Microsoft® Research
Faculty Summit 2010

Guarujá, Brasil | May 12 – 14 | In collaboration with FAPESP

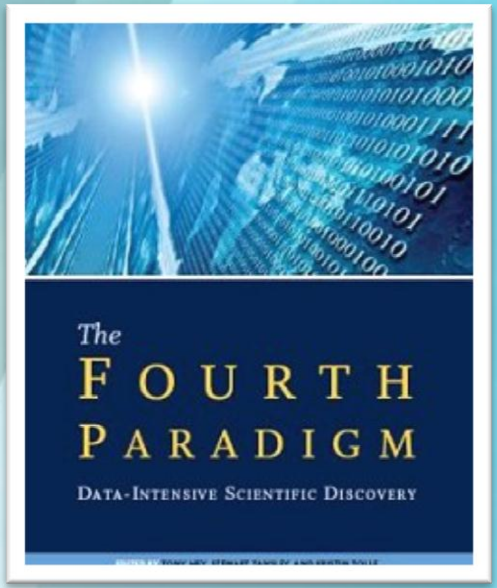
Microsoft® Research
Faculty Summit 2010

Guarujá, Brasil | May 12 – 14 | In collaboration with FAPESP

Bridging the Gaps:
Satellites to Science and Desktop to the Cloud

Catharine van Ingen
Partner Architect
eScience Group, Microsoft Research

The Data Flood:
Ecological Science and the 4^{th} Paradigm

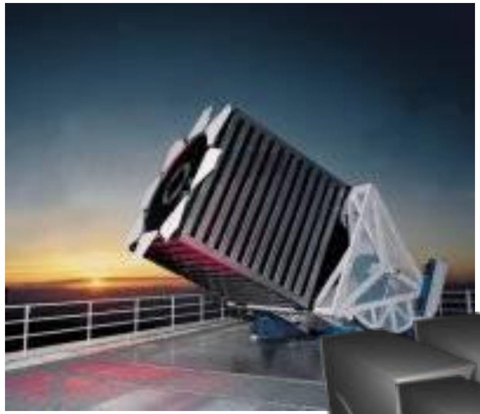


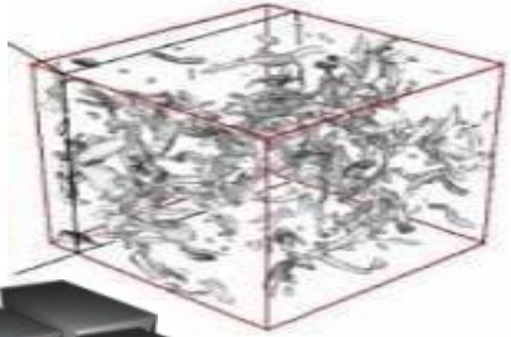
Small keys open big doors
Turkish Proverb

Emergence of a Fourth Paradigm

- Thousand years ago – **Experimental Science**
  - Description of natural phenomena
- Last few hundred years – **Theoretical Science**
  - Newton's Laws, Maxwell's Equations…
- Last few decades – **Computational Science**
  - Simulation of complex phenomena
- Today – **Data-Intensive Science**
  - Scientists overwhelmed with data sets from many different sources
    - Data captured by instruments
    - Data generated by simulations
    - Data generated by sensor networks
- eScience is the set of tools and technologies to support data federation and collaboration
  - For analysis and data mining
  - For data visualization and exploration
  - For scholarly communication and dissemination

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

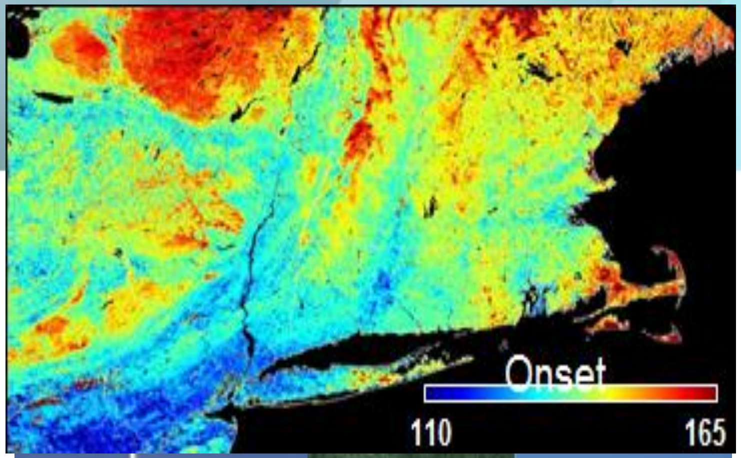




Jim Gray 2007

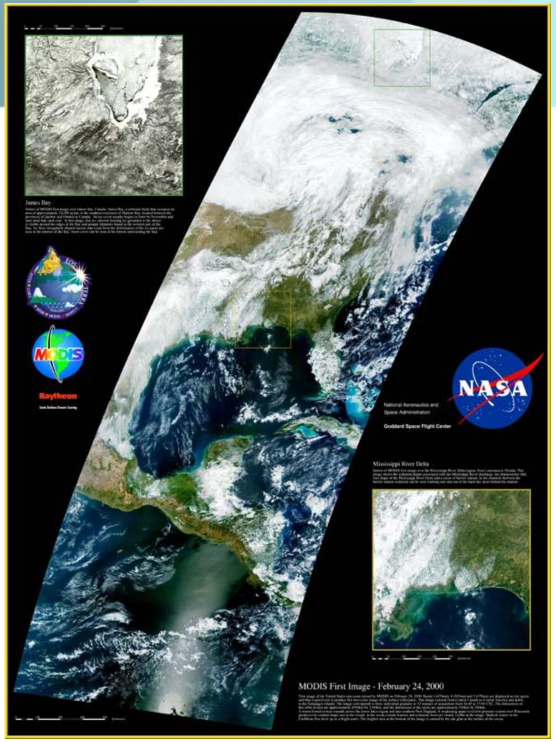
The Ecological Data Flood

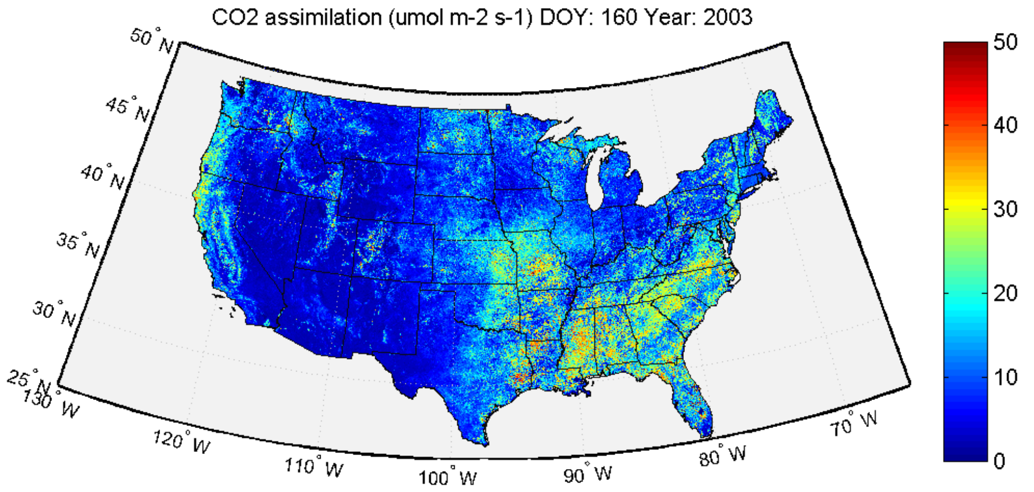
- We're living in a perfect storm of remote sensing, cheap ground-based sensors, internet data access, and commodity computing
- Yet deriving and extracting the variables needed for science remains problematic
  - Specialized knowledge for algorithms, internal file formats, data cleaning, etc, etc
  - Finding the right needle across the distributed heterogeneous and very rapidly growing haystacks



