Cloud, HPC, or Hybrid: A Case Study Involving Satellite Image Processing

Marty Humphrey*

Zach Hill *

Catharine van Ingen**

Keith Jackson***

Youngryel Ryu****

* Department of Computer Science, University of Virginia

** Microsoft Research, Microsoft Bay Area Research Center, San Francisco, CA

*** Lawrence Berkeley National Lab, Berkeley, CA

**** Harvard/Berkeley/Seoul National University

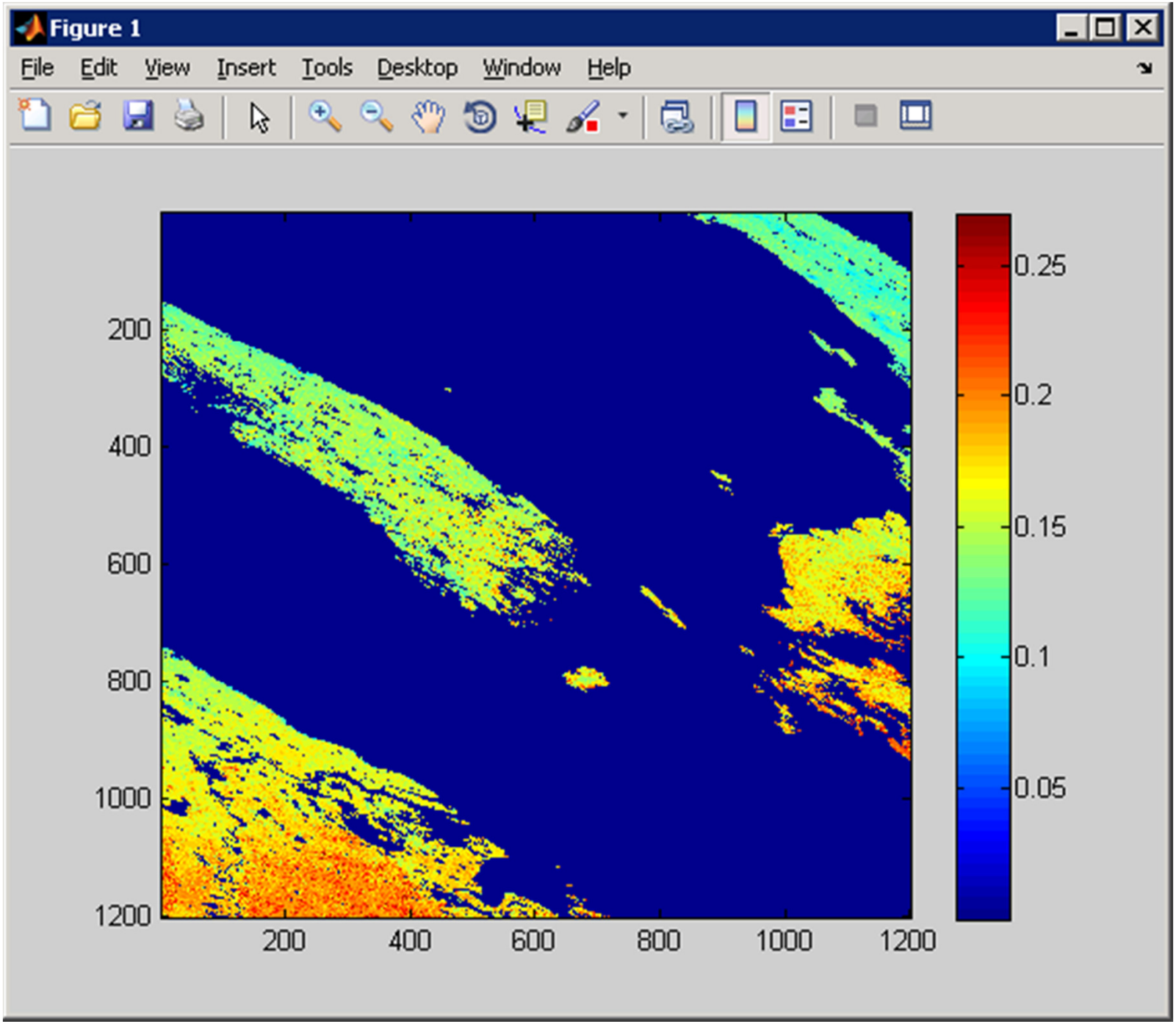
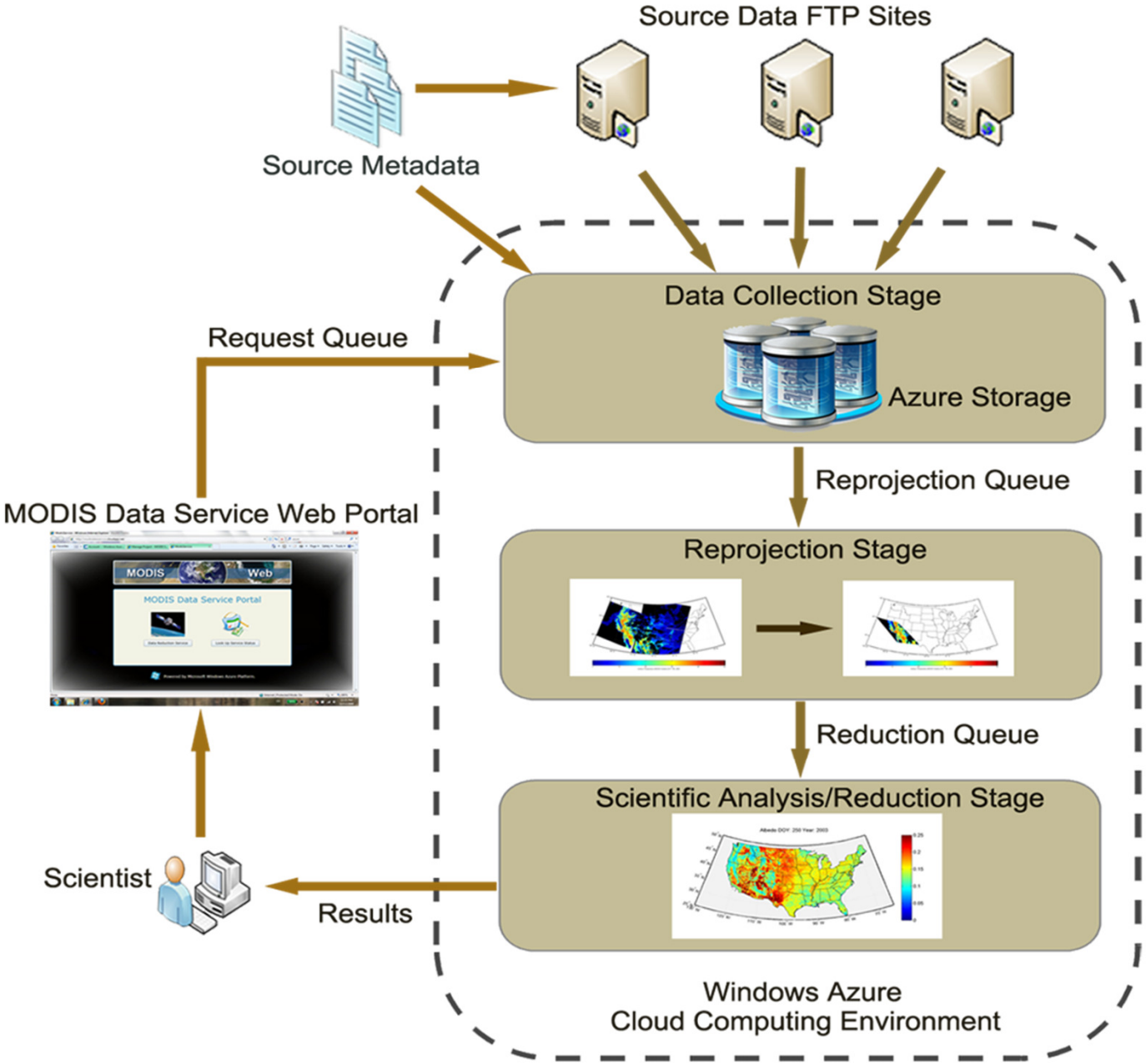
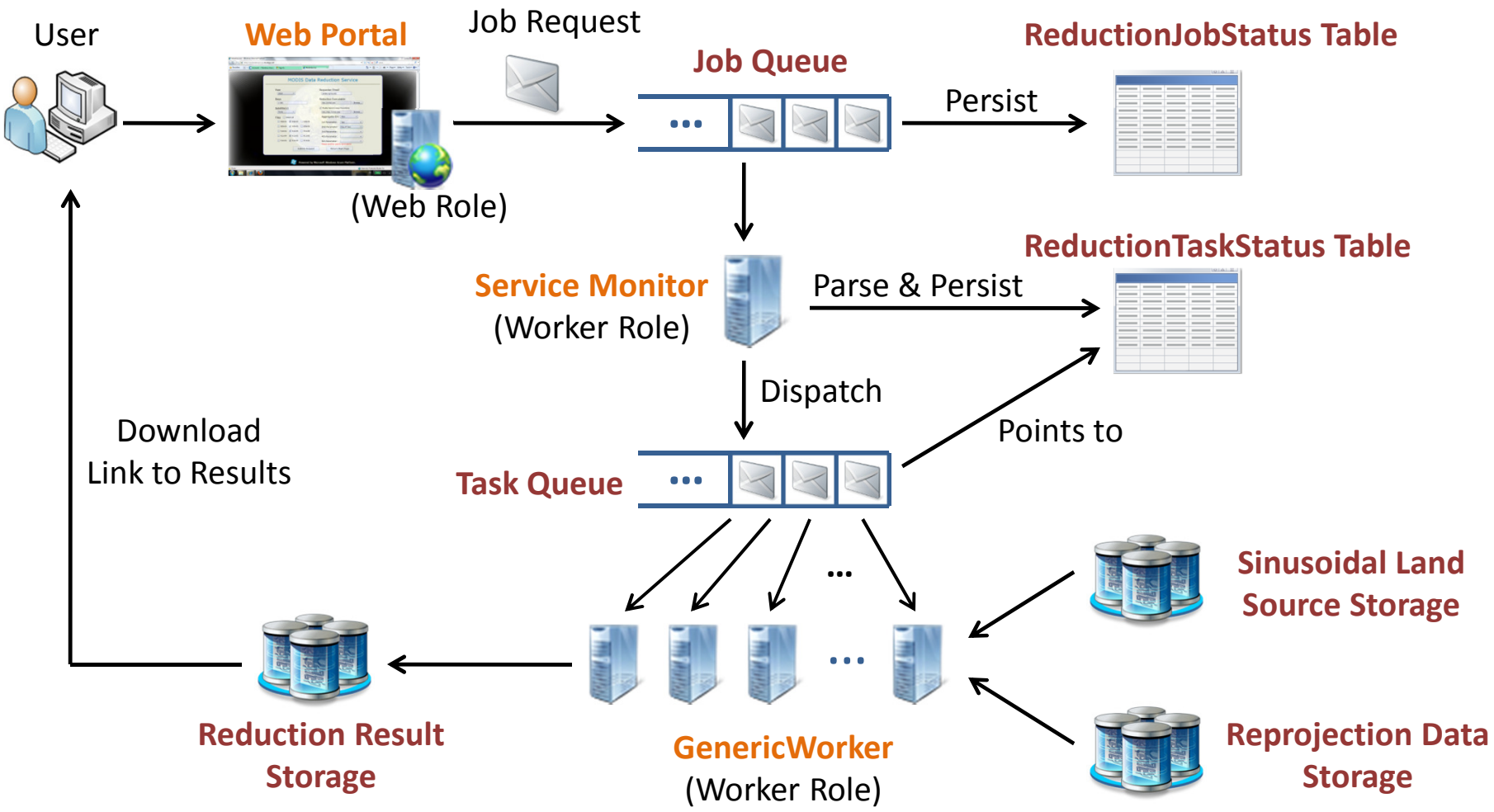


