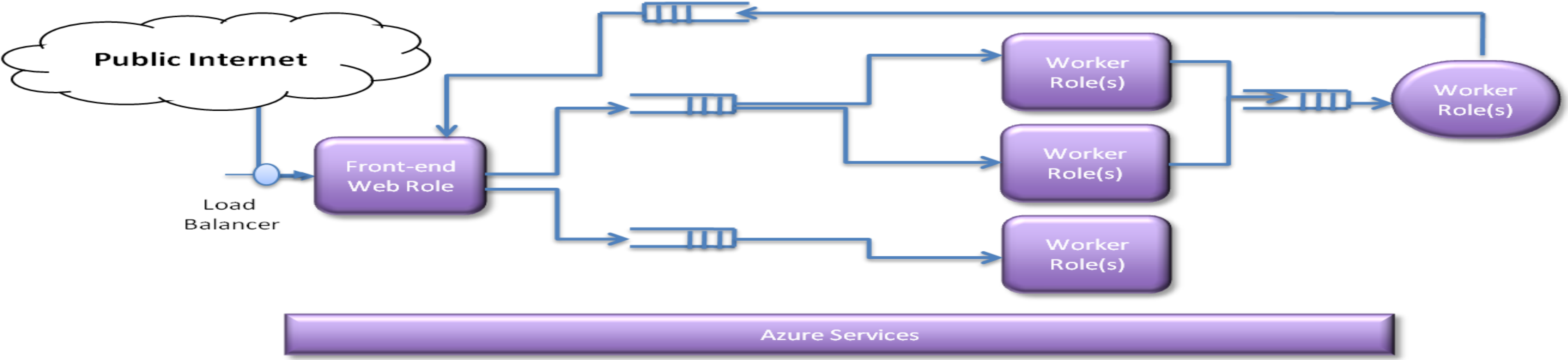


Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

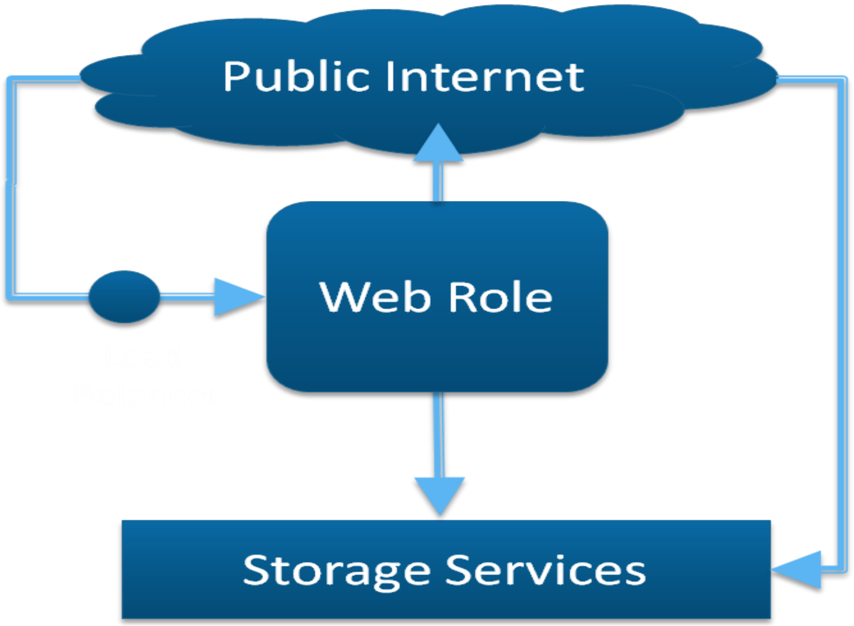
The Architecture of an Azure Application

Roles are a mostly stateless process running on a core.

- *Web Roles provide web service access to the app by the users. Web roles generate tasks for worker roles*
- *Worker Roles do "heavy lifting" and manage data in tables/blobs*
- *Communication is through queues*
- *The number of role instances should dynamically scale with load*



Web Role



Web farm that handles request from the internet

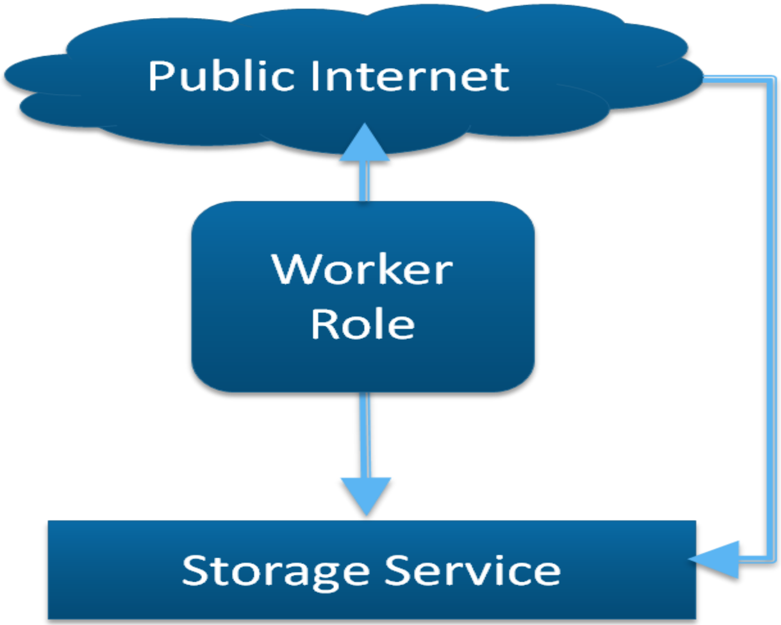
IIS7 hosted web core

- *Hosts ASP.NET*
- *XML based configuration of IIS7*
- *Integrated managed pipeline*
- *Supports SSL*
- *Windows Azure code access security policy for managed code*