Environmental Remote Sensing Data

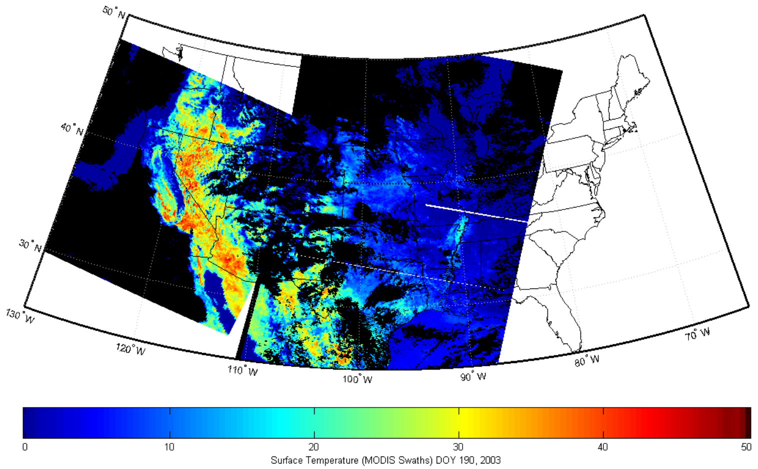
- Time series raster data
  - Over some period of time at some time frequency at some spatial granularity over some spatial area
  - Conversion from L0 data to L2 and beyond as well as reprojections still require specialized skills
  - Similar, but dirtier, than model output
- Can be “cut out” to create virtual sensors
- Today: PBs (L0) to TBs (L2+)



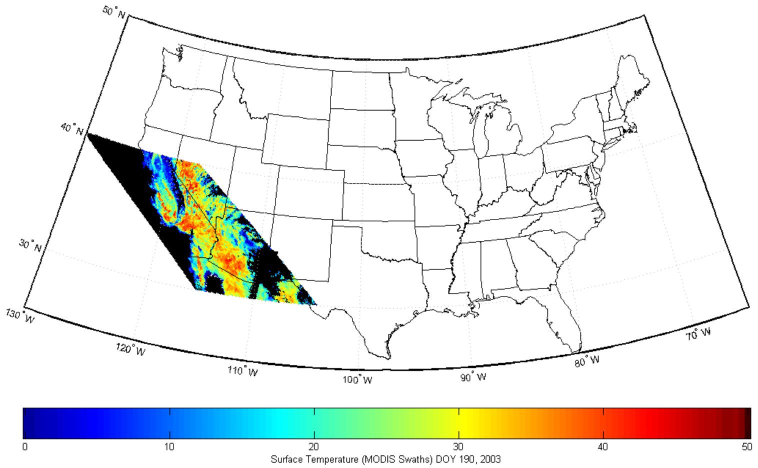


Tiling : Do Scientists Have to be Computer Scientists?

- Reprojection
  - Converts one geo-spatial representation to another.
  - Example is converting from latitude-longitude swaths to sinusoidal cells.
- Spatial resampling
  - Converts one spatial resolution to another.
  - Example is converting from 1 KM to 5 KB pixels.
- Temporal resampling
  - Converts one temporal resolution to another.
  - Example is converting from daily observation to 8 day averages.
- Gap filling
  - Assigns values to pixels without data either due to inherent data issues such as clouds or missing pixels introduced by one of the above.
- Masking
  - Eliminates uninteresting or unneeded pixels.
  - Examples are eliminating pixels over the ocean when computing a land product or eliminating pixels outside a spatial feature such as a watershed.



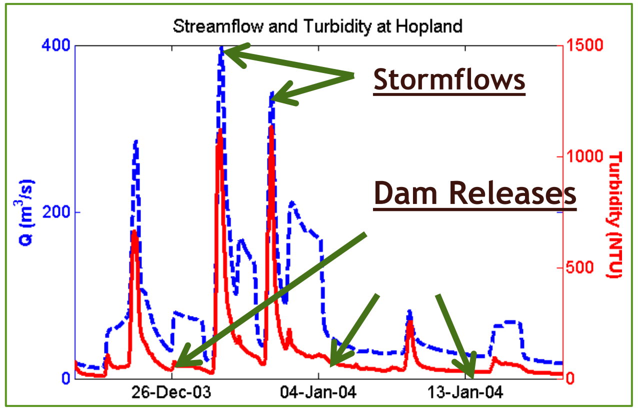
Source Data (Swath format)

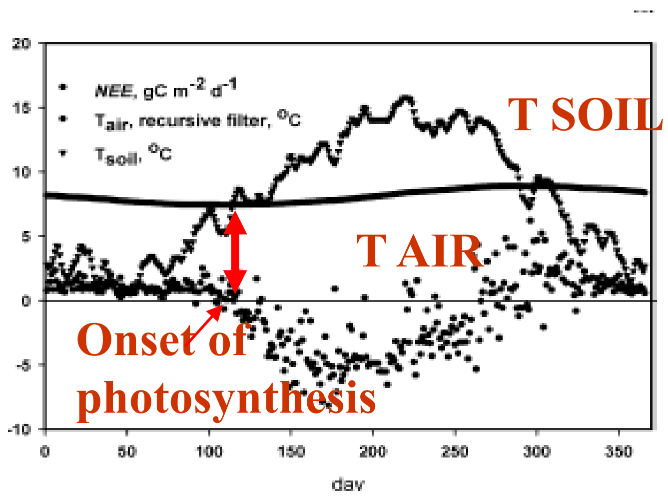


Reprojected Data (Sinusoidal format)

Environmental Sensor Data

- Time series data
  - Over some period of time at some time frequency at some spatial location.
  - May be actual measurement (L0) or derived quantities (L1+)
- (Re)calibrations, gaps and errors are a way of life.
  - Birds poop, batteries die, sensors fail.
  - Various quality assessment and signal correction algorithms.
  - Gap filling algorithms key as regular time series enable more analyzes
- Today: GBs to TBs