Figure 1
File Edit View Insert Tools Desktop Window Help
200
400
600
800
1000
1200
200 400 600 800 1000 1200
0.25
0.2
0.15
0.1
0.05

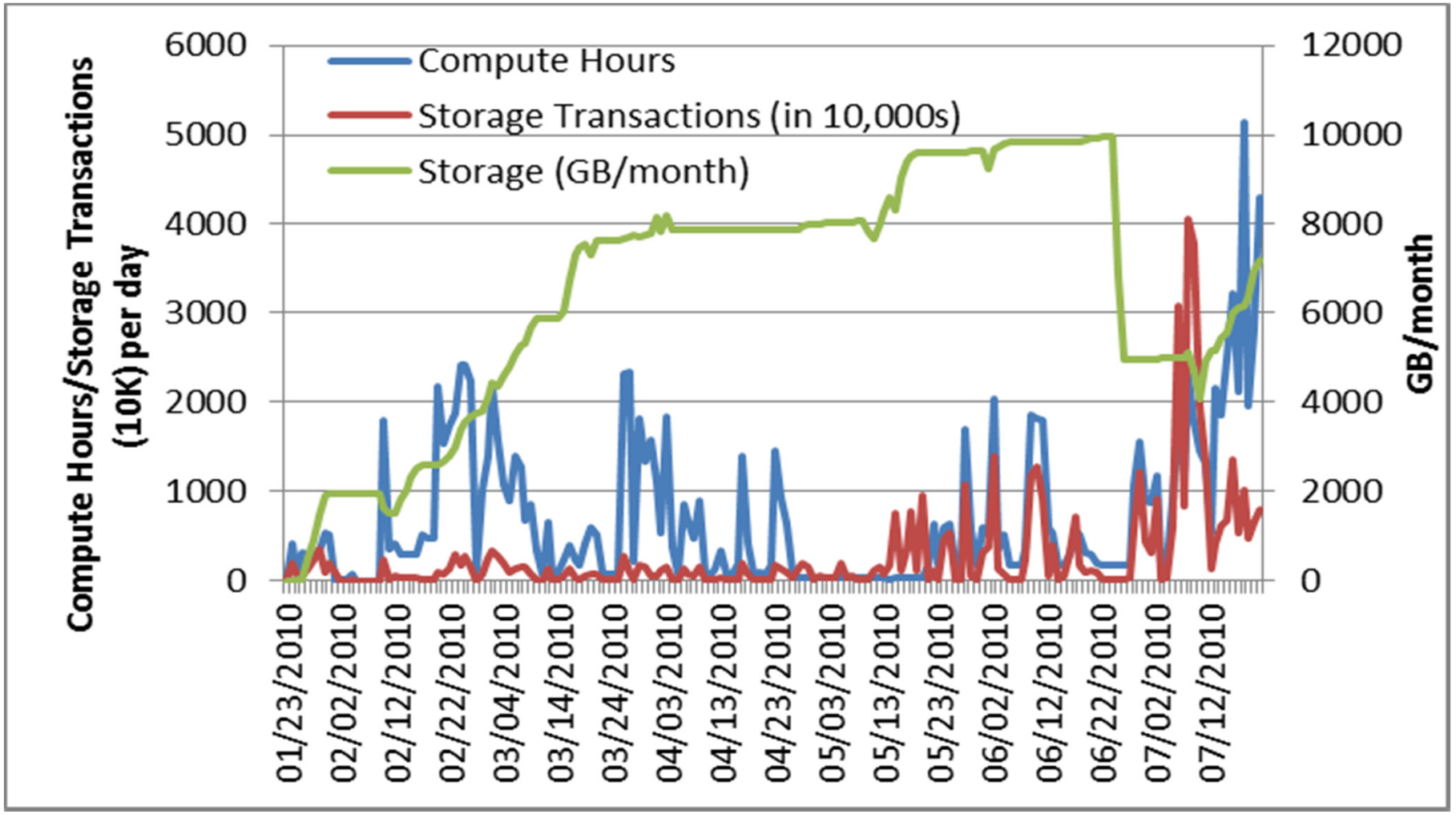


Data Processing Service Architecture

Internals


User
Web Portal
Job Request
Job Queue
ReductionJobStatus Table
Persist
(Web Role)
Service Monitor
(Worker Role)
Parse & Persist
ReductionTaskStatus Table
Dispatch
Download
Link to Results
Points to
Task Queue
Sinusoidal Land
Source Storage
Reduction Result
Storage
GenericWorker
(Worker Role)
Reprojection Data
Storage

MODISAzure


Compute Hours
Storage Transactions (in 10,000s)
Storage (GB/month)
Compute Hours/Storage Transactions (10K) per day
GB/month
6000
5000
4000
3000
2000
1000
0
12000
10000
8000
6000
4000
2000
0
01/23/2010
02/02/2010
02/12/2010
02/22/2010
03/04/2010
03/14/2010
03/24/2010
04/03/2010
04/13/2010
04/23/2010
05/03/2010
05/13/2010
05/23/2010
06/02/2010
06/12/2010
06/22/2010
07/02/2010
07/12/2010