Worker Role

- No inbound network connections
- Can read requests from queue in storage

Windows Azure specific policy for managed code



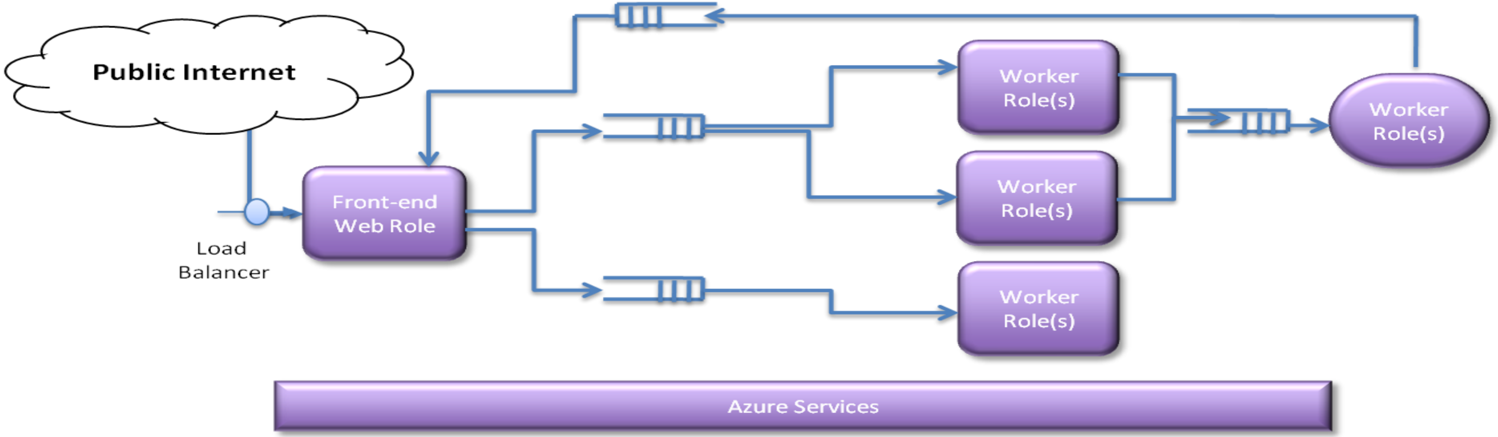
Service Runtime API

- Every role has access to APIs for common functionality needed for services
  - *Read configuration setting values*
  - *Write messages to set of standard logging streams*
    - "Printf" sitting on top of a lot of plumbing so logs are downloadable and archived easily
    - Critical messages generate live alerts
  - *Get access to unreliable local storage for caching*
- Defines interface for worker role

Azure Queues for Scalability and Availability

Scalability

- Queue length directly reflects how well backend processing is keeping up with overall workload
- Queues decouple different parts of the application, making it easier to scale parts of the application independently
- Flexible resource allocation, different priority queues and separation of backend servers to process different queues



Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Azure Queues for Scalability and Availability

Scalability

- Queue length directly reflects how well backend processing is keeping up with overall workload
- Queues decouple different parts of the application, making it easier to scale parts of the application independently
- Flexible resource allocation, different priority queues and separation of backend servers to process different queues

Decouple Front-End Roles from Back-End Roles

Handles Traffic Bursts

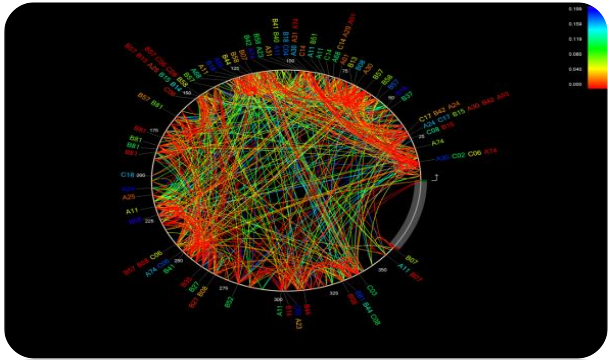
- *Queues provide a buffer to absorb traffic bursts and reduce the impact of individual component failures*

Reliable message delivery, exactly once execution

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Science Example/Demo *PhyloD as an Azure Service*

- Statistical tool used to analyze DNA of HIV from large studies of infected patients
- PhyloD was developed by Microsoft Research and has been highly impactful
- Small but important group of researchers
  - *100's of HIV and HepC researchers actively use it*
  - *1000's of research communities rely on results*



Cover of PLoS Biology November 2008

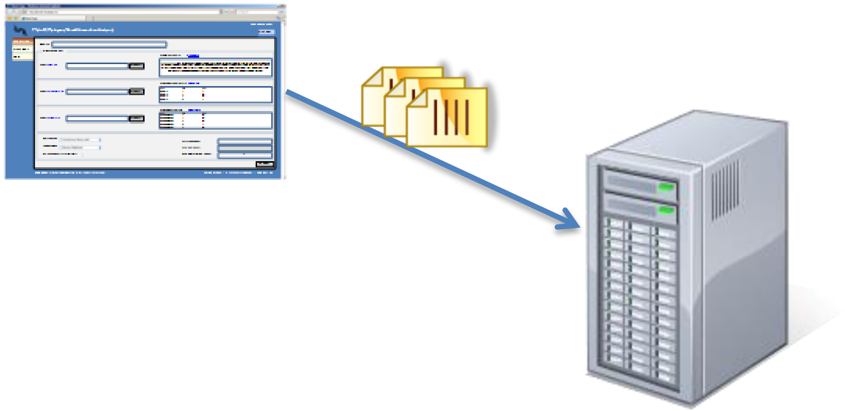
Typical job, 10 – 20 CPU hours, extreme jobs require 1K – 2K CPU hours