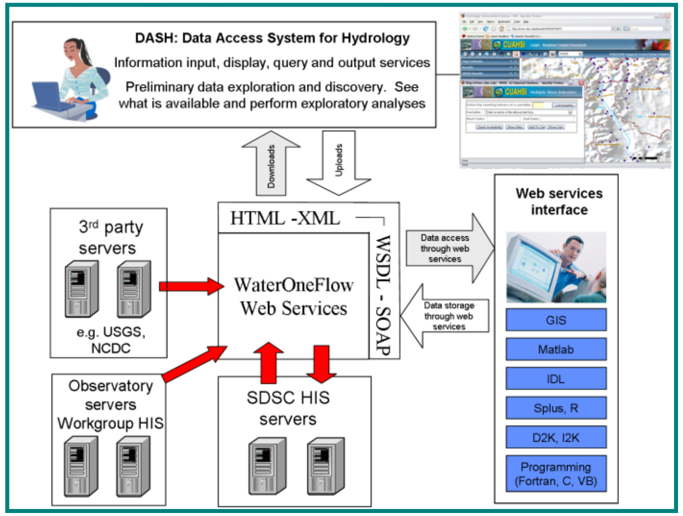


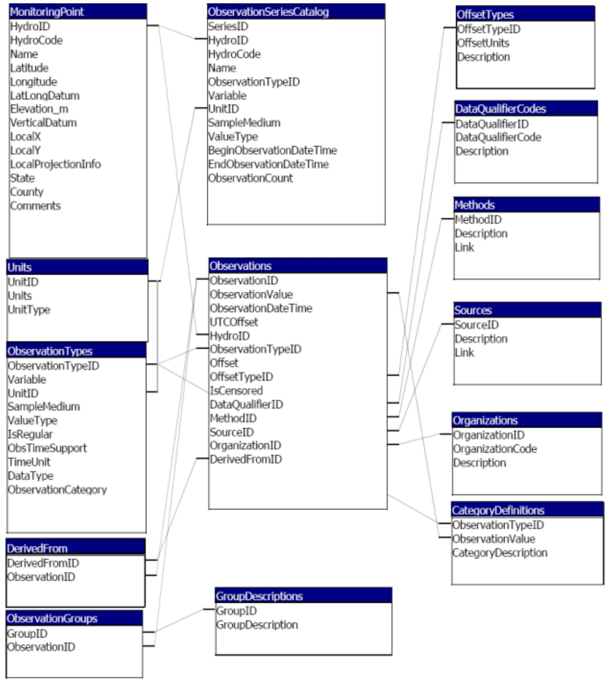


"Time is not just another axis"

Sensor Databases and Web Services

- Emerging trend is that groups use databases and web services to access, curate, and republish sensor data
- Most use a mostly normalized schema with the data in the center, but moving to putting the series catalog in the center
- Example is CUAHSI ODM
  - Initially to address internet access of US agency data – too hard to find, too hard to download all the data, too hard to get "just the new data"
  - Included water quality bottle samples, a notion of data revisions
  - 11 initial research sites growing over time





http://bwc.berkeley.edu   http://www.cuahsi.org

Environmental Ancillary Data

- Almost everything else!
  - ‘Constants’ such as latitude or longitude
  - Intermittent measurements such as grain size distributions or fish counts
  - Anecdotal descriptions such as “ripple” or “shaded”
  - Events such as algal blooms or leaf fall including those derived from sensor data such as “flood”
  - Disturbances such as a fire, harvest, landslide
- Not metadata such as instrument type, derivation algorithm, etc.
- Today: KBs to maybe GBs.

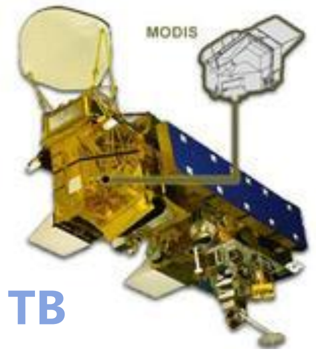
Ancillary Data is Different !

- Very hard won
  - Dig a pit or shoot an air rifle to get samples
  - Lab costs can be considerable
  - Gleaning from literature (and cross checking!)
- Very hard to curate
  - FLUXNET collection is currently ~30K numbers.
  - Often passed around in email and cut/pasted from web sites
- Very different usage patterns
  - Constant location attributes or aliases
  - Time series via splines or step functions
  - Filters for sensor data: periods before or after, sites with summer LAI > x, etc
  - Time benders: “since <event>”
- Often requires science judgment
  - Different scientists don’t always agree
  - Anecdotal reporting difficult to interpret
  - Citizen science contributions give important coverage but at quality?

Why Make this Distinction?

- Provenance and trust widely varies
  - Data acquisition, early processing, and reporting ranges from a large government agency to individual scientists.
  - Smaller data often passed around in email; big data downloads can take days (if at all)
- Data sharing concerns and patterns vary
  - Open access followed by (non-repeatable and tedious) pre-processing
  - True science ready data set but concerns about misuse, misunderstanding particularly for hard won data.
- Computational tools differ.
  - Not everyone can get an account at a supercomputer center
  - Very large computations require engineering (error handling)
  - Space and time aren't always simple dimensions

PB

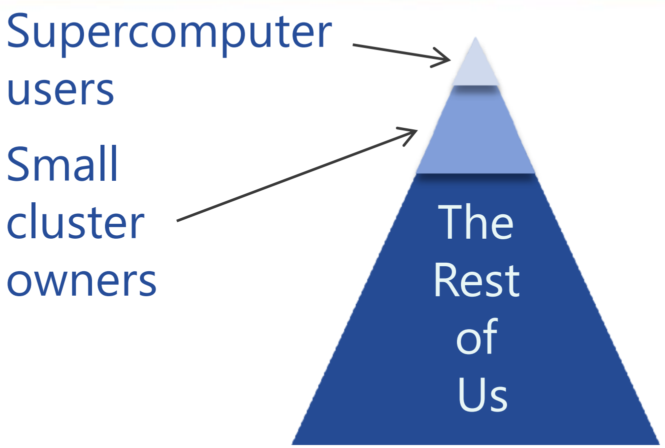


Complex shared detector | Simple instrument (if any)

Science happens when PBs, TBs, GBs, and KBs can be mashed up simply

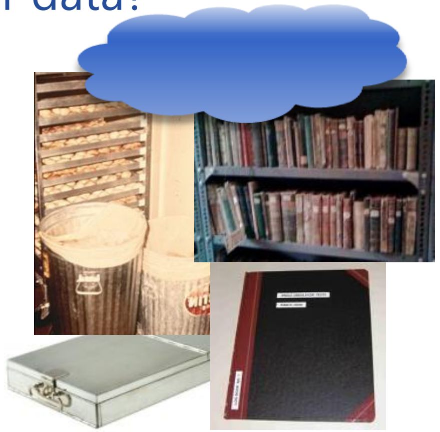
Complex and Heavy process by experts | Ad hoc observations and models

Bridging the Gap with the Cloud



- Barriers to Science:
  - Resource: compute, storage, networking, visualization capability
  - Complexity: specific cross-domain knowledge
  - Tedium: repetitive data gathering or preprocessing tasks
- With cloud computing, we can:
  - marshal needed storage and compute resources on demand without caring or knowing how that happens
  - access living curated datasets without having to find, educate, and reward a private data curator
  - run key common algorithms as Software as a Service without having to know the coding details or installing software
  - grow a given collaboration or share data and algorithms across science collaborations elastically

Where do you want your data?



