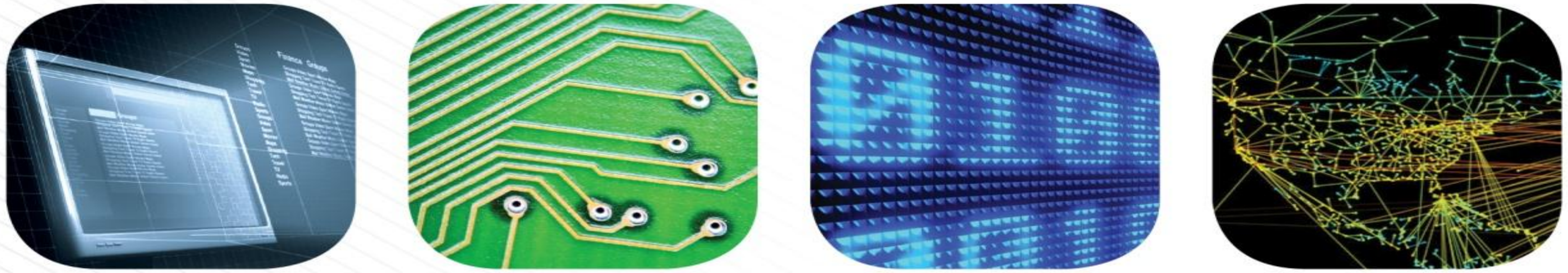


Data to INSIGHT CENTER

Towards Enabling Mid-Scale Geo-Science Experiments Through Microsoft Trident and Windows Azure

Eran Chinthaka Withana

Beth Plale

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

Agenda

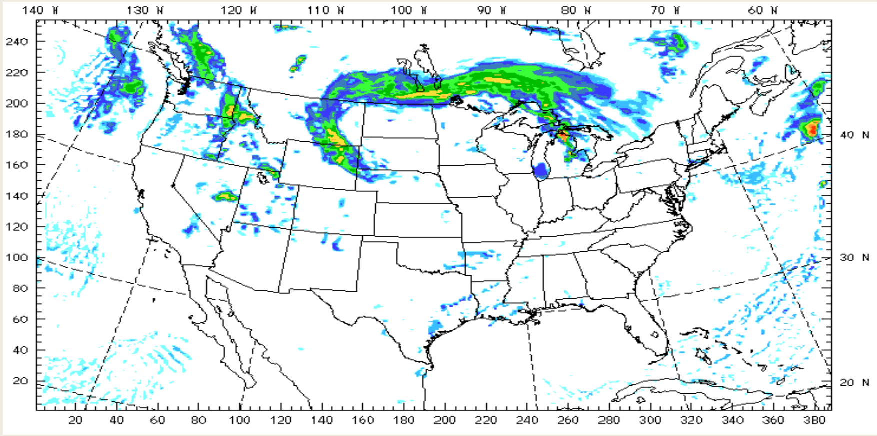
- Geo-Science Applications: Challenges and Opportunities
- Research Vision
- Proposed Framework
- Applications
  - Scheduling time-critical MPI applications in Windows Azure
  - Scheduling large number of small jobs (ensembles) in Windows Azure

Agenda

- Geo-Science Applications: Challenges and Opportunities
- Research Vision
- Proposed Framework
- Applications
  - Scheduling time-critical MPI applications in Windows Azure
  - Scheduling large number of small jobs (ensembles) in Windows Azure

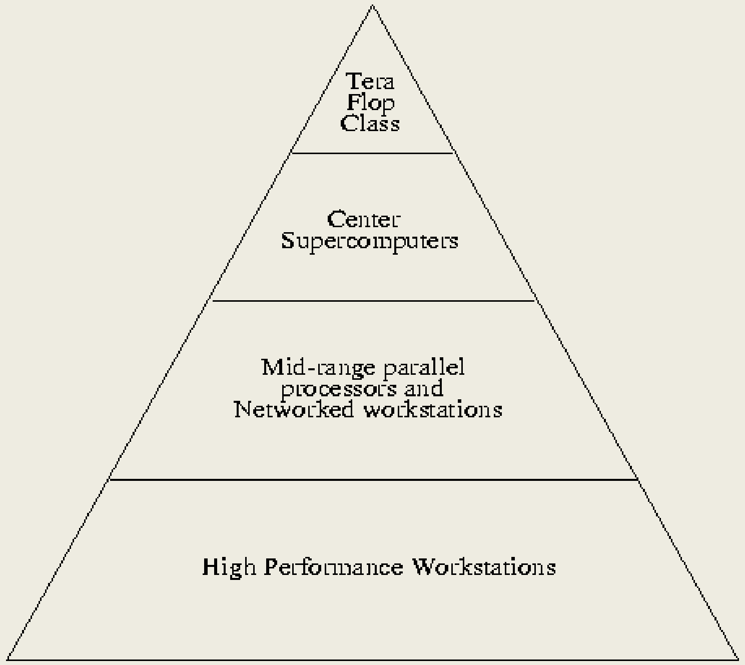
Geo-Science Applications

- High Resource Requirements
  - Compute intensive, dedicated HPC hardware
    - e.g. Weather Research and Forecasting (WRF) Model
- Emergence of ensemble applications
  - Large amount of small jobs
    - e.g. Examining each air layer, over a long period of time.
    - Single experiment = About 14000 jobs each taking few minutes to complete



Geo-Science Applications: Challenges

- Compute intensive applications
  - Mid-scale scientists
    - often scramble to find sufficient computational resources to test and run their codes
- Software requirements and platform dependence
  - MPI, Cygwin (if windows), Linux only binaries
- Management of large job executions
- Fault tolerance
- Reliability* of Grid computing resources and middleware
- Utilizing different compute resources


Tera Flop Class
Center Supercomputers
Mid-range parallel processors and Networked workstations
High Performance Workstations

*Marru S, Perera S, Feller M, Martin S. Reliable and Scalable Job Submission: LEAD Science Gateways Testing and Experiences with WS GRAM on TeraGrid Resources . TeraGrid Conference June 2008

Geo-Science Applications: Opportunities

- Cloud computing resources
  - On-demand access to “unlimited” resources
  - Flexibility
    - Worker roles and VM roles
- Recent porting of geo-science applications
  - WRF, WRF Preprocessing System (WPS) port to Windows
- Increased use of ensemble applications (large number of small runs)
- Production quality, opensource scientific workflow systems
  - Microsoft Trident

Agenda

- Geo-Science Applications: Challenges and Opportunities
- Research Vision
- Proposed Framework
- Applications
  - Scheduling time-critical MPI applications in Windows Azure
  - Scheduling large number of small jobs (ensembles) in Windows Azure

Research Vision

- Enabling geo-science experiments
  - Type of applications
    - Compute intensive, ensembles
  - Type of scientists
    - Meteorologists, atmospheric scientists, emergency management personnel, geologists
- Utilizing both Cloud computing and Grid computing resources
- Utilizing opensource, production quality scientific workflow environments
- Improved data and meta-data management

Geo-Science Applications

SLOSH

WRF

Scientific Workflows

Microsoft Project Trident

Compute Resources

Windows Azure

TeraGrid

amazon web services

Existing Approaches

- GRAM
  - Features
    - Coordinates job submissions to Grid computing resources
  - Limitations
    - Scalability and reliability issues
    - Ease of installation and maintenance

- CARMEN project
  - Features
    - Concentrates on building a cloud environment for neuroscientists
      - Provide data sharing and analysis capabilities
    - Encapsulates tools as WS-I compliant web services
    - Dynamic deployments using Dynasoar
  - Limitations
    - Strict application requirements
    - Ability to support wide variety of compute resources