- Very CPU efficient
- Requires a large number of test runs for a given job (1 – 10M tests)
- Highly compressed data per job (~100 KB per job)

Demo Presenter
Wei Lu, Cloud Computing Futures

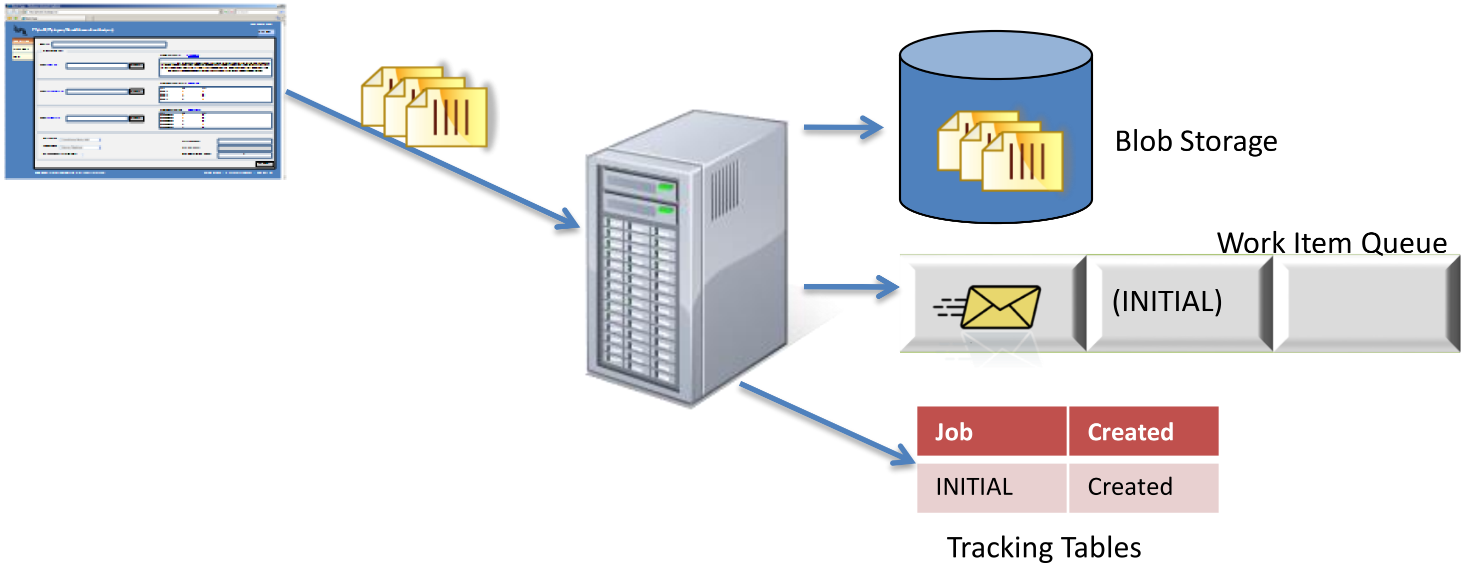
PhyloD as an Azure Service

- The web role provides an interface to the clients. Worker roles perform actual computation. Web role and worker roles share information using blobs, queue and tables.



PhyloD as an Azure Service

- Web role copies input tree, predictor and target files to blob storage, enqueues INITIAL work item and updates tracking tables.

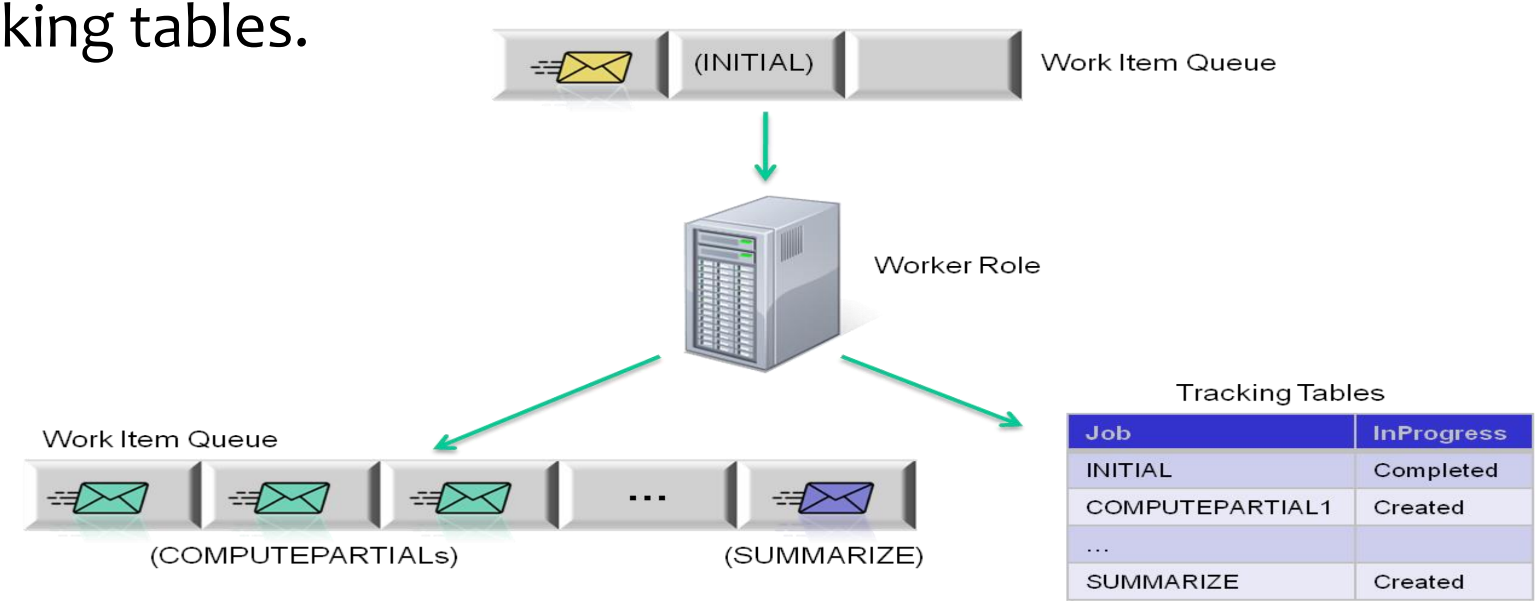


Job	Created
INITIAL	Created

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

PhyloD as an Azure Service

- Worker role enqueues a COMPUTEPARTIAL work item for each partition of the input problem followed by a SUMMARIZE work item to aggregate the partial results and finally updates the tracking tables.