Democratizing science analysis by fostering sharing and reuse

Azure and Cloud Computing

Ideas rose in clouds; I felt them collide until pairs interlocked, so to speak, making a stable combination.
Henri Poincare

The Cloud

- A model of computation and data storage based on "pay as you go" access to "unlimited" remote data center capabilities
- A cloud infrastructure provides a framework to manage scalable, reliable, on-demand access to applications
- A cloud is the "invisible" backend to many of our mobile applications
- Historical roots in today's Internet apps
  - Search, email, social networks
  - File storage (Live Mesh, Mobile Me, Flickr, …)

The Cloud Landscape

IaaS: Provide a data center and a way to host client VMs and data

PaaS: Provide a programming environment to build and manage the deployment of a cloud application

Saas: Delivery of software from the cloud to the desktop

Infrastructure as a Service

Platform as a Service

Software as a Service

Eucalyptus

flexiscale™

RIGHT SCALE™

RESERVOIR

IBM®

amazon web services™ EC2

GOGRID

Windows Azure

Google™ App Engine

cloudera

hadoop

Microsoft .NET Services

Microsoft SQL Services

Google™ Docs

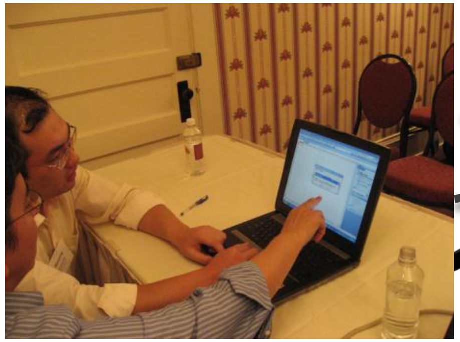
Microsoft Dynamics CRM Services

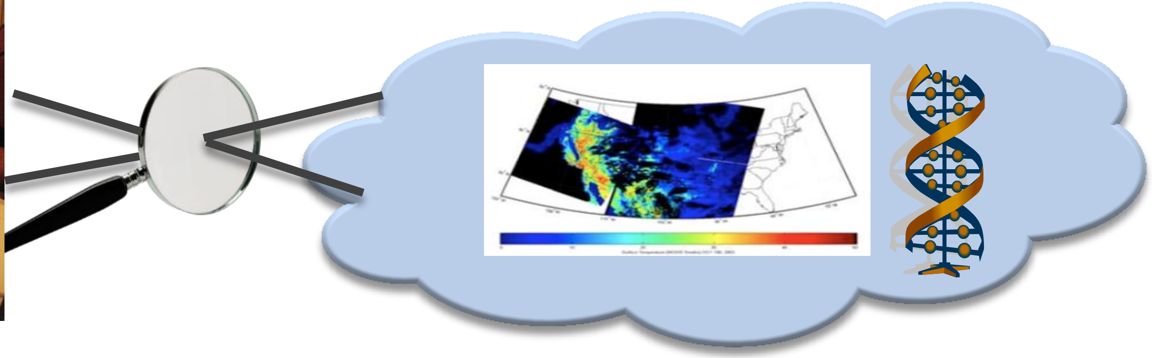
Microsoft SharePoint Services

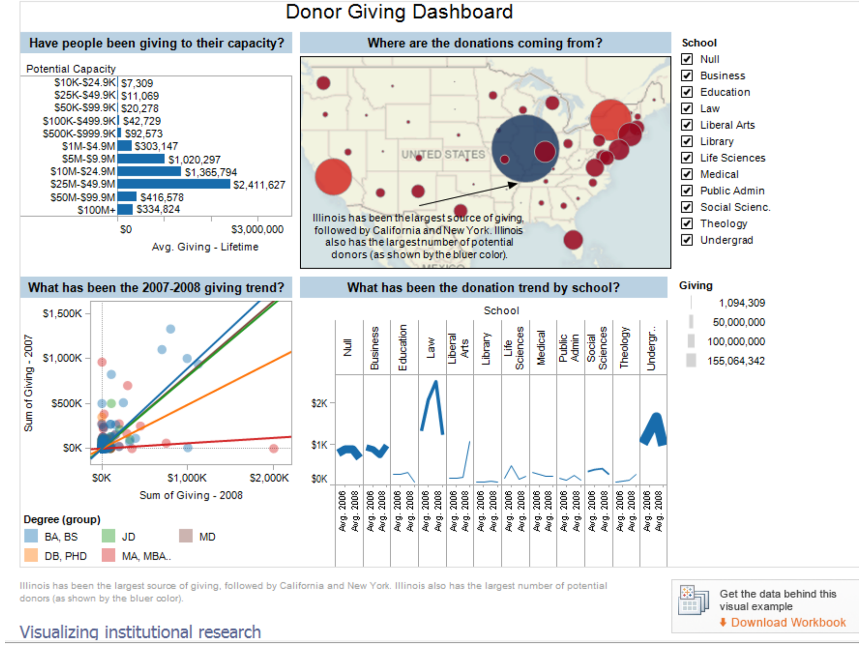
salesforce.com® Success On Demand.

Research Clients for A Cloud Research Platform

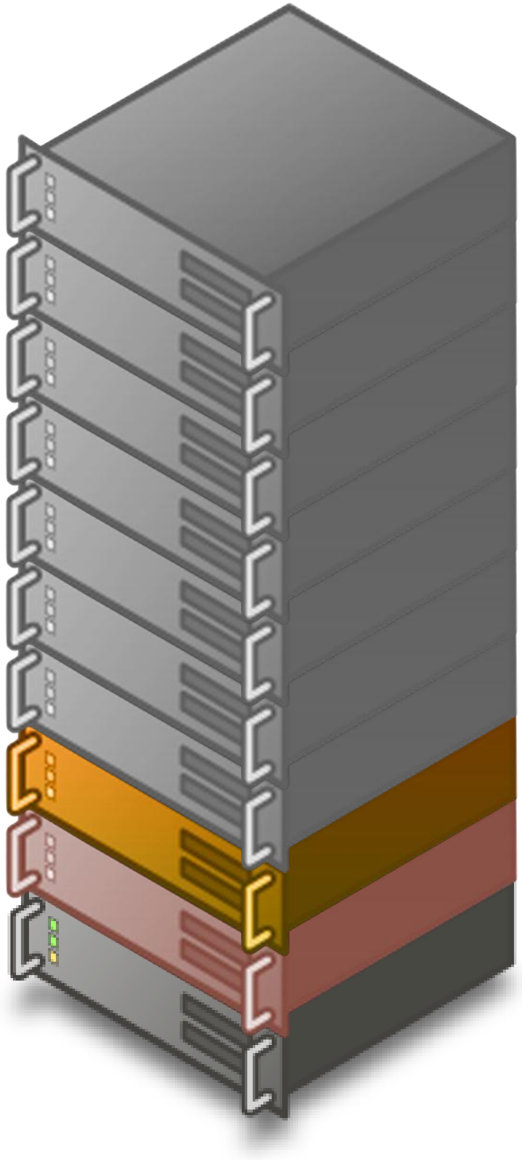
- Seamless interaction is crucial
  - Cloud is the lens that magnifies the power of desktop
  - Persist and share data from client in the cloud
  - Analyze data initially captured in client tools, such as Excel
    - Analysis as a service (SQL, Map-Reduce, R/MatLab).
    - Data visualization generated in the cloud, display on client
    - Provenance, collaboration, other core services…







Azure Configuration by the Fabric Controller (FC)



Each Guest VM has:

- 1-8 CPU cores: 1.5-1.7 GHz x64
- Memory: 1.7-14.2 GB
- Network: 100+ Mbps
- Local Storage: 500GB – 2 TB