- Condor
  - Features
    - Enables creation of resource pools from grid and cloud computing resources
  - Limitations
    - On-demand resource allocation and management
    - Ease of integration with workflow environments
- GridWay, SAGA, Falcon
  - Limitations
    - tightly integrated with complex middleware to address a broad range of problems

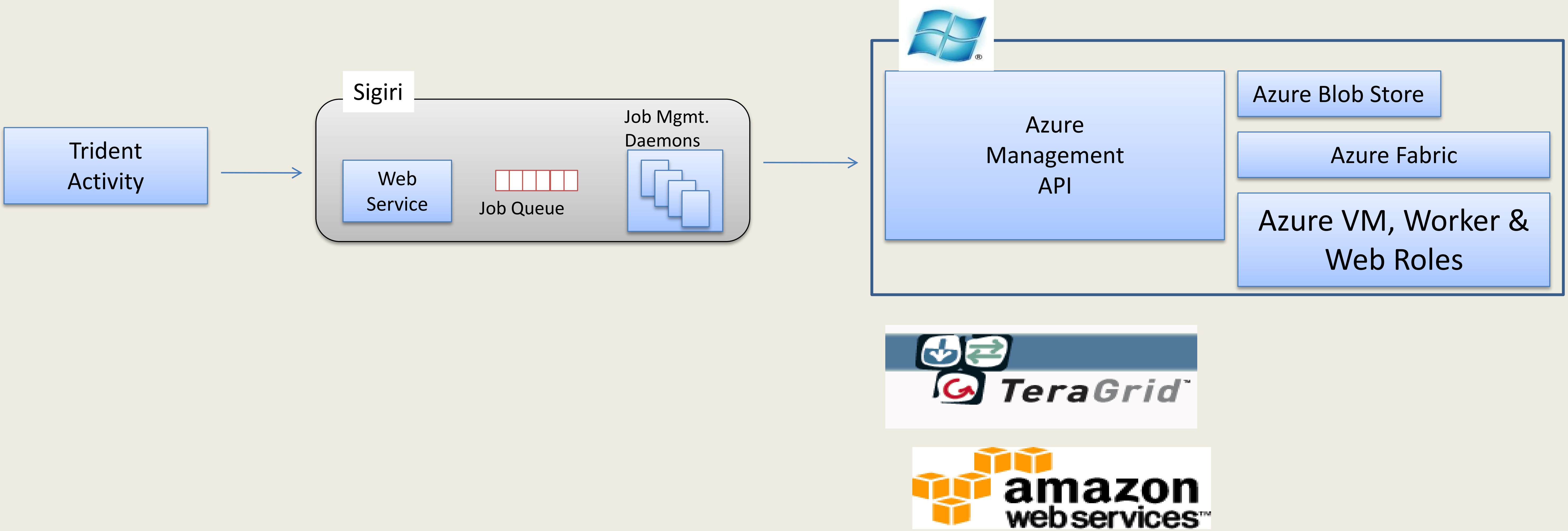
Agenda

- Geo-Science Applications: Challenges and Opportunities
- Research Vision
- Proposed Framework
- Applications
  - Scheduling time-critical MPI applications in Windows Azure
  - Scheduling large number of small jobs (ensembles) in Windows Azure

Design Decisions

- Decoupled architecture with low turnaround time
- Web services interfaces for interactions
  - Ease of integrating with workflow engines and tools
- Ability to support multiple job description languages
  - e.g. JSDL and RSL
- Flexibility to support various security protocols
  - transport level security and WS-Security
- Extensibility to support a range of compute resources
  - Should support grid and cloud resources
  - Should be able to schedule and monitor jobs
- Robust management of scientific jobs
  - Experiences with GRAM2 and GRAM4
- Ease of installation and maintenance