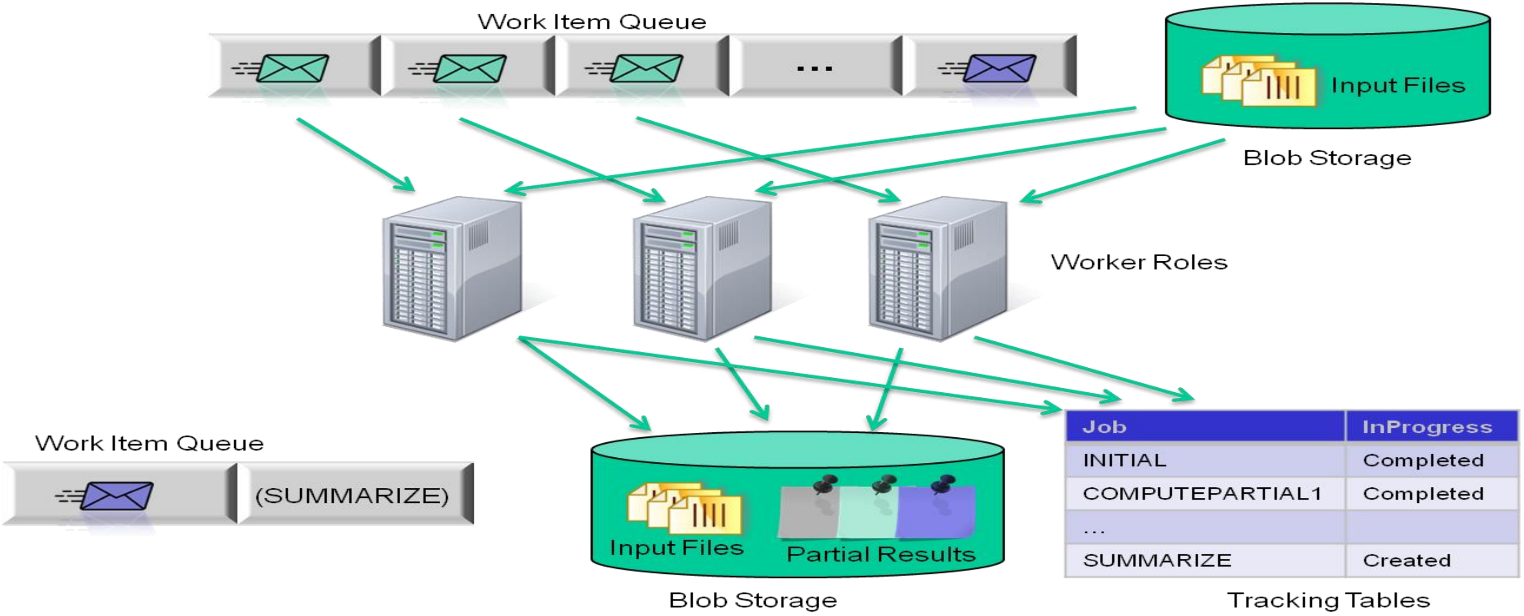


Job	InProgress
INITIAL	Completed
COMPUTEPARTIAL1	Created
…	
SUMMARIZE	Created

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

PhyloD as an Azure Service

- Worker role copies the input files to its local storage, computes p-values for a subset of the allele-codon pairs, copies the partial results back to blob storage and updates the tracking tables.