Concerns, Limitations and Questions

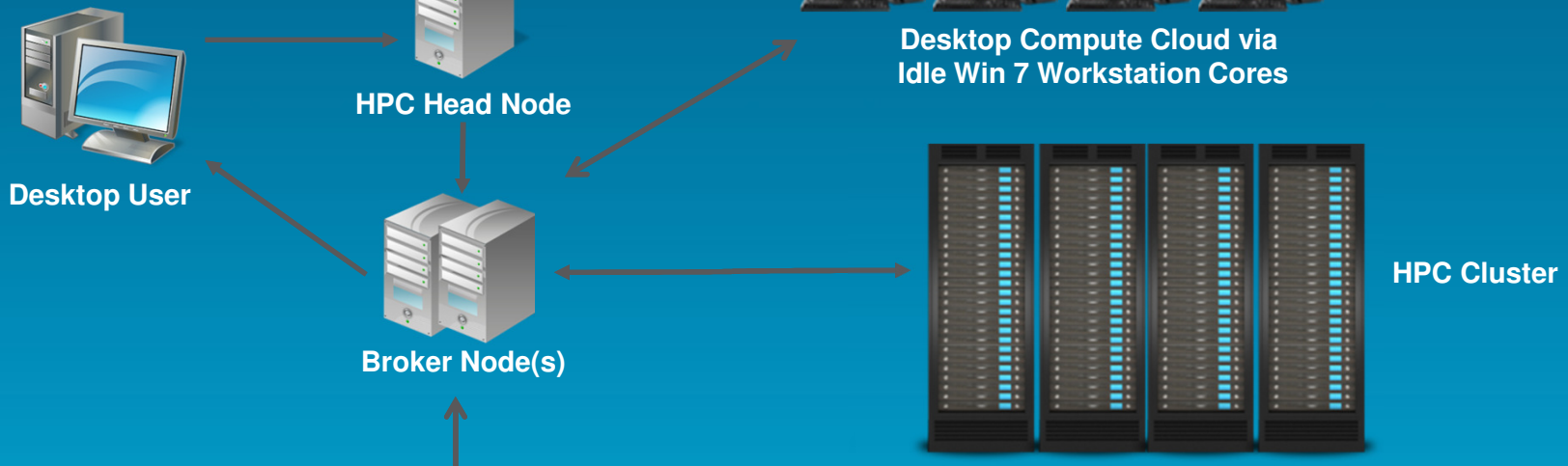
- Dev cycle… not great
  - How to debug?
- Our own queuing system.. ugh
- Performance?
- Dynamic (?) scalability

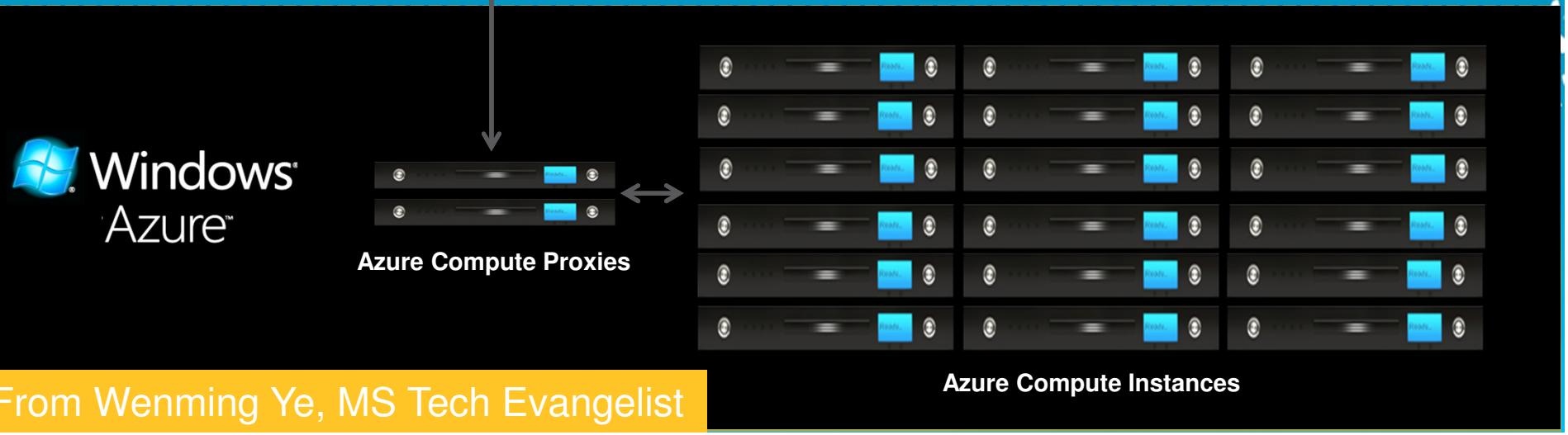
- Our enterprise and “The Cloud” : *The great divide*

- *Let’s “port” it to Win HPC (and more..)*

Windows HPC + Windows Azure Burst!

On-premise





From Wenming Ye, MS Tech Evangelist

MODISAzure: To Cloudburst or Not Cloudburst

- Dev/debug experience?
- Cost?
- Performance?
- Reliability?

- *Speed to science?*

Porting Our App *from* Windows Azure

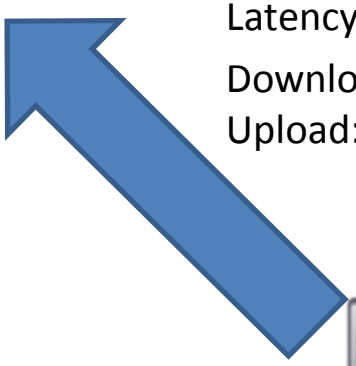
- Platform-specific behavior
  - In-line : if (“host in Azure”) { …} Else { … }
  - App.config
- Dev
  - Small cluster and RDP
- Issues
  - 8 / 16 threads trying to read/write a file
  - Built-in app “fault-tolerance”: ugh
  - $$$: no longer a concern

Wincluster (and Azure)

Chicago (~550 miles)

Latency: 21 ms

Download: 91.1 Mbps
Upload: 30.1 Mbps



Netgear
10/100/1000

San Antonio (~1300 miles)

Latency: 43 ms

Download: 61.7 Mbps
Upload: 15.9 Mbps

- Shuttle SN78S
- Dual core (AMD Athlon X2 2.8GHz)
- 4GB RAM
- C: 640GB (7200)

- Dual quad-core (AMD Opteron 2344 HE 1.7GHz)
- 16GB RAM
- C: 150GB (10000), D: 640GB (7200)

Adding Azure Node for MODIS

- Boot node (15-20 min)
- Install VPN code/endpoint (*Connect)*
- Create f: and hpcsync
- Install matlab runtime
- *Overall: 35 min [manual]*

Our packages

	Size	NC upload	SC upload
Modis app	1.33 MB	8.32s (7.9 – 8.9)	8.64s (7.4 – 10.8)
Hpc_client	17 MB	86.8 (28.2 – 128.2)	30.9 (25.5 – 35.2)
matlab	260 MB	628 (385.7 – 961.6)	295 (202.6 – 436.7)
Default input files	1.5 GB	2405 (1122 – 6948)	1391 (1130 – 2116)

Roll our own VM ?