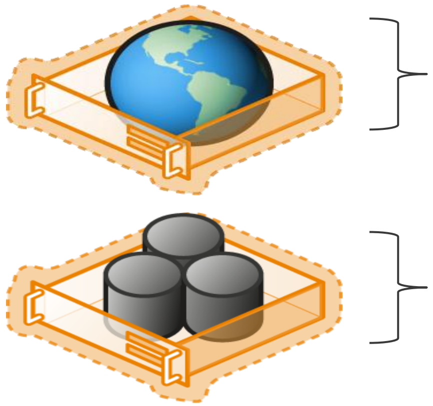
Configured with:

- .NET framework
- IIS 7.0
- 64-bit Windows Server 2008 Enterprise
- Azure platform

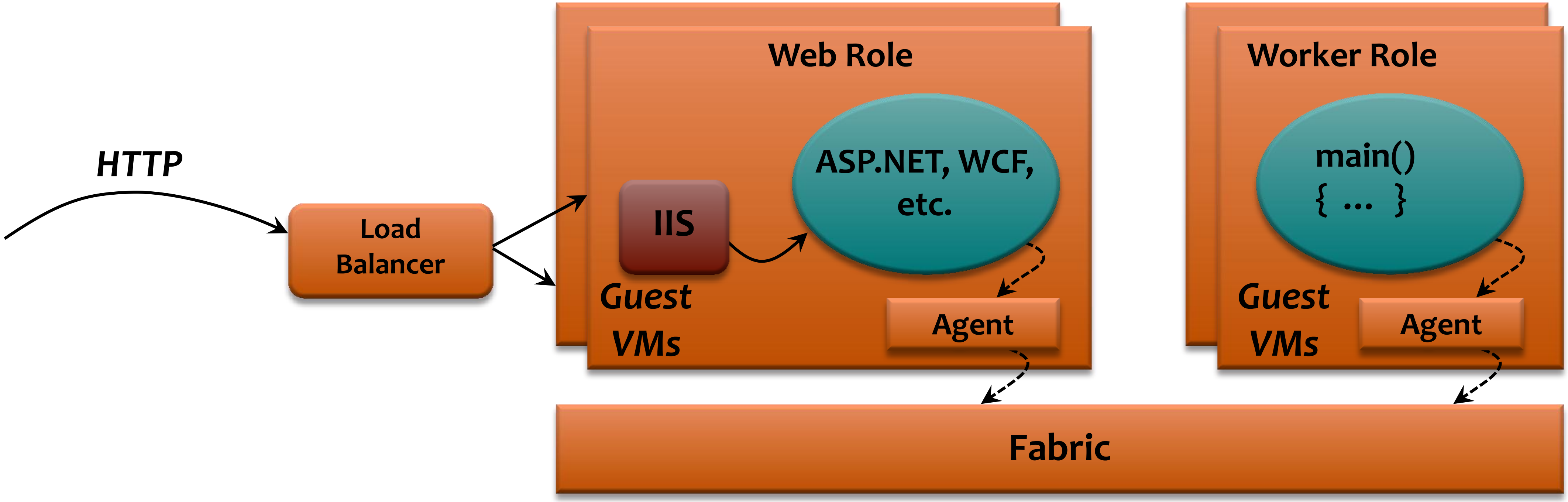
Microsoft® .NET Framework

Windows Server® Internet Information Services 7.0

Windows Server® 2008 Enterprise



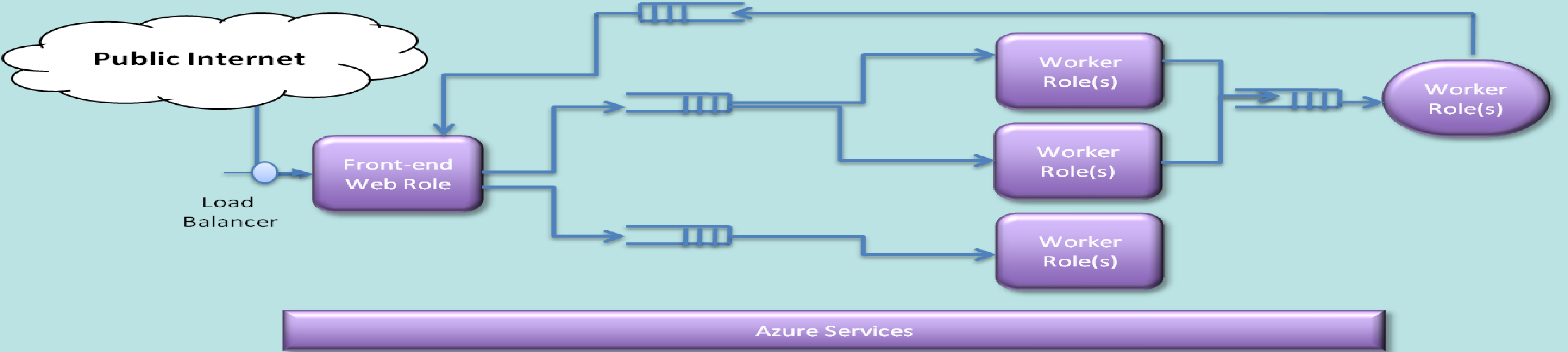
Windows Azure Compute Service



- Web Role provides client access web presence
- Worker Role does all heavy lifting
- Each can scale independently