Proposed Framework



Trident Activity

Sigiri

Web Service

Job Queue

Job Mgmt. Daemons

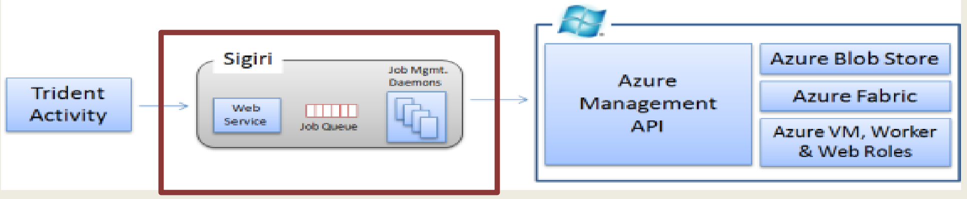
Azure Management API

Azure Blob Store

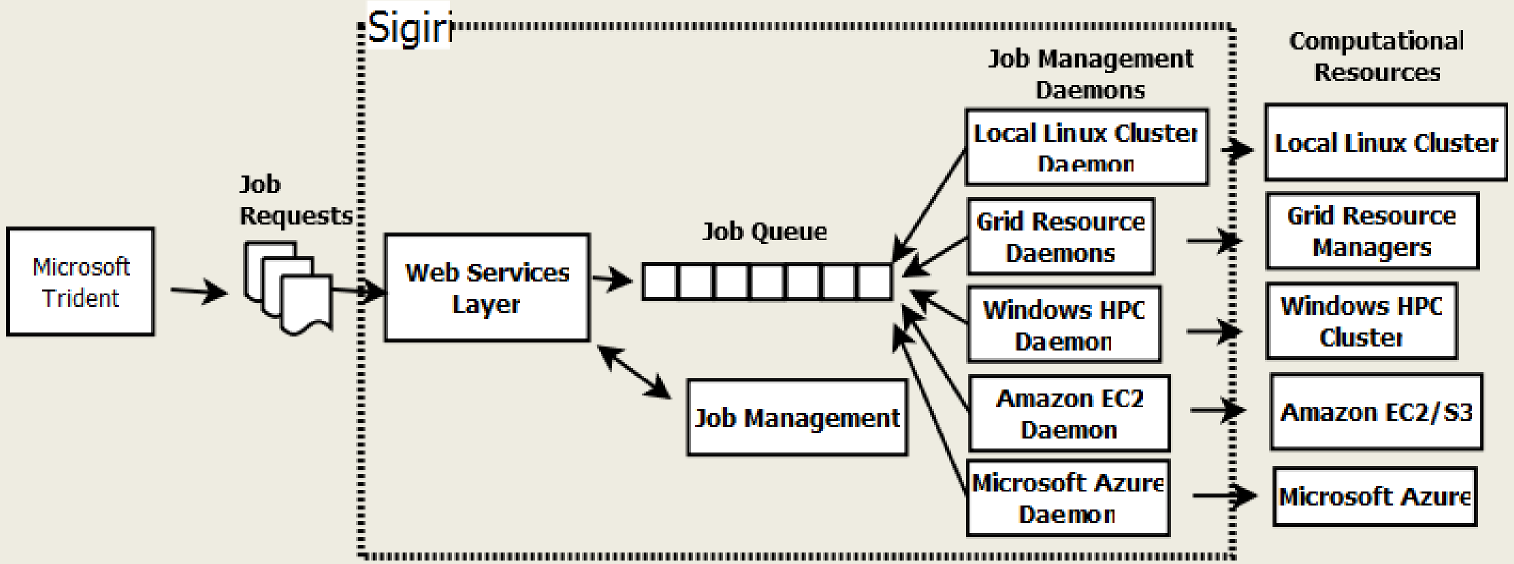
Azure Fabric

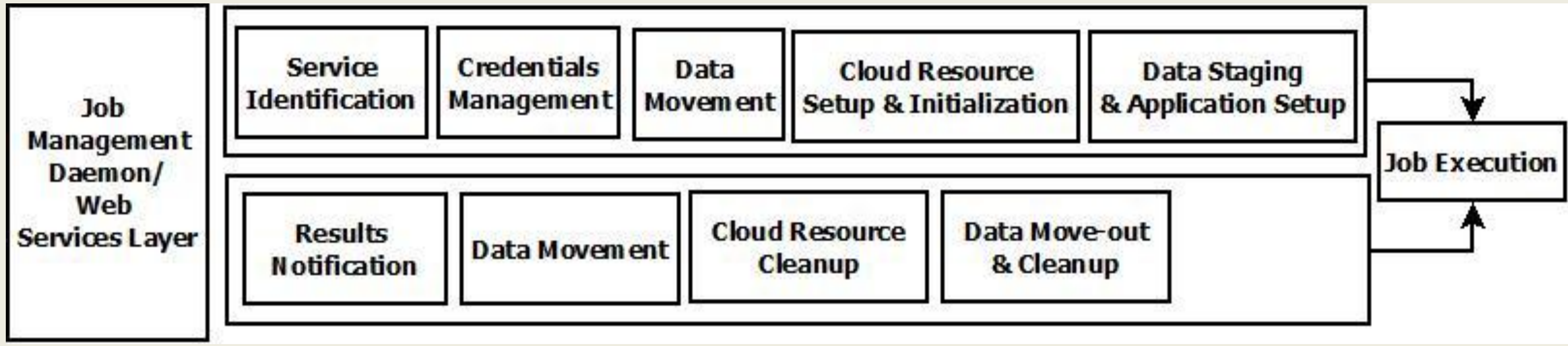
Azure VM, Worker & Web Roles

Proposed Framework

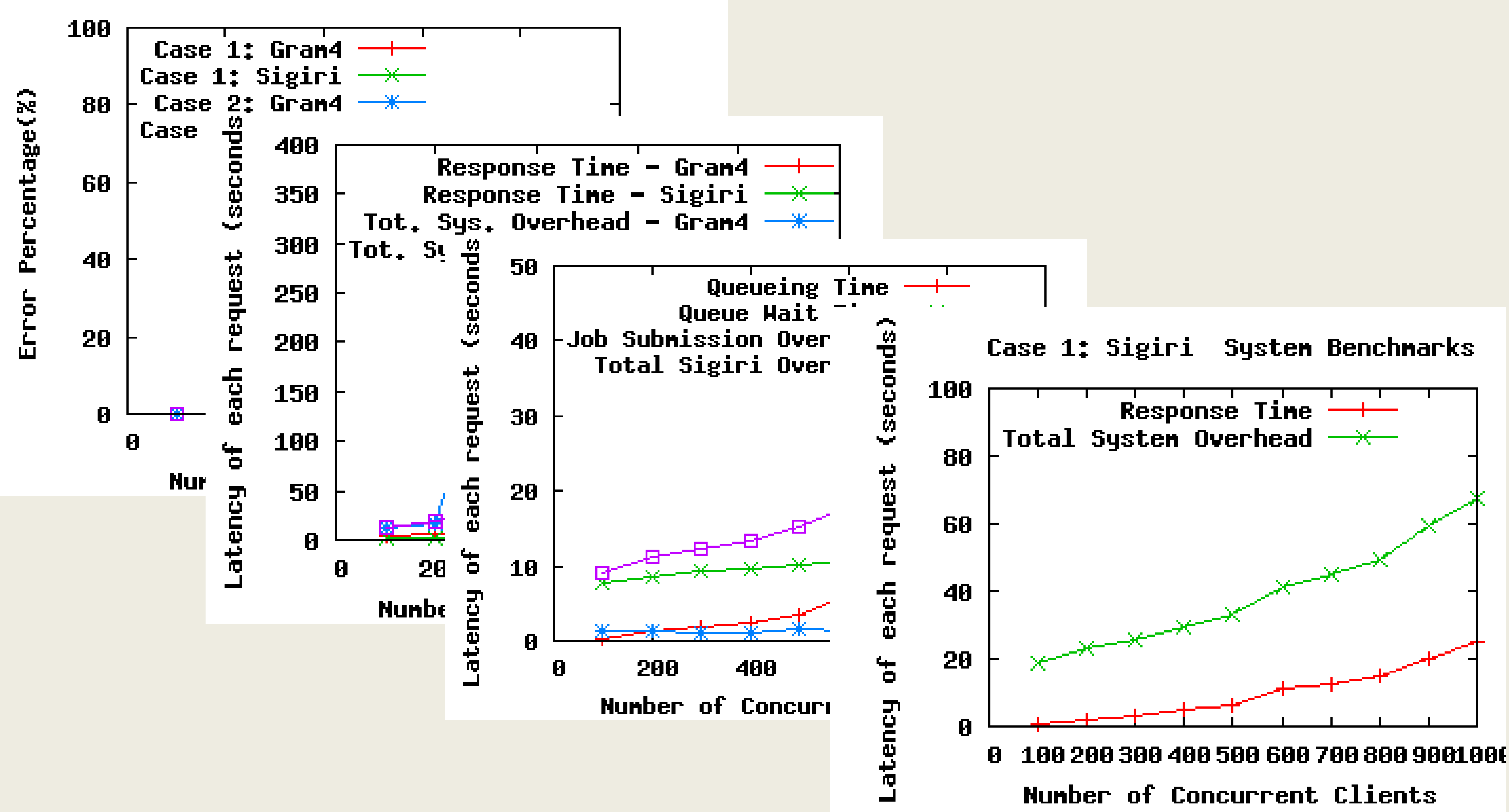


- Sigiri – Abstraction for grids and clouds
  - Web service
    - Decouples job acceptance from execution and monitoring
  - Daemons
    - Manages compute resource interactions
      - Job submissions and monitoring
      - Cleaning up resources
      - Efficient allocation of resources
    - Template based approach for cloud computing resources

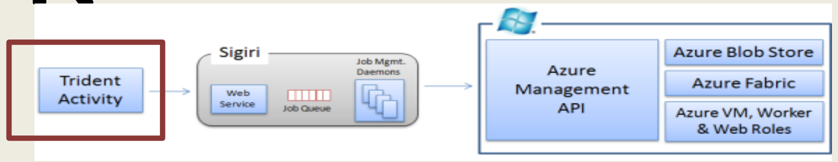




Performance Evaluation



Proposed Framework



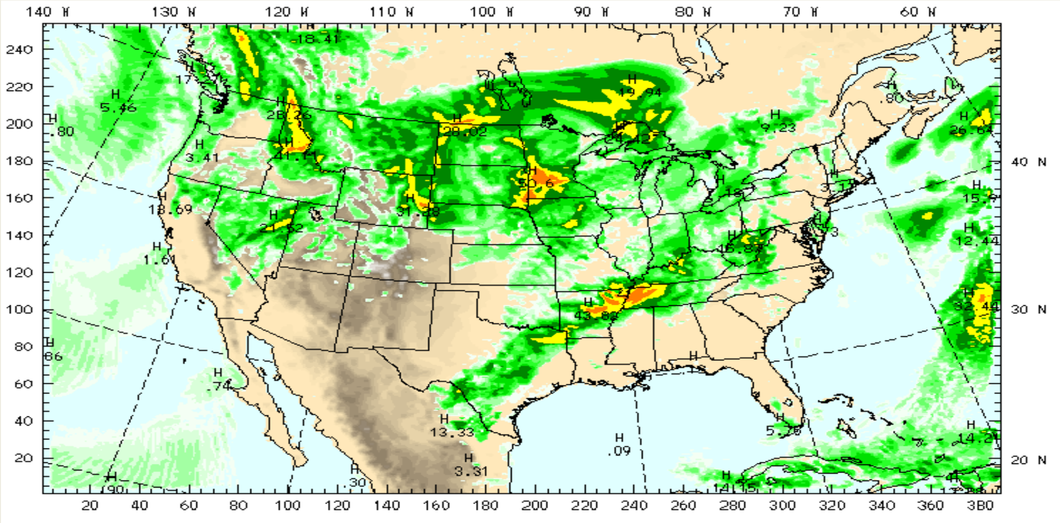
- Trident Activity
  - Activities compose a workflow
  - Activity wraps a task / application
    - Input parameter collection and validation
    - Request composition
    - Invocation and monitoring
  - Framework activities
    - Interacts with Sigiri to schedule jobs and monitor the progress
    - Data movement to / from cloud storage (Windows Blob Store or Amazon S3)
    - Visualization