- Task 1: Install Hyper-V and build VM (13 steps)
- Task 2: Preparing Base Image for Deployment (16 steps)
- Task 3: Installing the Windows Azure VM Role Integration Components (13 steps)
- Task 4: Uploading the Disk Image to Windows Azure (8 steps)
- Task 5: Creating the Service Model (17 steps)
- Task 6: Creating the Hosted Service and Deploying the Package (10 steps)

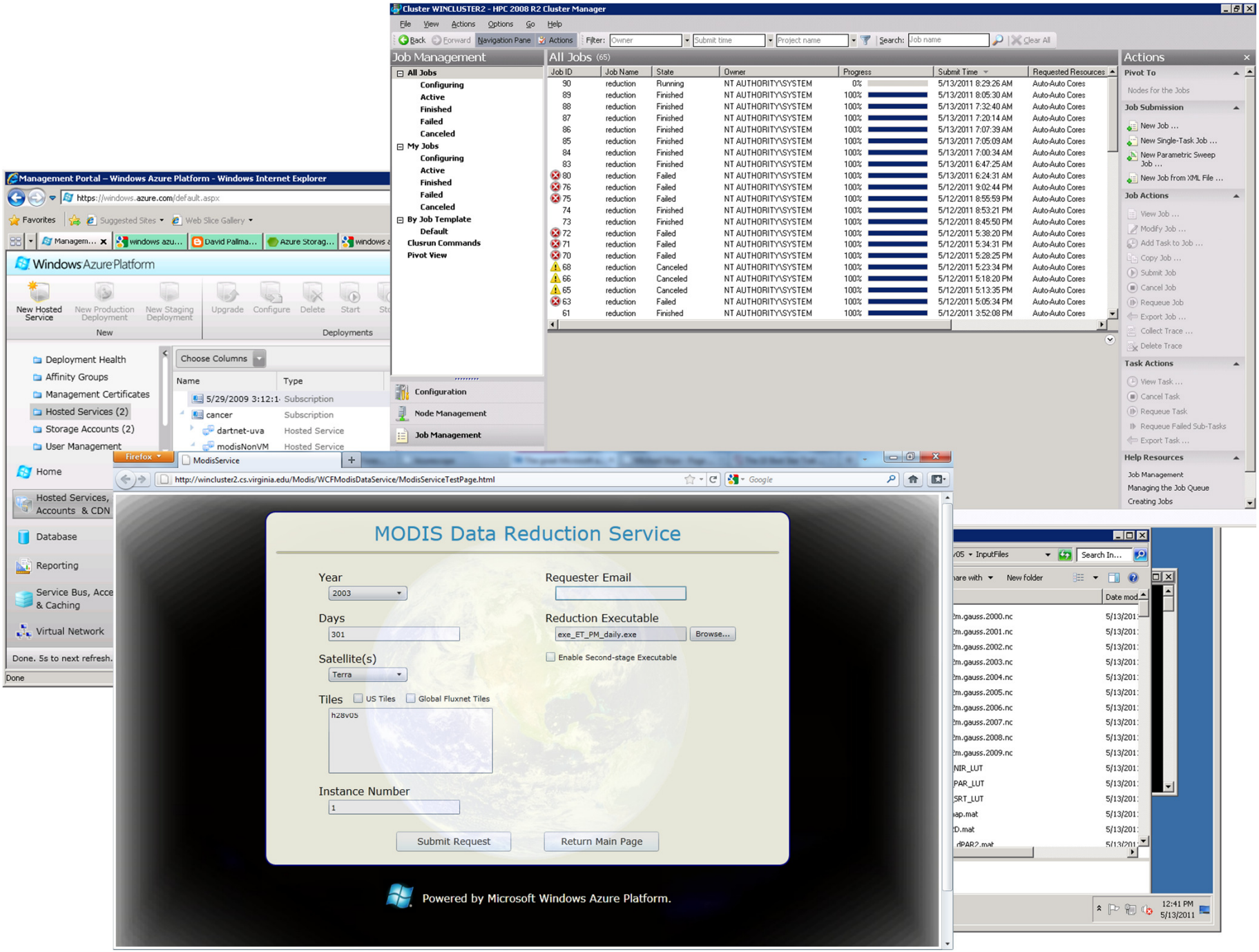
Azure Compute Instances

Compute Instance Size	CPU	Memory	Instance Storage	I/O Performance	Cost per hour
Extra Small	1.0 GHz	768 MB	20 GB	Low	$0.05
Small	1.6 GHz	1.75 GB	225 GB	Moderate	$0.12
Medium	2 x 1.6 GHz	3.5 GB	490 GB	High	$0.24
Large	4 x 1.6 GHz	7 GB	1,000 GB	High	$0.48
Extra Large	8 x 1.6 GHz	14 GB	2,040 GB	High	$0.96

Note: for $1K I can *buy* an ex-large-equivalent node or rent an ex-large VM for 1042 hrs (43 days)

Azure Costs – pay as you go (note: “subscription” offers exist)

- Storage
  - $0.15 per GB stored per month
  - $0.01 per 10K storage transactions
- SQL Azure
  - Web ed.: $9.99 per DB up to 1 GB /month
  - Business ed.: $ 99.99 per DB up to 10GB /month
- Data transfers
  - $0.10 per GB in
  - $0.15 per GB out