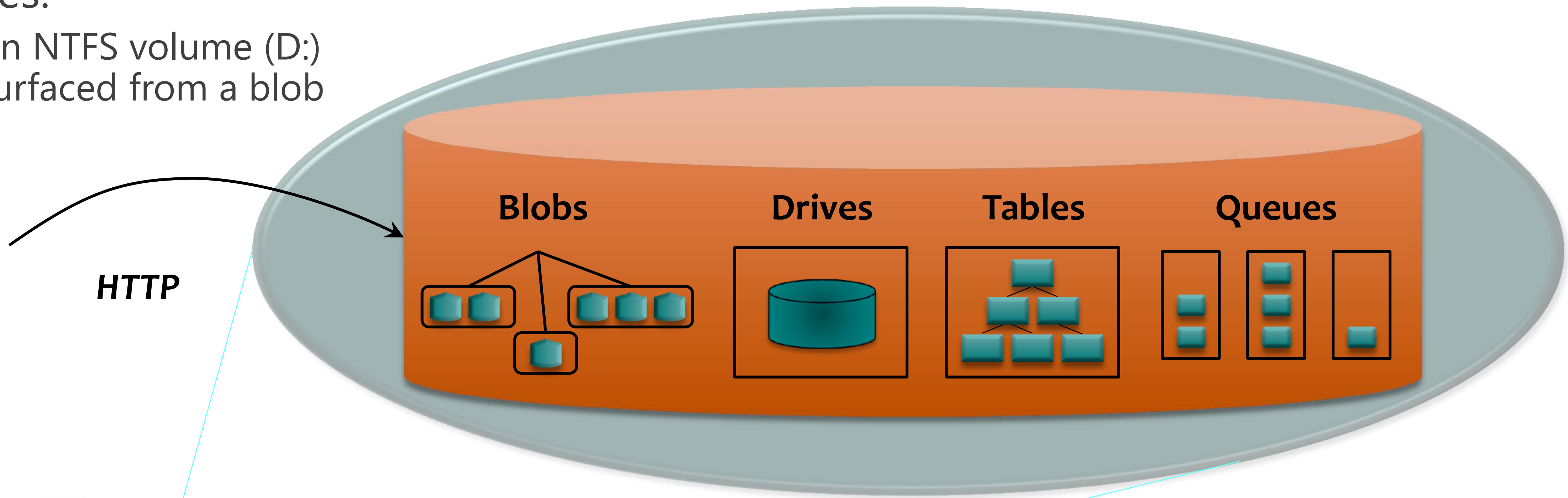
Scalable, Fault Tolerant Applications

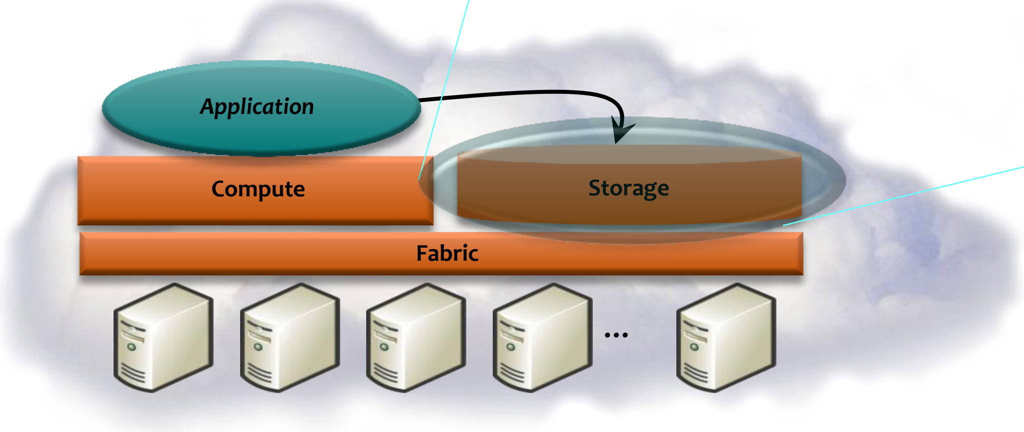
- Queues are the application glue for loosely coupled applications
  - Link application components, enabling each to scale independently
  - Resource allocation, different priority queues and backend servers
  - Mask faults in worker roles through reliable messaging and retries
- Use Inter-role communication for performance
  - TCP communication between role instances



Windows Azure Storage Service

- Drives:
  - An NTFS volume (D:) surfaced from a blob





- Blobs, Tables, and Queues:
  - are exposed via .NET and RESTful interfaces
  - can be accessed by Windows Azure apps, other cloud applications or non-cloud client applications

MODISAzure :
Computing Evapotranspiration (ET) in The Cloud

You never miss the water till the well has run dry
Irish Proverb

Computing ET From Historical Sensor Data

$$ET = P - R - \frac{dS}{dt}$$

Simple Water Balance

ET: Evapotranspiration or release of water to the atmosphere by evaporation from open water bodies and transpiration by plants

P: Precipitation including snowfall

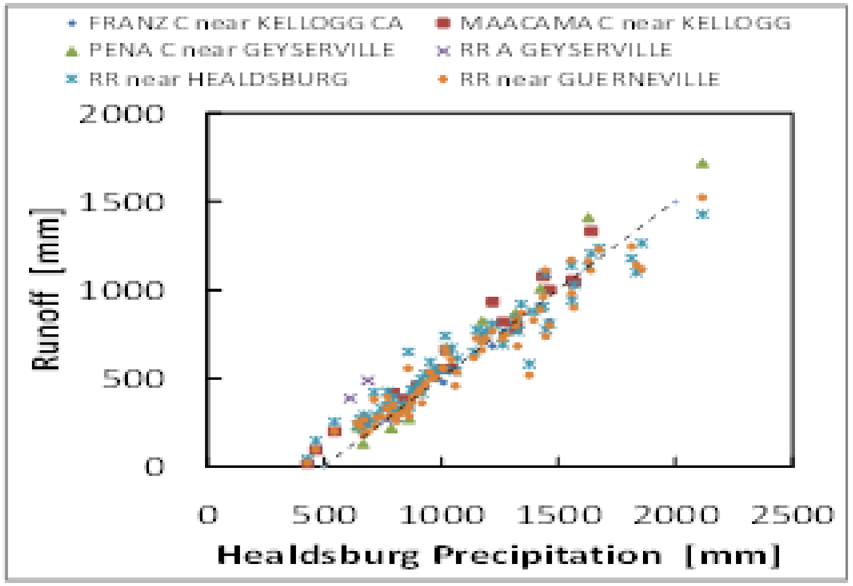
R: Surface runoff in streams and rivers

dS/dt: change in water storage over time such as increase in lakes or groundwater levels

P: http://www.ncdc.noaa.gov/oa/ncdc.html

R: http://waterdata.usgs.gov/nwis

- Easy to do (with a digital watershed)
- Long term trends only



In Mediterranean climates such as California, a long term equilibrium may exist. The ecosystem determines ET by soils and climate and the lowest recorded annual rainfall may determines vegetation.

~400 MB of data reduced to ~1KB

Computing ET from First Principles

$$ET = \frac{\Delta Rn + \rho_a c_p (\delta q) g_a}{(\Delta + \gamma(1 + g_a/g_s))\lambda_v}$$

ET = Water volume evapotranspired ($m^3\ s^{-1}\ m^{-2}$)

Δ = Rate of change of saturation specific humidity with air temperature.(Pa K^{-1})

λ_v = Latent heat of vaporization (J/g)

R_n = Net radiation (W m^{-2})

c_p = Specific heat capacity of air (J $kg^{-1}\ K^{-1}$)

ρ_a = dry air density (kg m^{-3})

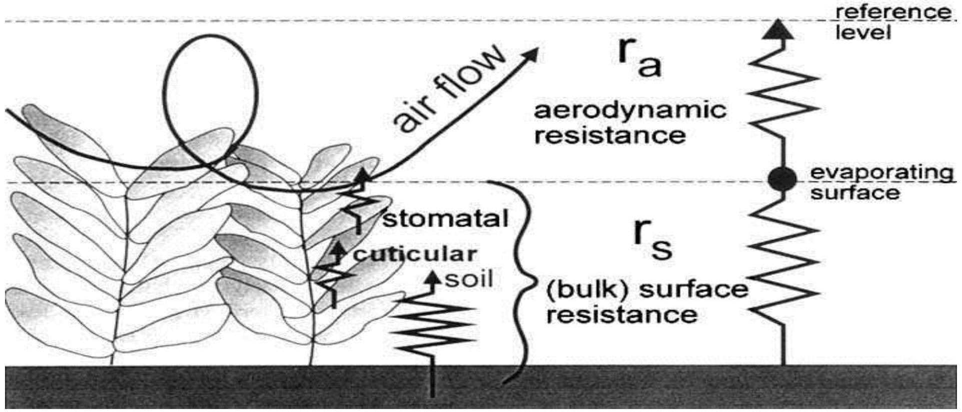
δq = vapor pressure deficit (Pa)

g_a = Conductivity of air (inverse of r_a) (m s^{-1})

g_s = Conductivity of plant stoma, air (inverse of r_s) (m s^{-1})

γ = Psychrometric constant ($\gamma \approx 66$ Pa K^{-1})