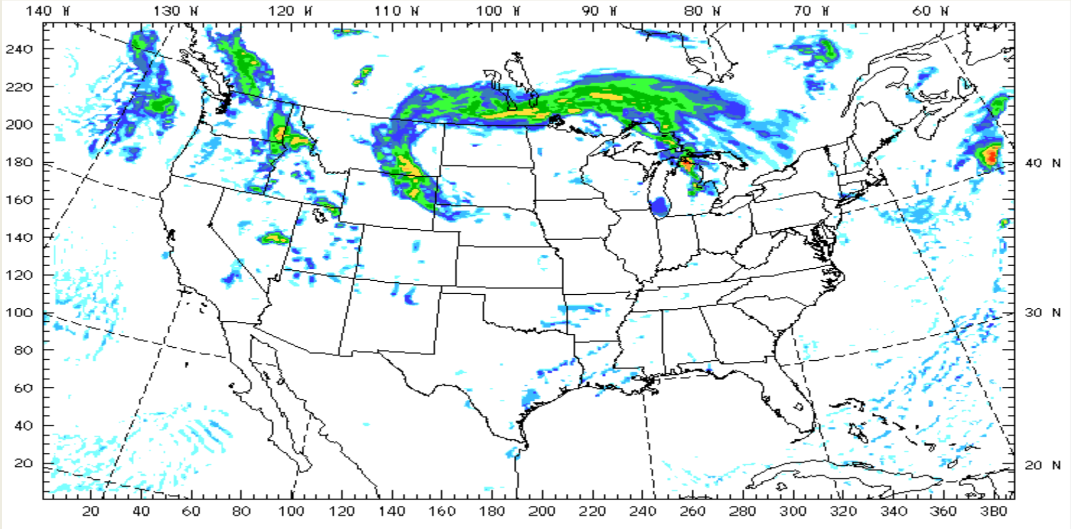
Agenda

- Geo-Science Applications: Challenges and Opportunities
- Research Vision
- Proposed Framework
- Applications
  - Scheduling time-critical MPI applications in Windows Azure
  - Scheduling large number of small jobs (ensembles) in Windows Azure

Weather Research and Forecast Model (WRF)

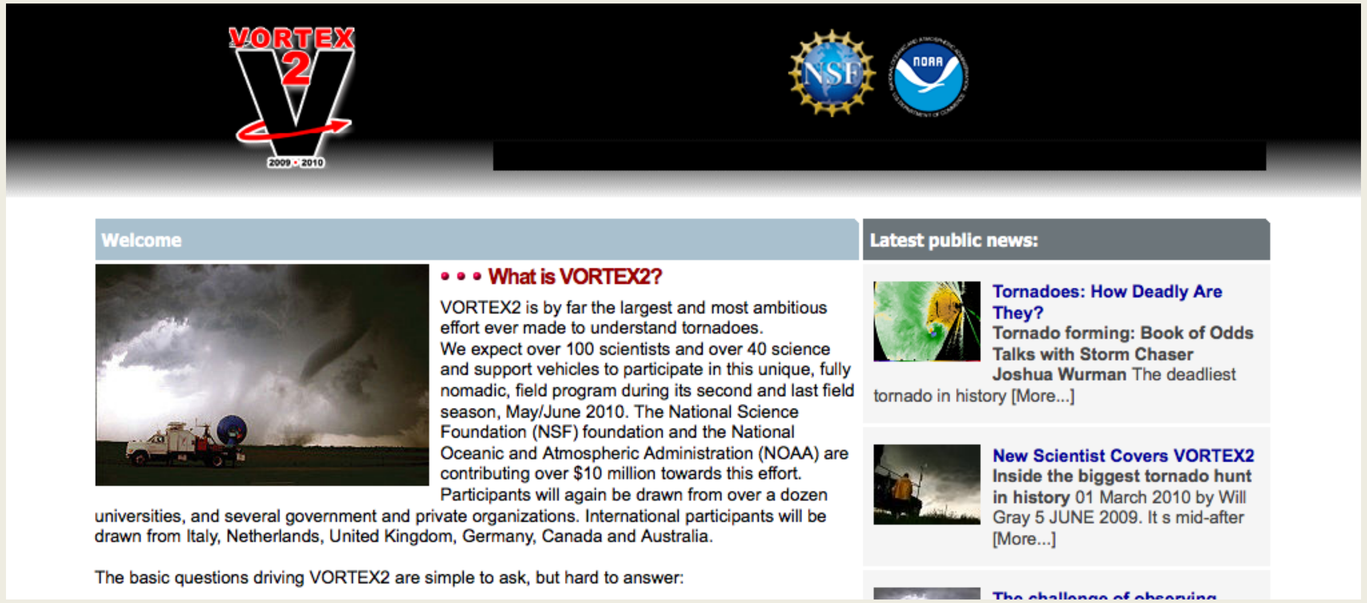
- Mesoscale numerical weather prediction system
- Designed to serve both operational forecasting and atmospheric research needs.
- A software architecture allowing for computational parallelism and system extensibility





Background: LEAD II and Vortex2 Experiment

- May 1, 2010 to June 15, 2010
- ~6 weeks, 7-days per week
- Workflow started on the hour every hour each morning.
- Had to find and bind to latest model data (i.e., RUC 13km and ADAS data) to set initial and boundary conditions.
  - If model data was not available at NCEP and University of Oklahoma, workflow could not begin.
- Execution of complete WRF stack within 1 hour



Welcome

• • • What is VORTEX2?

VORTEX2 is by far the largest and most ambitious effort ever made to understand tornadoes. We expect over 100 scientists and over 40 science and support vehicles to participate in this unique, fully nomadic, field program during its second and last field season, May/June 2010. The National Science Foundation (NSF) foundation and the National Oceanic and Atmospheric Administration (NOAA) are contributing over $10 million towards this effort. Participants will again be drawn from over a dozen universities, and several government and private organizations. International participants will be drawn from Italy, Netherlands, United Kingdom, Germany, Canada and Australia.