Cluster WINCLUSTER2 - HPC 2008 R2 Cluster Manager
File View Actions Options Go Help
Back Forward Navigation Pane Actions Filter: Owner Submit time Project name Search: Job name Clear All
Job Management
All Jobs
Configuring
Active
Finished
Failed
Canceled
My Jobs
Configuring
Active
Finished
Failed
Canceled
By Job Template
Default
Clusrun Commands
Pivot View
Configuration
Node Management
Job Management
All Jobs (65)
Job ID Job Name State Owner Progress Submit Time Requested Resources
90 reduction Running NT AUTHORITY\SYSTEM 0% 5/13/2011 8:29:26 AM Auto-Auto Cores
89 reduction Finished NT AUTHORITY\SYSTEM 100% 5/13/2011 8:05:30 AM Auto-Auto Cores
88 reduction Finished NT AUTHORITY\SYSTEM 100% 5/13/2011 7:32:40 AM Auto-Auto Cores
87 reduction Finished NT AUTHORITY\SYSTEM 100% 5/13/2011 7:20:14 AM Auto-Auto Cores
86 reduction Finished NT AUTHORITY\SYSTEM 100% 5/13/2011 7:07:39 AM Auto-Auto Cores
85 reduction Finished NT AUTHORITY\SYSTEM 100% 5/13/2011 7:05:09 AM Auto-Auto Cores
84 reduction Finished NT AUTHORITY\SYSTEM 100% 5/13/2011 7:00:34 AM Auto-Auto Cores
83 reduction Finished NT AUTHORITY\SYSTEM 100% 5/13/2011 6:47:25 AM Auto-Auto Cores
80 reduction Failed NT AUTHORITY\SYSTEM 100% 5/13/2011 6:24:31 AM Auto-Auto Cores
76 reduction Failed NT AUTHORITY\SYSTEM 100% 5/12/2011 9:02:44 PM Auto-Auto Cores
75 reduction Failed NT AUTHORITY\SYSTEM 100% 5/12/2011 8:55:59 PM Auto-Auto Cores
74 reduction Finished NT AUTHORITY\SYSTEM 100% 5/12/2011 8:53:21 PM Auto-Auto Cores
73 reduction Finished NT AUTHORITY\SYSTEM 100% 5/12/2011 8:45:50 PM Auto-Auto Cores
72 reduction Failed NT AUTHORITY\SYSTEM 100% 5/12/2011 5:38:20 PM Auto-Auto Cores
71 reduction Failed NT AUTHORITY\SYSTEM 100% 5/12/2011 5:34:31 PM Auto-Auto Cores
70 reduction Failed NT AUTHORITY\SYSTEM 100% 5/12/2011 5:28:25 PM Auto-Auto Cores
68 reduction Canceled NT AUTHORITY\SYSTEM 100% 5/12/2011 5:23:34 PM Auto-Auto Cores
66 reduction Canceled NT AUTHORITY\SYSTEM 100% 5/12/2011 5:18:20 PM Auto-Auto Cores
65 reduction Canceled NT AUTHORITY\SYSTEM 100% 5/12/2011 5:13:35 PM Auto-Auto Cores
63 reduction Failed NT AUTHORITY\SYSTEM 100% 5/12/2011 5:05:34 PM Auto-Auto Cores
61 reduction Finished NT AUTHORITY\SYSTEM 100% 5/12/2011 3:52:08 PM Auto-Auto Cores
Actions
Pivot To
Nodes for the Jobs
Job Submission
New Job ...
New Single-Task Job ...
New Parametric Sweep Job ...
New Job from XML File ...
Job Actions
View Job ...
Modify Job ...
Add Task to Job ...
Copy Job ...
Submit Job
Cancel Job
Requeue Job
Export Job ...
Collect Trace ...
Delete Trace
Task Actions
View Task ...
Cancel Task
Requeue Task
Requeue Failed Sub-Tasks
Export Task ...
Help Resources
Job Management
Managing the Job Queue
Creating Jobs
Management Portal – Windows Azure Platform - Windows Internet Explorer
https://windows.azure.com/default.aspx
Favorites Suggested Sites Web Slice Gallery
Windows Azure Platform
New Hosted Service
New Production Deployment
New Staging Deployment
Upgrade Configure Delete Start
New
Deployments
Deployment Health
Affinity Groups
Management Certificates
Hosted Services (2)
Storage Accounts (2)
User Management
Choose Columns
Name Type
5/29/2009 3:12:1 Subscription
cancer Subscription
dartnet-uva Hosted Service
modisNonVM Hosted Service
Home
Hosted Services, Accounts & CDN
Database
Reporting
Service Bus, Acce & Caching
Virtual Network
Done. 5s to next refresh.
Done
Firefox
ModisService
http://wincluster2.cs.virginia.edu/Modis/WCFModisDataService/ModisServiceTestPage.html
MODIS Data Reduction Service
Year
2003
Days
301
Satellite(s)
Terra
Tiles
US Tiles
Global Fluxnet Tiles
h28v05
Instance Number
1
Requester Email
Reduction Executable
exe_ET_PM_daily.exe
Browse...
Enable Second-stage Executable
Submit Request
Return Main Page
Powered by Microsoft Windows Azure Platform.
InputFiles
Search In...
New folder
Date mod
2m.gauss.2000.nc 5/13/201
2m.gauss.2001.nc 5/13/201
2m.gauss.2002.nc 5/13/201
2m.gauss.2003.nc 5/13/201
2m.gauss.2004.nc 5/13/201
2m.gauss.2005.nc 5/13/201
2m.gauss.2006.nc 5/13/201
2m.gauss.2007.nc 5/13/201
2m.gauss.2008.nc 5/13/201
2m.gauss.2009.nc 5/13/201
NIR_LUT 5/13/201
PAR_LUT 5/13/201
SRT_LUT 5/13/201
ap.mat 5/13/201
D.mat 5/13/201
dPAR2.mat
12:41 PM
5/13/2011