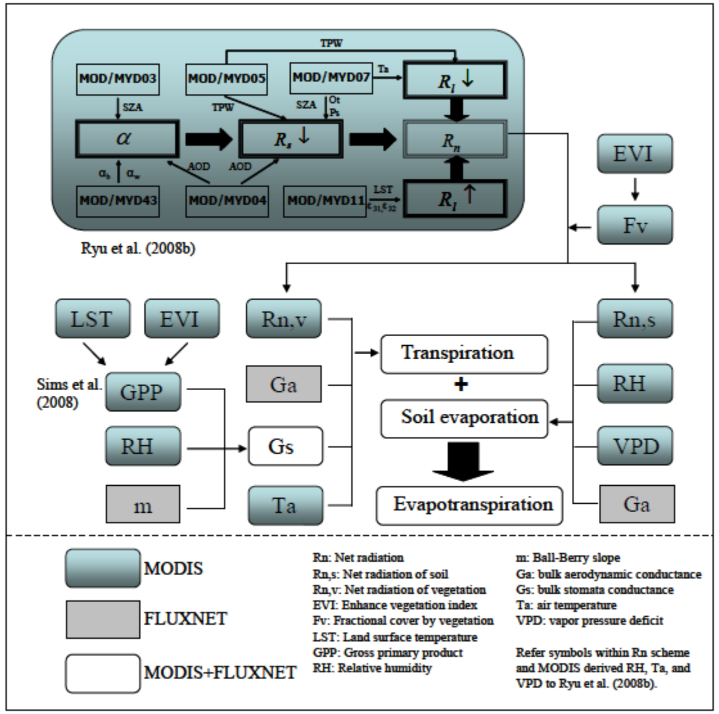
- Lots of inputs : big reduction
- Some of the inputs are not so simple



Estimating resistance/conductivity across a catchment can be tricky

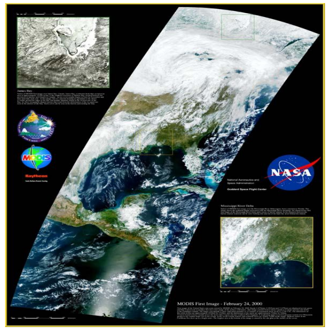
Computing ET from Imagery, Sensors and Field Data

- Modification of Penman-Monteith
  - Additions to handle for dry region leaf/air temperature differences, snow cover, leaf area fill, and temporal upscaling
  - All time value inputs (including meterology) from MODIS
  - Conductance from biome aggregate flux tower properties
  - Not a simple matrix computation due to above science needs
- Validation by comparison with flux tower data from 74 US towers (299 site years)



FLUXNET curated sensor dataset (30GB, 960 files)

NASA MODIS imagery source archives 5 TB (600K files)



FLUXNET curated field dataset 2 KB (1 file)

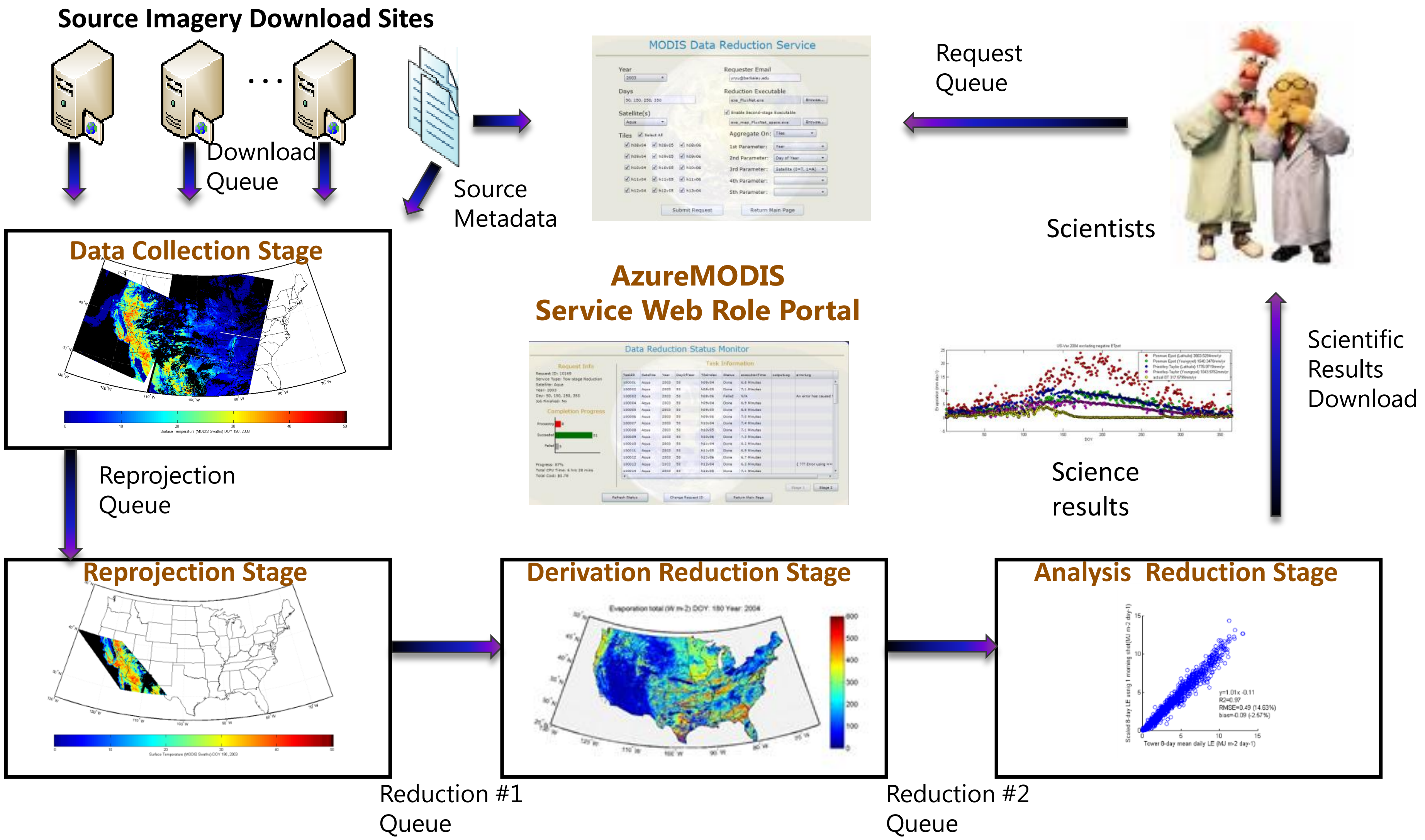
MODISAzure: Four Stage Image Processing Pipeline

Data collection stage
- Downloads requested input tiles from NASA ftp sites
- Includes geospatial lookup for non-sinusoidal tiles that will contribute to a reprojected sinusoidal tile

Reprojection stage
- Converts source tile(s) to intermediate result sinusoidal tiles
- Simple nearest neighbor or spline algorithms

Derivation reduction stage
- First stage visible to scientist
- Computes ET in our initial use