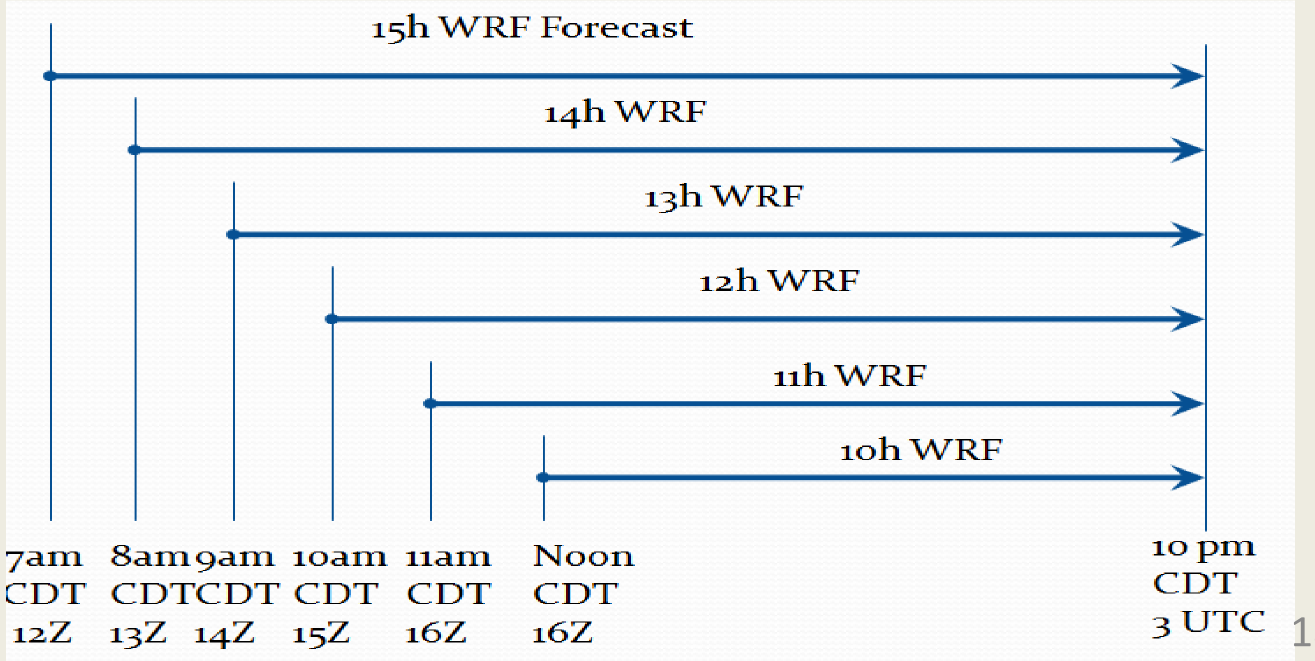
The basic questions driving VORTEX2 are simple to ask, but hard to answer:

Latest public news:

Tornadoes: How Deadly Are They?
Tornado forming: Book of Odds Talks with Storm Chaser Joshua Wurman The deadliest tornado in history [More...]

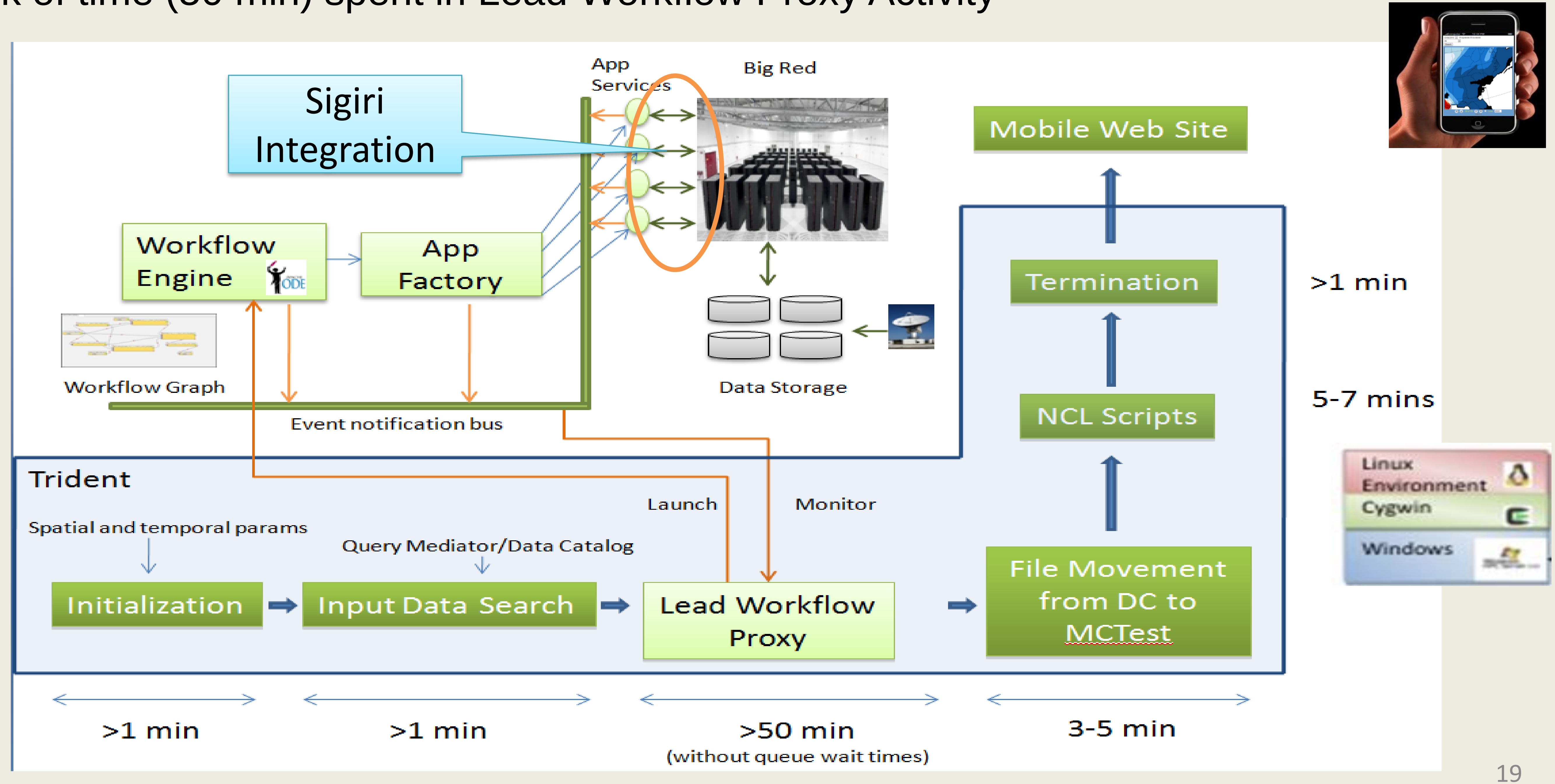
New Scientist Covers VORTEX2
Inside the biggest tornado hunt in history 01 March 2010 by Will Gray 5 JUNE 2009. It s mid-after [More...]