Results

Single Day Computation

- 2003, DOY=301, h28v05, Terra
- ET_PM_daily: 2,543 KB
- Input: 219 files, 3.0 GB total
  - Pre-staged: 76 files, 2.49 GB total
  - HDF: 143 files, 517 MB total
- Temp files: 173 files, 18.1 MB total
- Output: 32 files, 358K total

2003, DOY=301, h28v05, Terra Execution Time (minutes)

	Stage-in	Compute	Total
local	0:38 (z:)	6.59	7:40
Medium (local)	0:46	1.73	3:36
Medium (blob)	10:52	1.73	13:36
Medium (UVa)	14:21	1.73	16:59
Large (local)	0:47	1.70	3:35
Large (blob)	11:01	1.70	13:36
Large (UVa)	14:29	1.70	16:37
Ex-Large (local)	0:36	1.75	3:33
Ex-Large (blob)	10:50	1.75	13:29
Ex-Large (UVa)	14:06	1.75	16:49

2003, DOY=301-316, h28v05, Terra

[*w/ stage 2 reduction*]

- Total input: 216 files, 859 MB [+ pre-staged]
- Total output: 118 files, 1.58 MB

	Execution Time (minutes)	Network ($)	Storage ($)	Compute ($)	Total ($)
Local (1x2x4)	22:34	--	--	--	--
4 Medium (blob,cold)	19:19	$ 0.08	$ 0.45 *per month*	$ 0.31	$ 0.38
2 Large (blob, cold)	18:41	$ 0.08	$ 0.45 *per month*	$ 0.30	$ 0.38
1 Ex-Large (blob, cold)	20:39	$ 0.08	$ 0.45 *per month*	$ 0.33	$ 0.41

Note: Network cost is $$ to upload 216 files (once) to Windows Azure blob storage

2003, Full Year, h28v05, Terra

- Total input: 2517 files, 9.46 GB [+ pre-staged]
  - Cost to upload: $ 0.95; cost per month to store: $ 1.42
- 190 GB (301K files/links) total written by daily computations
- Total output: 1st stage: 11680 files, 126 MB; 2nd stage: 118 files, 5.79 MB

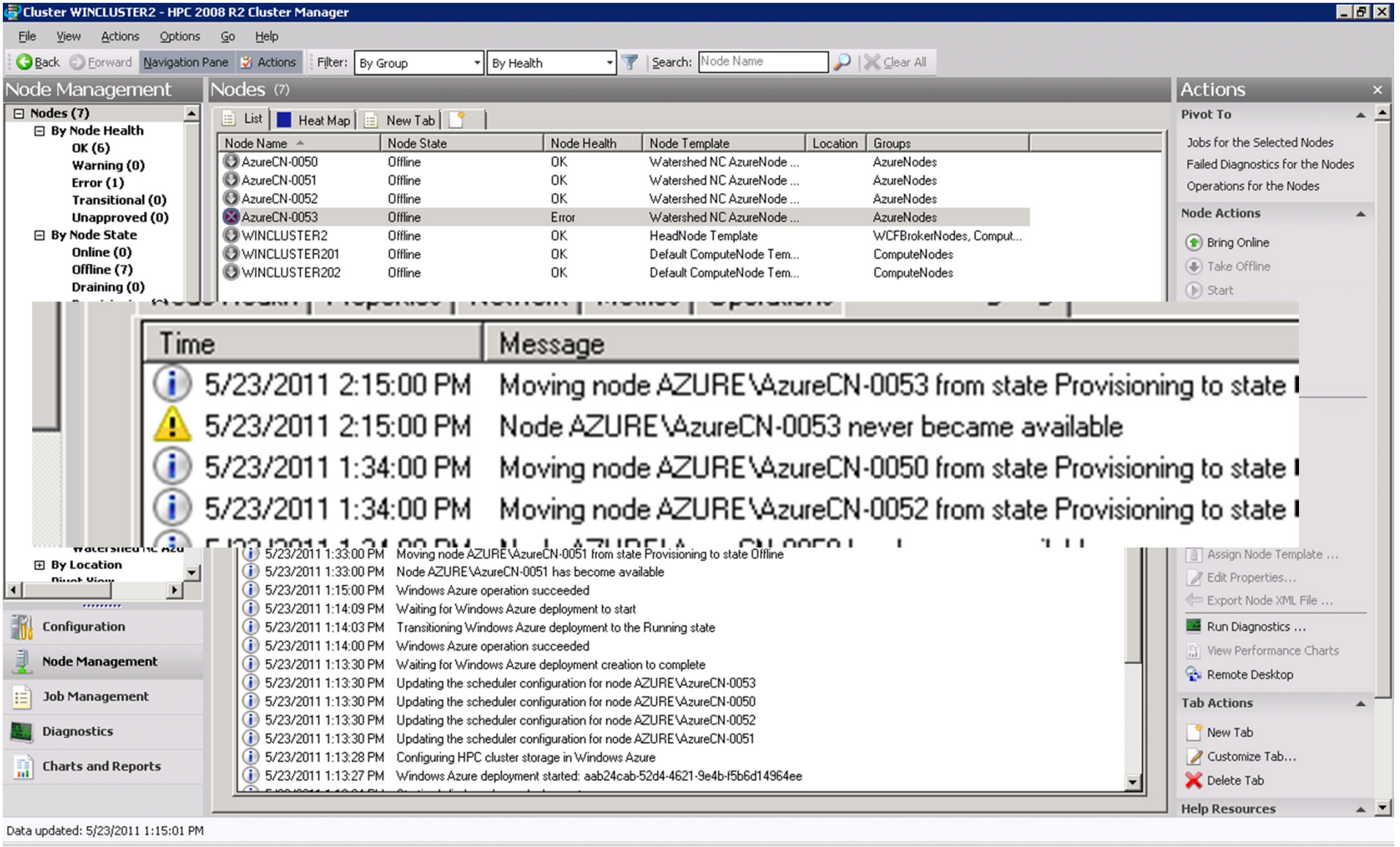
	Execution Time	Compute ($)
Local (2x2x4 = 16 cores)	2 hrs 39	$ 0
1 Ex-large (8 cores)	5 hrs 21	1 * $0.96 * 5.35 = $ 5.14
2 Ex-large	2 hrs 49	2 * $0.96 * 2.82 = $ 4.65
4 Ex-large	1 hr 32	4 * $0.96 * 1.53 = $ 5.88
Hybrid: Local + 4 Ex-large	1 hr 29	4 * $0.96 * 1.48 = $ 5.70
8 Ex-large	1 hr 5	8 * $0.96 * 1.08 = $ 8.32
Hybrid: Local + 8 Ex-large	1 hr 13	8 * $0.96 * 1.22 = $ 9.37