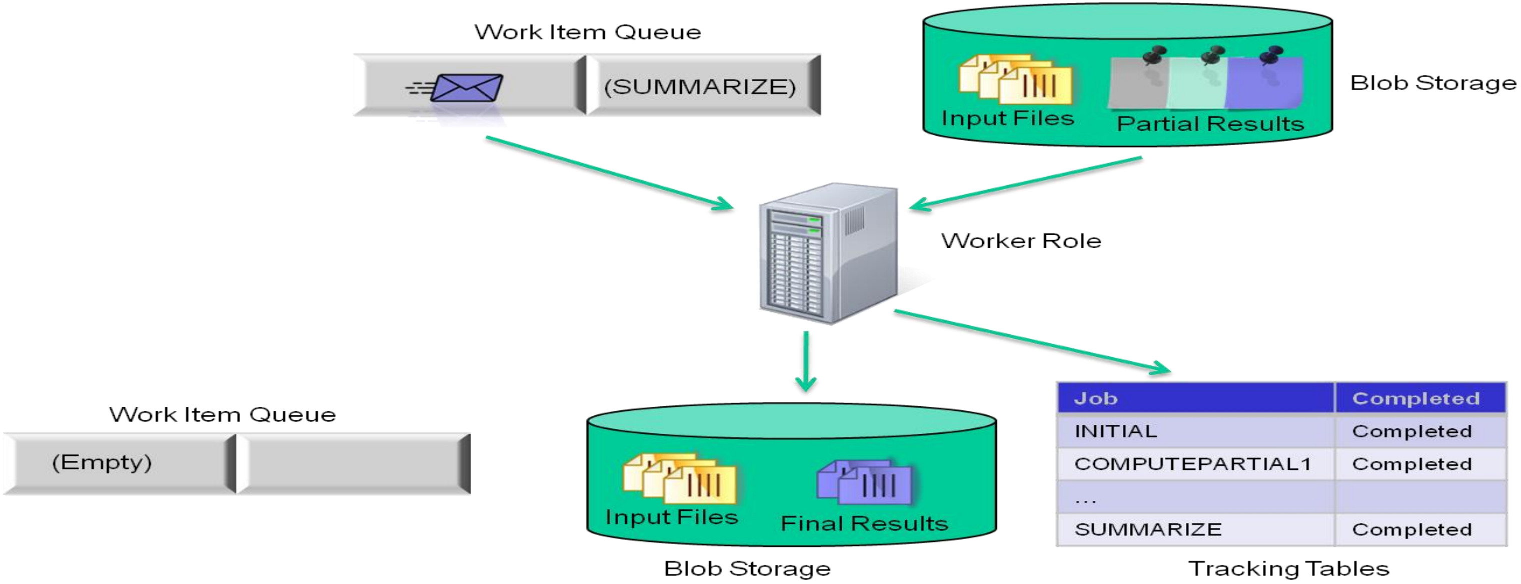


Job	InProgress
INITIAL	Completed
COMPUTEPARTIAL1	Completed
...	
SUMMARIZE	Created

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

PhyloD as an Azure Service

- Worker role copies input files and COMPUTEPARTIAL outputs to local storage, computes q-values for each allele-codon pair, copies results to the blob storage and updates tracking tables.

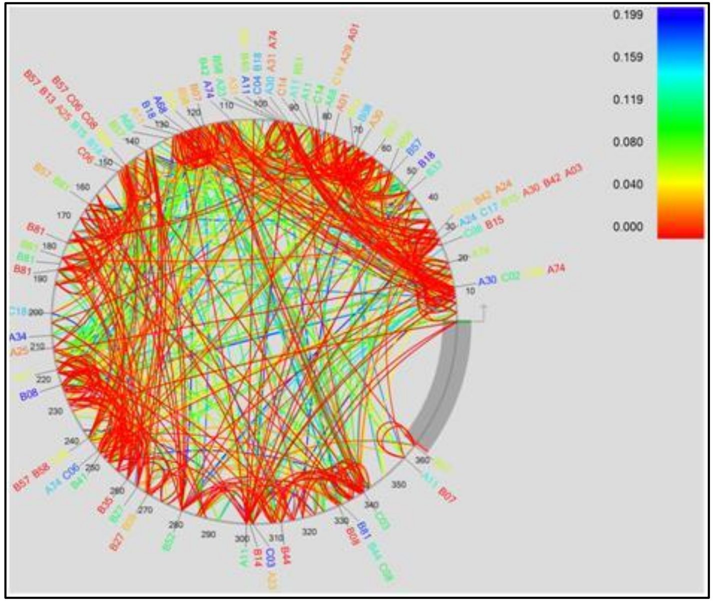


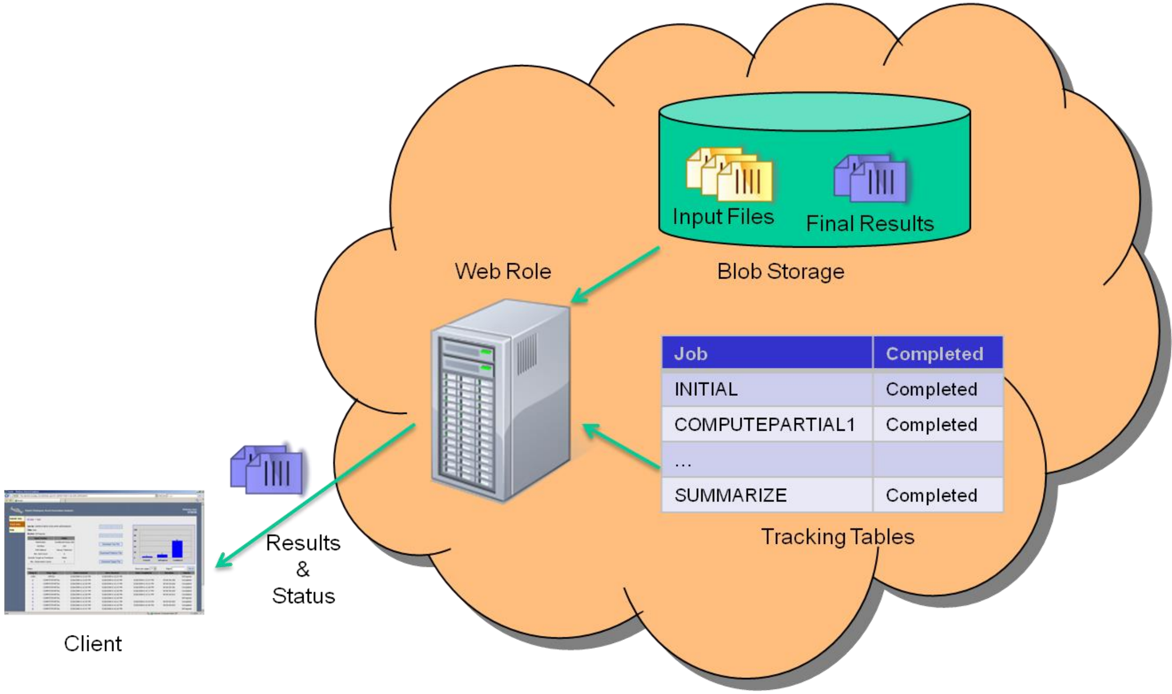
Job	Completed
INITIAL	Completed
COMPUTEPARTIAL1	Completed
…	
SUMMARIZE	Completed

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

PhyloD as an Azure Service

- Web role serves the final results from blob storage and status reports from tracking tables.