The challenge of observing



The Trident Vortex2 Workflow: Timeline

Bulk of time (50 min) spent in Lead Workflow Proxy Activity



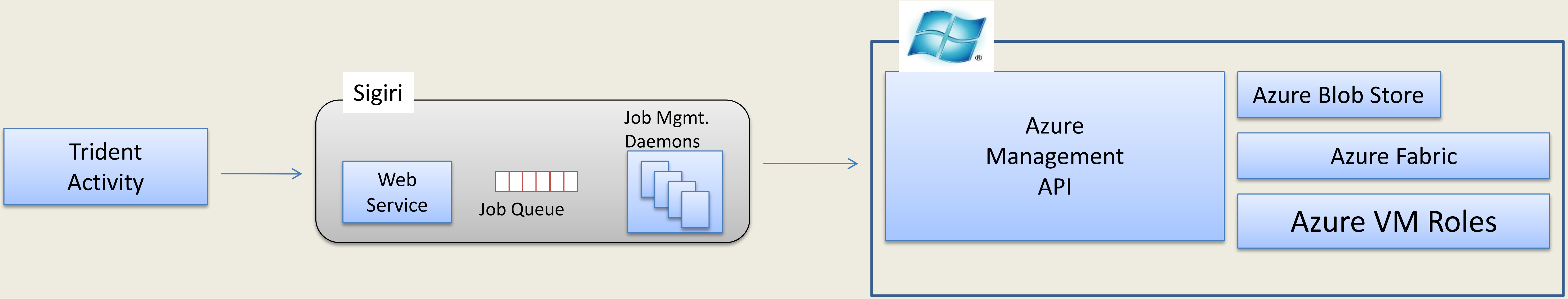
Agenda

- Geo-Science Applications: Challenges and Opportunities
- Research Vision
- Proposed Framework
- Applications
  - Scheduling time-critical MPI applications in Windows Azure
  - Scheduling large number of small jobs (ensembles) in Windows Azure

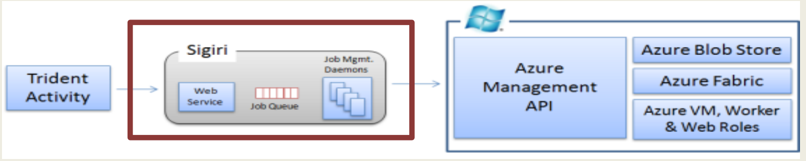
Moving WRF Stack to Windows Azure

- Opportunities
  - Reliability of grid computing resources
  - WRF and WRF Preprocessing System (WPS) ported to Windows
  - Enable midscale scientists to exploit the capabilities of WRF
- Concerns
  - Porting of WRF to Azure to run on multiple nodes
  - Strict software requirements and the choice between worker and VM Roles
    - Need of MPI, Cygwin
    - Restricted to single virtual machine

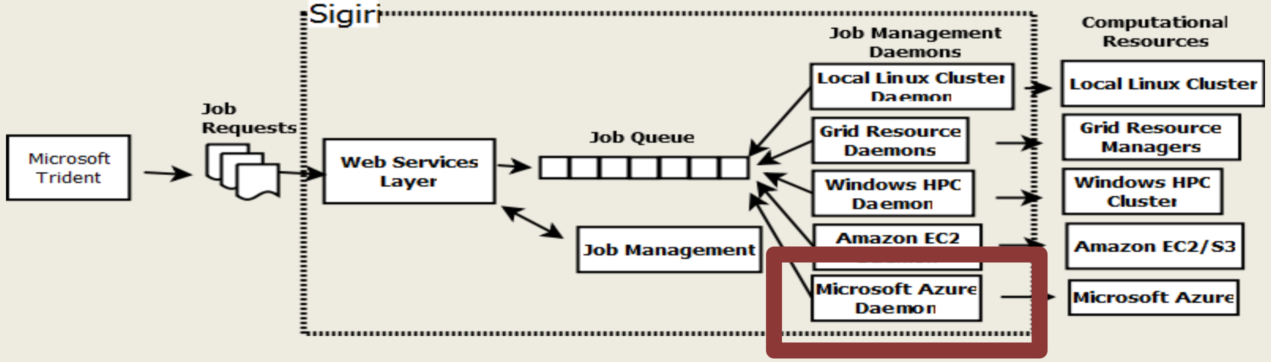
Enabling WRF Stack on Azure



Enabling WRF Stack on Azure



- Sigiri – Microsoft Azure Daemon
  - Maintains applications to virtual machine mappings
    - Can use an external service as well
  - Interacts with Windows Azure API to
    - Deploy and start hosted services
    - Handle Azure security credentials
    - Maintain Azure VM pools
    - Monitor job executions

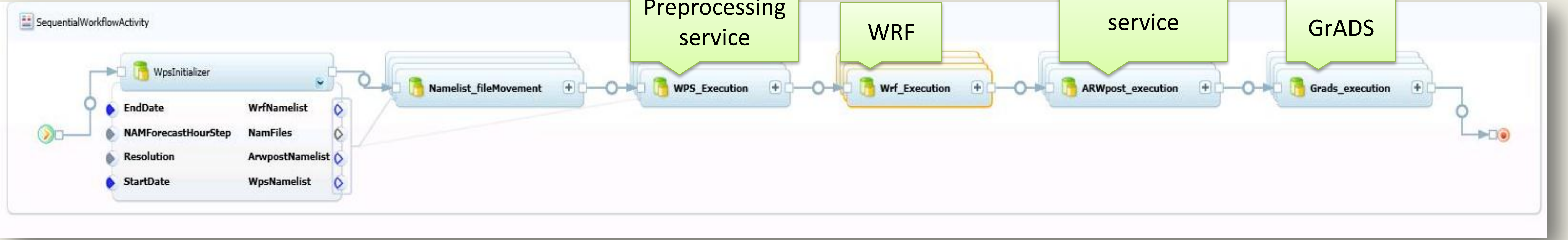


- Sigiri – Microsoft Azure Service
  - Deployed inside virtual machines
  - Accepts job submission requests from Microsoft Azure daemon
  - Launches jobs and monitors them
  - Enables status queries

Working with Azure VM Roles

- All the related applications are installed on a virtual machine using Hyper-V
  - Virtual machine image sizes are limited 35GB to enable a wide variety of instance types
- Custom virtual machines (VHD files) are uploaded and stored in Azure Blob Store
  - Managed by azure command line tools
- Custom hosted service deployments are needed to start VM roles with custom VM images
  - Dynamic configuration of service descriptors to support service requirements
    - Configuration of VHD files, number of instances, certificate associations
- Hosted services are deployed and started on-demand using Azure management API
- Sigiri Azure daemon
  - manages the interactions with Azure management API
  - Manages the life cycle of virtual machines
- Light-weight Sigiri service within the started VM role instances acts as job managers
- Azure blob store is used for all data transfers to and from virtual machines