Cloudbursting Our App

- Cloudbursting fits well for us
  - better than Azure-only (dev experience)
  - better than Win HPC enterprise-only (generic cloud arguments)
- Win HPC only Azure nodes or Win HPC enterprise and Azure? (aka: should we even bother with our own existing HW?)
    - Disclaimer: 16 nodes, circa 2008, “personal cluster” (NOT your average case?)
  - Yes: they’re “free”
  - No: ugh – heterogeneity in code, wide-area run-time complexity (worst of both worlds?)

Cloudbursting Your App

- Generic cloud arguments apply
  - Don't leave VMs running unused, don't forget about stuff you've stored in the cloud, speed of light is not getting any faster…
- DON'T roll your own queuing system in the cloud!
  - Esp. if you need it to be multi-user
  - However, no policy to "only run the apps that 'make sense' in the cloud"
- You might start seeing time-outs in your code
  - And in the cluster boot as well ….


Cluster WINCLUSTER2 - HPC 2008 R2 Cluster Manager
File View Actions Options Go Help
Back Forward Navigation Pane Actions Filter: By Group By Health Search: Node Name Clear All
Node Management
Nodes (7)
By Node Health
OK (6)
Warning (0)
Error (1)
Transitional (0)
Unapproved (0)
By Node State
Online (0)
Offline (7)
Draining (0)
By Location
Configuration
Node Management
Job Management
Diagnostics
Charts and Reports
Nodes (7)
List Heat Map New Tab
Node Name Node State Node Health Node Template Location Groups
AzureCN-0050 Offline OK Watershed NC AzureNode ... AzureNodes
AzureCN-0051 Offline OK Watershed NC AzureNode ... AzureNodes
AzureCN-0052 Offline OK Watershed NC AzureNode ... AzureNodes
AzureCN-0053 Offline Error Watershed NC AzureNode ... AzureNodes
WINCLUSTER2 Offline OK HeadNode Template WCFBrokerNodes, Comput...
WINCLUSTER201 Offline OK Default ComputeNode Tem... ComputeNodes
WINCLUSTER202 Offline OK Default ComputeNode Tem... ComputeNodes
Time Message
5/23/2011 2:15:00 PM Moving node AZURE\AzureCN-0053 from state Provisioning to state
5/23/2011 2:15:00 PM Node AZURE\AzureCN-0053 never became available
5/23/2011 1:34:00 PM Moving node AZURE\AzureCN-0050 from state Provisioning to state
5/23/2011 1:34:00 PM Moving node AZURE\AzureCN-0052 from state Provisioning to state
5/23/2011 1:33:00 PM Moving node AZURE\AzureCN-0051 from state Provisioning to state Offline
5/23/2011 1:33:00 PM Node AZURE\AzureCN-0051 has become available
5/23/2011 1:15:00 PM Windows Azure operation succeeded
5/23/2011 1:14:09 PM Waiting for Windows Azure deployment to start
5/23/2011 1:14:03 PM Transitioning Windows Azure deployment to the Running state
5/23/2011 1:14:00 PM Windows Azure operation succeeded
5/23/2011 1:13:30 PM Waiting for Windows Azure deployment creation to complete
5/23/2011 1:13:30 PM Updating the scheduler configuration for node AZURE\AzureCN-0053
5/23/2011 1:13:30 PM Updating the scheduler configuration for node AZURE\AzureCN-0050
5/23/2011 1:13:30 PM Updating the scheduler configuration for node AZURE\AzureCN-0052
5/23/2011 1:13:30 PM Updating the scheduler configuration for node AZURE\AzureCN-0051
5/23/2011 1:13:28 PM Configuring HPC cluster storage in Windows Azure
5/23/2011 1:13:27 PM Windows Azure deployment started: aab24cab-52d4-4621-9e4b-f5b6d14964ee
Actions
Pivot To
Jobs for the Selected Nodes
Failed Diagnostics for the Nodes
Operations for the Nodes
Node Actions
Bring Online
Take Offline
Start
Assign Node Template ...
Edit Properties...
Export Node XML File ...
Run Diagnostics ...
View Performance Charts
Remote Desktop
Tab Actions
New Tab
Customize Tab...
Delete Tab
Help Resources
Data updated: 5/23/2011 1:15:01 PM

Cloudbursting Your App (cont.)

- Does your app fit this pattern?

```
#!\bin\bash
scp foo.cs.virginia.edu:input.zip .
foo.exe
scp output.txt foo.cs.virginia.edu:
```

- App source code can get ugly (a la "#ifdef cloud")
  - *UVa PhD student Zach Hill: CSAL*
- Homogeneity is good
  - E.g., make the Azure nodes look like Enterprise nodes (e.g., F:)
- Hmm… Dryad ("Linq to HPC") knows where your data is (to some degree) but HPC Scheduler does not
- *But my app is MPI…*

Summary

- MODISAzure: bag-of-tasks, large input, small computation, small output
- Good fit for Windows Azure cloudbursting
  - Good performance, good ease of use, good dev exp, good $$, good management exp, no cap-ex
- Need more studies on long-term reliability, how to get “fastest results for cheapest $$$”
  - Hmmm.. “#ifdef cloud” or not?
- *Border between enterprise and cloud is still too thick*