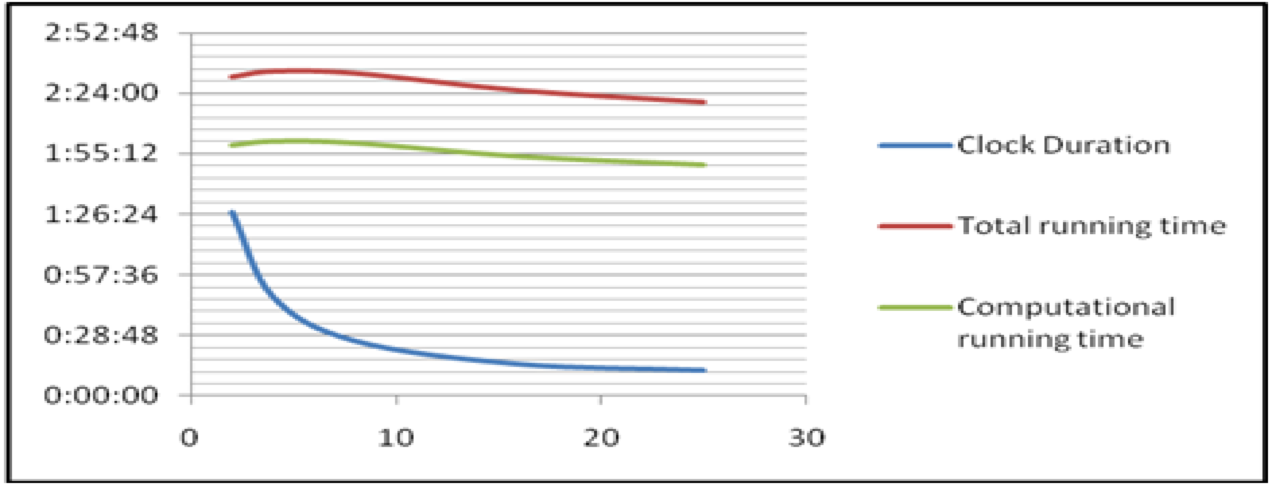




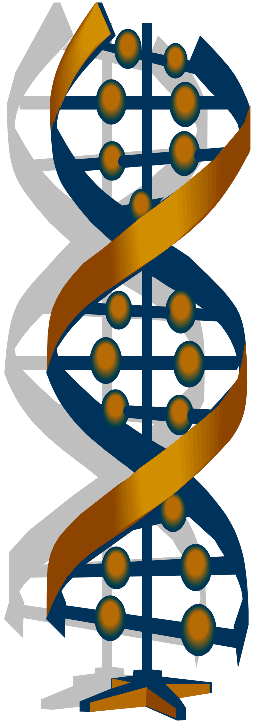
Job	Completed
INITIAL	Completed
COMPUTEPARTIAL1	Completed
…	
SUMMARIZE	Completed

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

PhyloD as an Azure Service



Workers	Clock Duration	Total running time	Computational running time
25	0:12:00	2:19:39	1:49:43
16	0:15:00	2:25:12	1:53:47
8	0:26:00	2:33:23	2:00:14
4	0:47:00	2:34:17	2:01:06
2	1:27:00	2:31:39	1:59:13



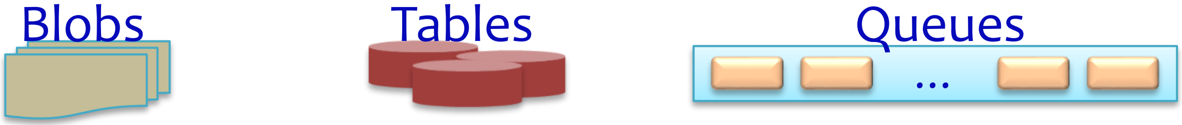
Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Azure Data Storage

Two levels

1. Basic Azure storage

Three abstractions:



Blobs

Tables

Queues

- Three replicas of everything
- Shared key authentication
- REST API

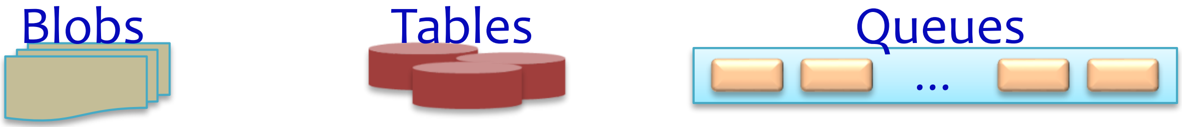
Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Azure Data Storage

Two levels

1. Basic Azure storage

Three abstractions:



Blobs

- Simple interface for storing named files along with metadata for the file
- Up to 50GB each
- 8KB metadata each
- Stored in containers
- Public or private access at container level
- Standard REST Interface: **PutBlob, GetBlob, DeleteBlob**

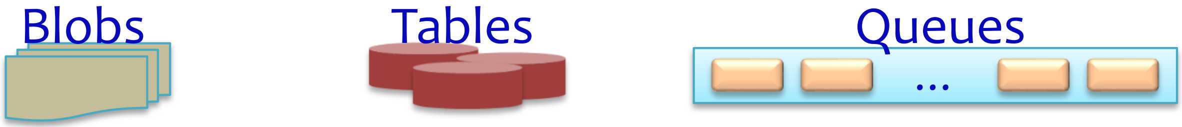
Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Azure Data Storage

Two levels

1. Basic Azure storage

Three abstractions:



Blobs Tables Queues

…

Tables

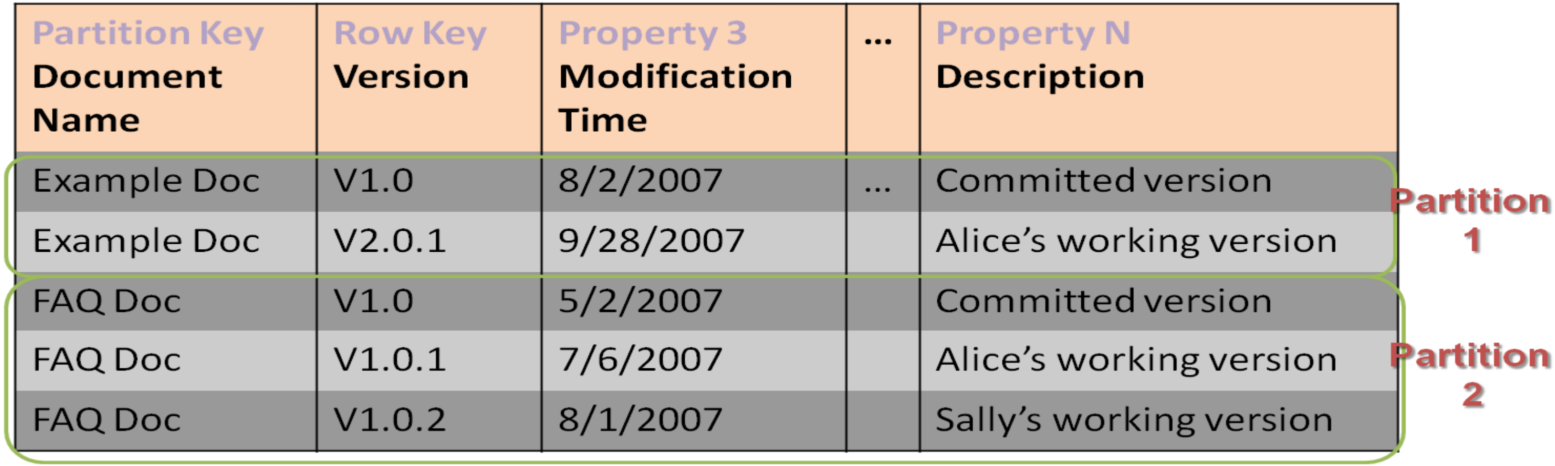
- provide structured storage. A table is a set of entities, which contain a set of properties;
- Non-relational;
- Partitioned for scale;
- No fixed schema;
- ADO.NET Data Services.

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Partition Key and Partition

Every table has a partition key

- It is the first property (column) of your table
- All entities in table with same partition key value live in the same partition (locality for storage and efficient retrieval)
- Azure will automatically load balance partitions



Partition Key	Row Key	Property 3	...	Property N
Document Name	**Version**	**Modification Time**		**Description**
Example Doc	V1.0	8/2/2007	...	Committed version
Example Doc	V2.0.1	9/28/2007		Alice's working version
FAQ Doc	V1.0	5/2/2007		Committed version
FAQ Doc	V1.0.1	7/6/2007		Alice's working version
FAQ Doc	V1.0.2	8/1/2007		Sally's working version

Partition 1

Partition 2

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Partition Key and Partition

Every table has a partition key

- It is the first property (column) of your table
- All entities in table with same partition key value live in the same partition (locality for storage and efficient retrieval)
- Azure will automatically load balance partitions

Need to choose partitioning scheme to make data access scalable (Details beyond scope of this talk)

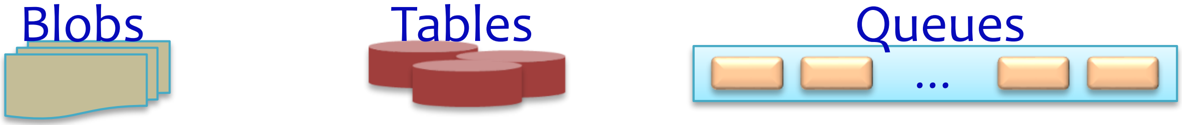
- See Brad Calder’s talk
  “Essential Cloud Storage Services”
- As well as Pablo Castro’s talk
  “Modeling Data For Efficient Access At Scale”

Azure Data Storage

Two levels

1. Basic Azure storage

Three abstractions:



Blobs

Tables

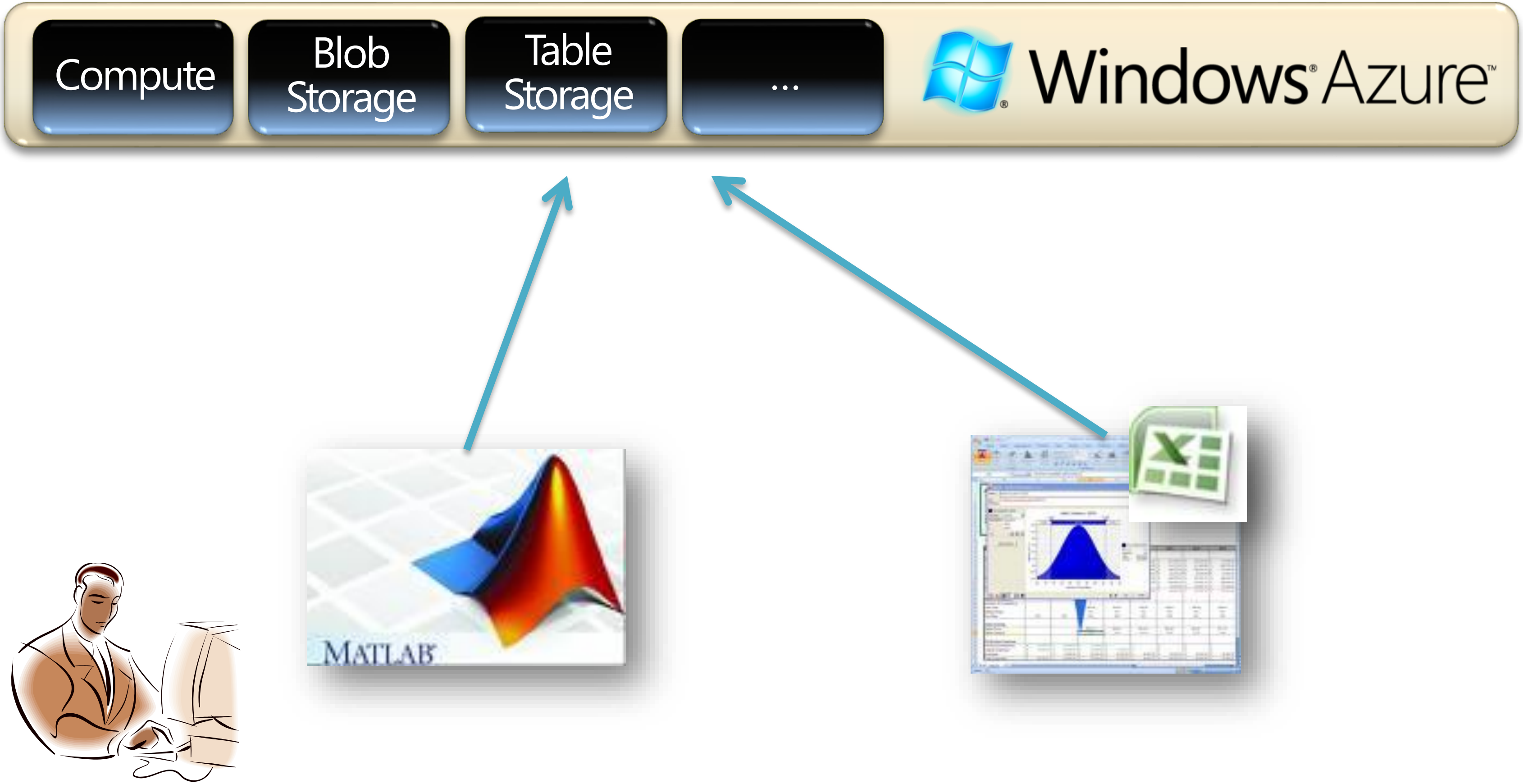
Queues

Queues

- provide reliable storage and delivery of messages for an application
- Asynchronous message passing
- 8KB messages
- Two-phase commit

Barga, Gannon: Cloud Computing Presentation, MSR Faculty Summit 2009

Demo: MATLAB on Azure



Compute

Blob Storage

Table Storage

…

Demo Presenter

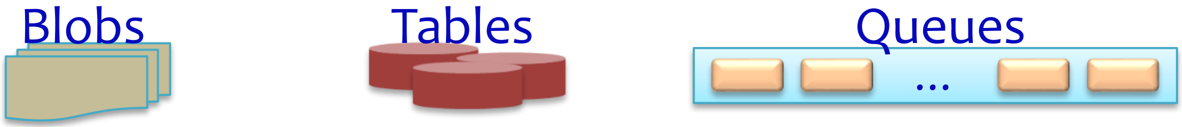
Wei Lu, Cloud Computing Futures

Azure Data Storage

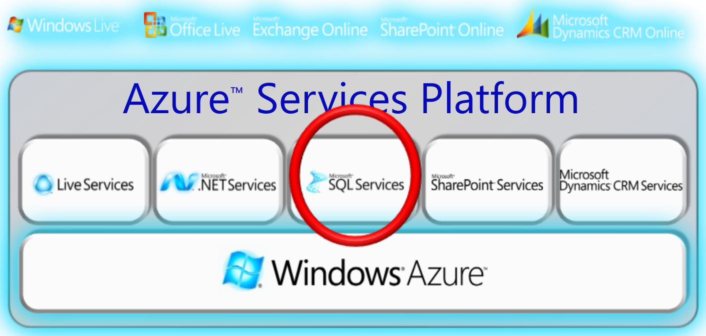
Two levels

1. Basic Azure storage

   Three abstractions:



Blobs Tables Queues



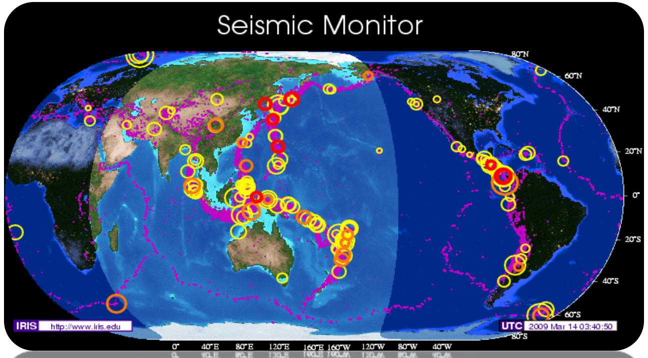
2. *SQL Azure*

   - Relational storage in the cloud

Geophysical Data in Azure

IRIS is a Seattle based consortium, sponsored by the National Science Foundation to collect and distribute global seismological data

- *Two Petabytes of seismic data collected*
- *Includes HD videos, seismograms, images, and data from major earthquakes*



High research value worldwide, frequently used in academia

Consume in any language, any tool, any platform in S+S scenario

Today Is An Inflection Point

- Economic challenges
  - Research efficiency
  - Infrastructure scaling
- Technology transition
  - Cloud software + services
  - Multicore architectures, scalable storage and computation
- Opportunity to play an active role in defining the future of both education and research
- Microsoft Research is engaging the community
  - Looking for opportunities to collaborate
  - Develop cloud services for research and technical computing at scale

Questions?