WRF Workflow

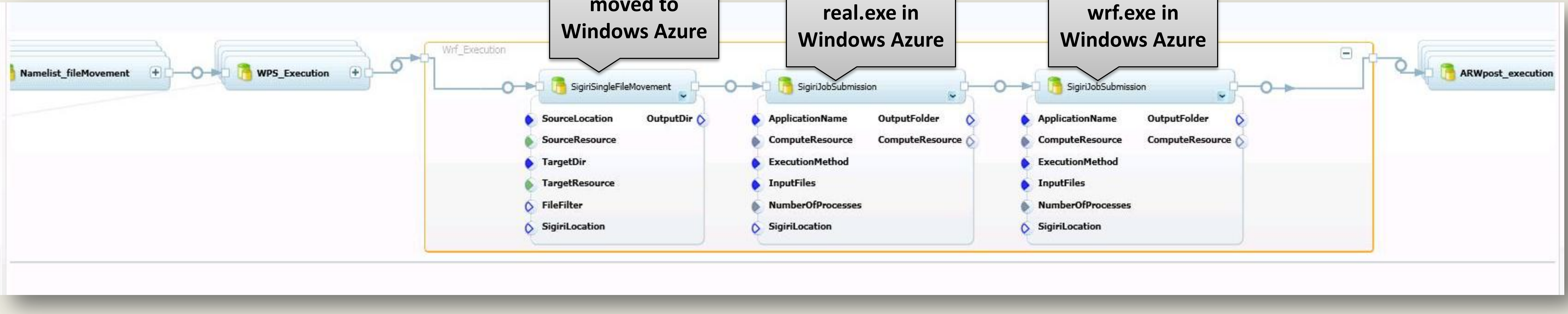


WRF Preprocessing service

WRF

ARWPost service

GrADS

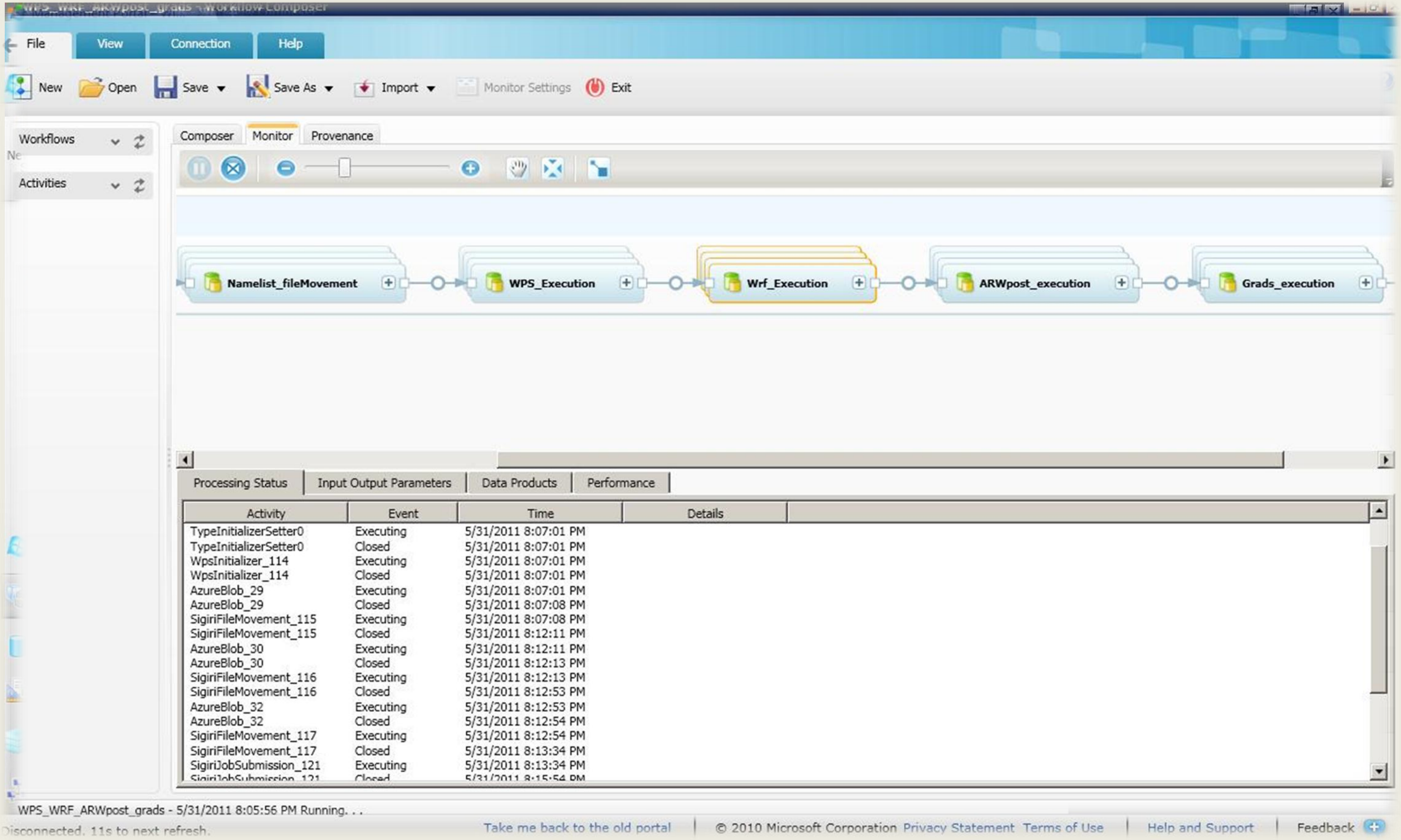


Input files are moved to Windows Azure

Execution of real.exe in Windows Azure

Execution of wrf.exe in Windows Azure

WRF Job Execution in Windows Azure using Microsoft Trident



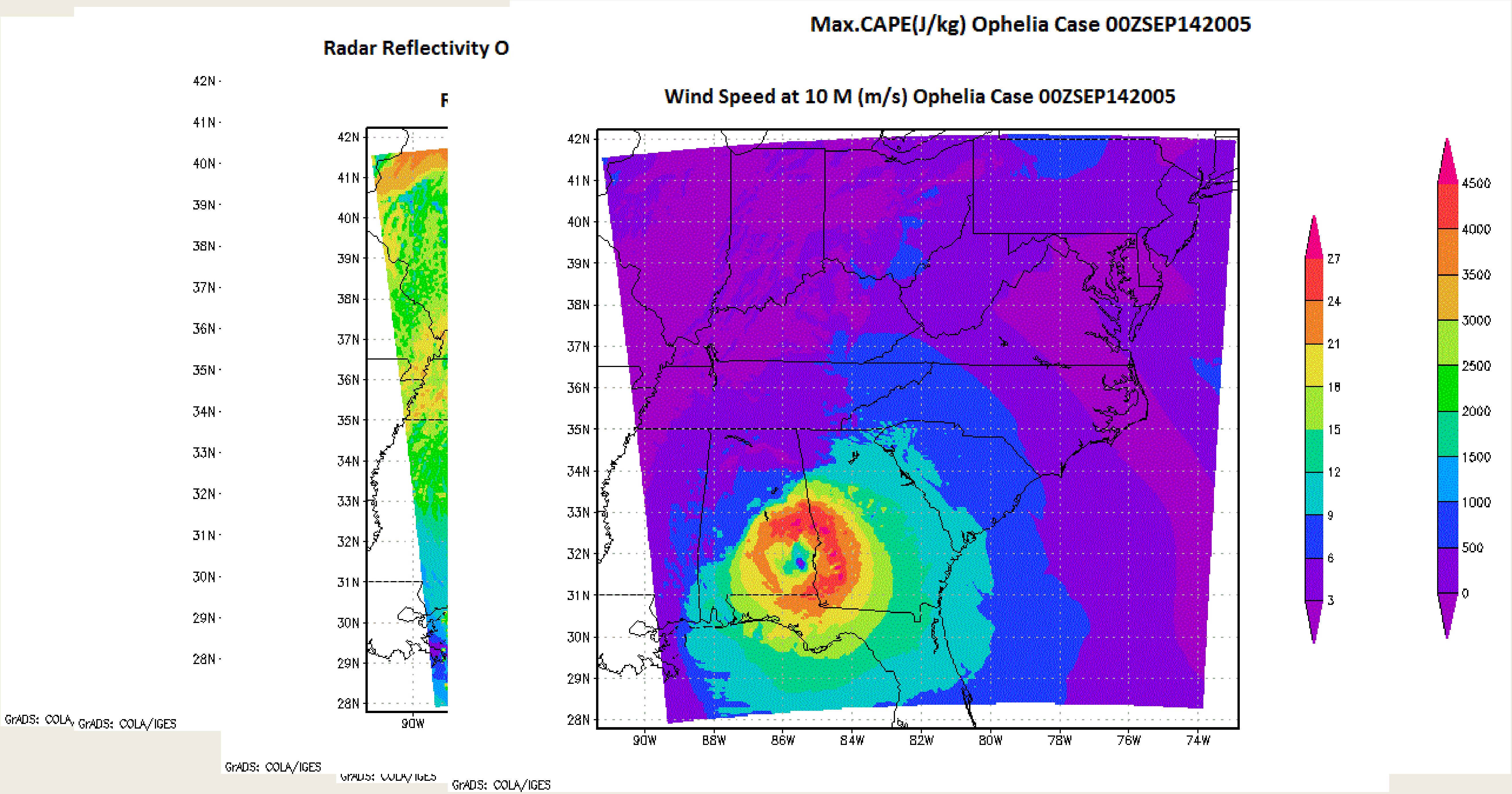
Activity	Event	Time	Details
TypeInitializerSetter0	Executing	5/31/2011 8:07:01 PM	
TypeInitializerSetter0	Closed	5/31/2011 8:07:01 PM	
WpsInitializer_114	Executing	5/31/2011 8:07:01 PM	
WpsInitializer_114	Closed	5/31/2011 8:07:01 PM	
AzureBlob_29	Executing	5/31/2011 8:07:01 PM	
AzureBlob_29	Closed	5/31/2011 8:07:08 PM	
SigiriFileMovement_115	Executing	5/31/2011 8:07:08 PM	
SigiriFileMovement_115	Closed	5/31/2011 8:12:11 PM	
AzureBlob_30	Executing	5/31/2011 8:12:11 PM	
AzureBlob_30	Closed	5/31/2011 8:12:13 PM	
SigiriFileMovement_116	Executing	5/31/2011 8:12:13 PM	
SigiriFileMovement_116	Closed	5/31/2011 8:12:53 PM	
AzureBlob_32	Executing	5/31/2011 8:12:53 PM	
AzureBlob_32	Closed	5/31/2011 8:12:54 PM	
SigiriFileMovement_117	Executing	5/31/2011 8:12:54 PM	
SigiriFileMovement_117	Closed	5/31/2011 8:13:34 PM	
SigiriJobSubmission_121	Executing	5/31/2011 8:13:34 PM	
SigiriJobSubmission_121	Closed	5/31/2011 8:15:54 PM	

WPS_WRF_ARWpost_grads - 5/31/2011 8:05:56 PM Running. . .

isconnected. 11s to next refresh.

Take me back to the old portal | © 2010 Microsoft Corporation Privacy Statement Terms of Use | Help and Support | Feedback

Test Run with Ophelia (14-Sep-2005)



Max.CAPE(J/kg) Ophelia Case 00ZSEP142005

Radar Reflectivity O

Wind Speed at 10 M (m/s) Ophelia Case 00ZSEP142005

Using Windows Azure for WRF Executions: Concerns and Experiences

- Default environment has no support for MPI executions
- Limitations of MPI on Azure
  - Limited to single node, shared memory execution
  - Only small scale experiments are possible within a single node
- Execution of Linux binaries are limited to the capabilities of platform emulators (cygwin)
- Windows Azure VM roles are in beta stage
  - Debugging is hard
  - Support
  - VM creation has about 20 to 30 steps (Marty Humprey “Cloud, HPC or Hybrid: A Case Study Involving Satellite Image Processing”)
- Windows management API is not well documented
  - Certain semantics of the API related to VM roles is ambiguous
- Increased startup overheads of virtual machines

Agenda

- Geo-Science Applications: Challenges and Opportunities
- Research Vision
- Proposed Framework
- Applications
  - Scheduling time-critical MPI applications in Windows Azure
  - Scheduling large number of small jobs (ensembles) in Windows Azure

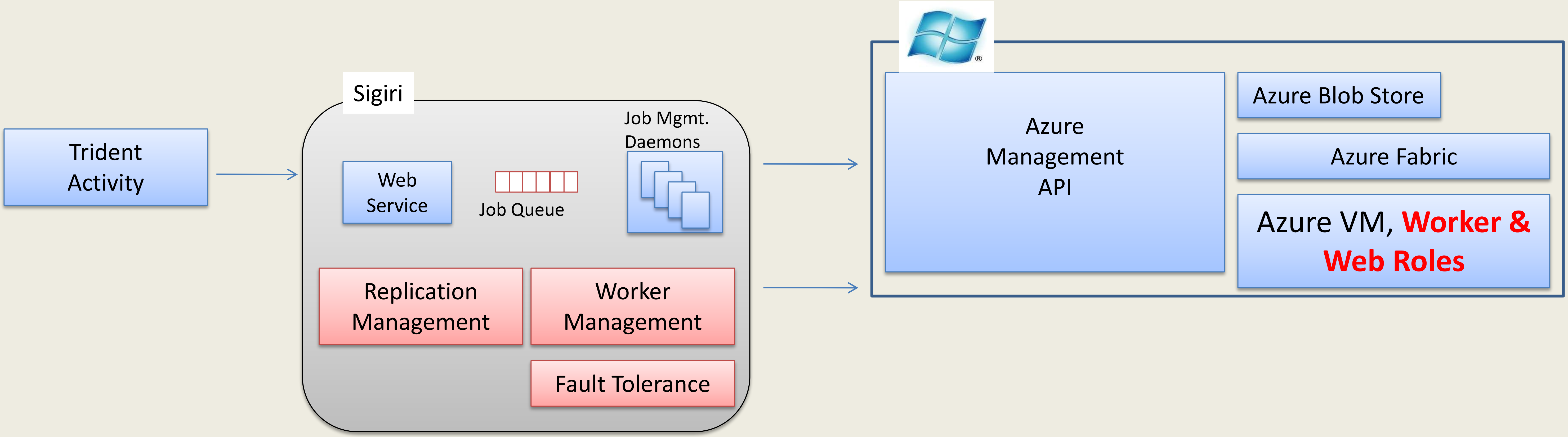
Towards Enabling Ensemble Runs in Geo-Science

- Search for a proper use of worker roles for geo-science applications
- Sample Application
  - enables the study of change in the strength and impact of storms that start over the oceans
  - has given access to and manipulation of climate model scenarios for emergency management and personnel and local government officials
  - Typical simulation only takes a few minutes to run on a medium-sized workstation (Input: 3GB, Output: 8GB)
  - Complete experiment sweeps both temporal and spatial parameters
    - Air layers at different heights over a period of time
    - About 14000 – 15000 jobs per experiment and then aggregation of data
- Research Focus
  - Fault tolerant ensemble execution using Azure worker roles
  - Management of large number of workers
  - Orchestrated through Trident
  - Downstream workflow management of the data results

Towards Enabling Ensemble Runs in Geo-Science

- Framework Extensions
  - Management of large number of job submissions and their life cycles
    - Optimal allocation of workers for jobs
    - Using a combination of worker and VM roles
  - Fault tolerance
    - Replication of stragglers and failing jobs
  - Management of data movements
  - Management of resources before and after job executions
- Making experiment outputs available for scientists and interested parties
  - Data catalogues
  - Meta-data management

Extensions to the Framework to Enable Ensemble Runs



Summary

- Geo-Science Applications: Challenges and Opportunities
- Research Vision
- Proposed Framework
- Applications
  - Scheduling time-critical MPI applications in Windows Azure
  - Scheduling large number of small jobs (ensembles) in Windows Azure

Further Information

- Please visit our website: http://pti.iu.edu/d2i/leadII-home
- LEAD II and Vortex2 video: http://pti.iu.edu/video/vortex2
- Contact us
  - Eran Chinthaka Withana (echintha@cs.indiana.edu)
  - Beth Plale (plale@cs.indiana.edu)

Team Members: Indiana University (lead) University of Miami, University of Oklahoma

Questions … ??

- Further Information
  - Please visit our website: http://pti.iu.edu/d2i/leadII-home
  - LEAD II and Vortex2 video: http://pti.iu.edu/video/vortex2
  - Contact us
    - Eran Chinthaka Withana (echintha@cs.indiana.edu)
    - Beth Plale (plale@cs.indiana.edu)