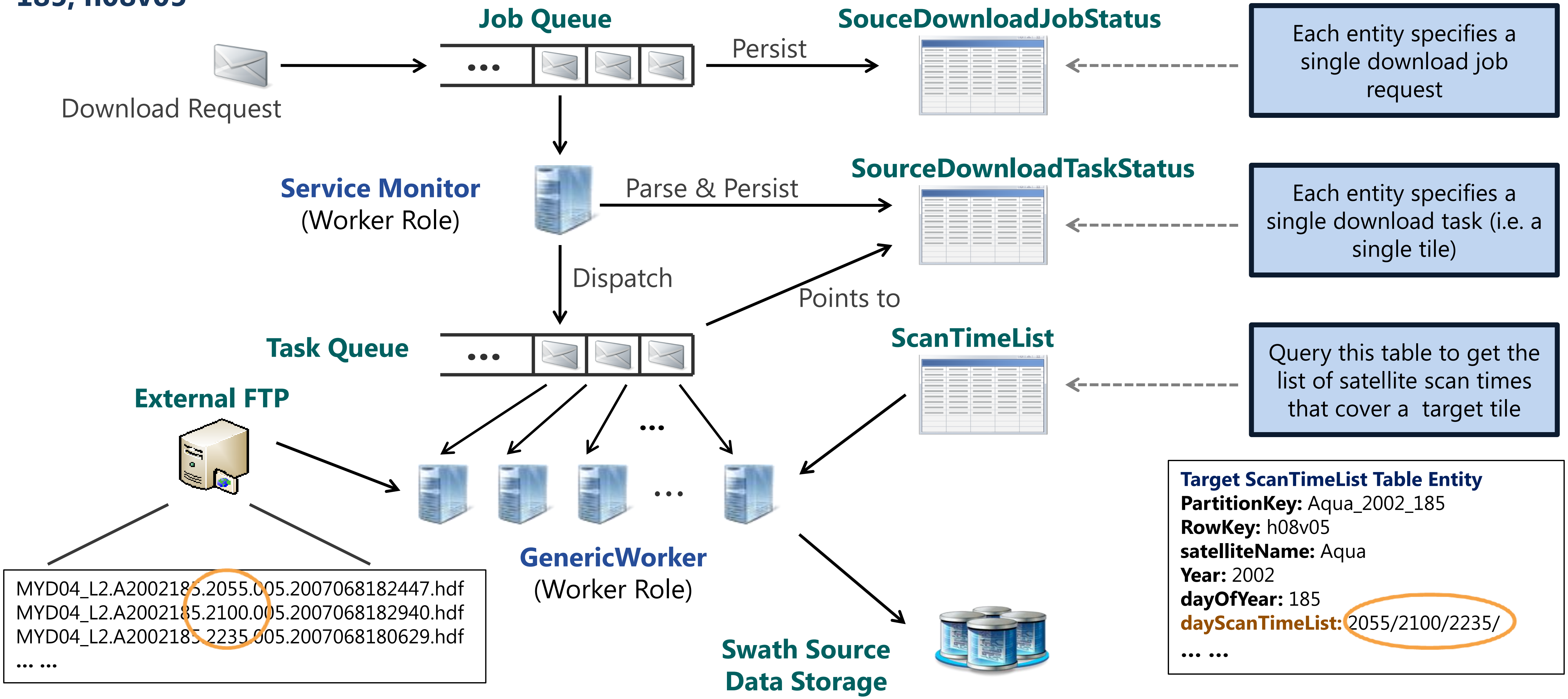
Analysis reduction stage
- Optional second stage visible to scientist
- Enables production of science analysis artifacts such as maps, tables, virtual sensors



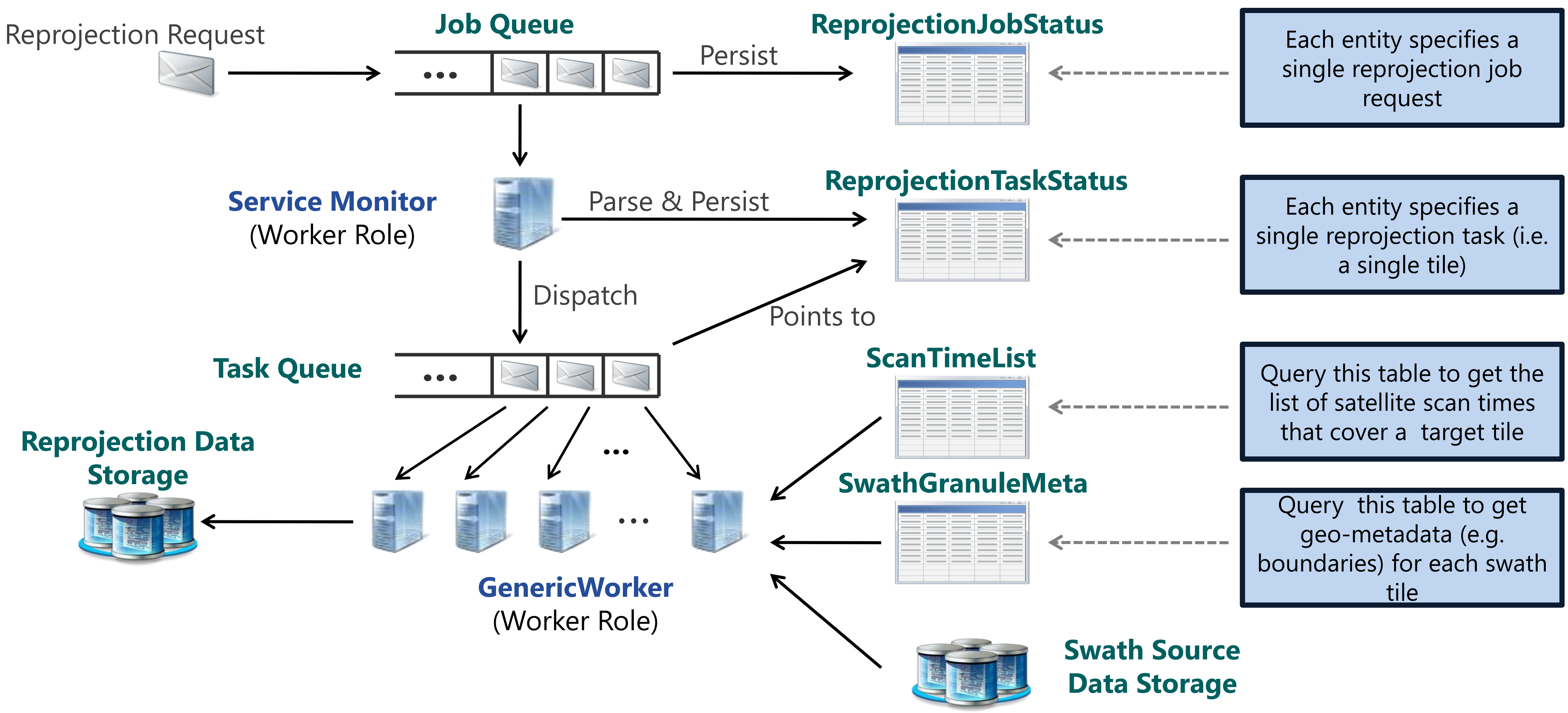
http://research.microsoft.com/en-us/projects/azure/azuremodis.aspx

Source Data Download Service

Example: Download the required source files for reprojecting the target sinusoidal tile: **MYD04_L2, Year 2002, Day 185, h08v05**

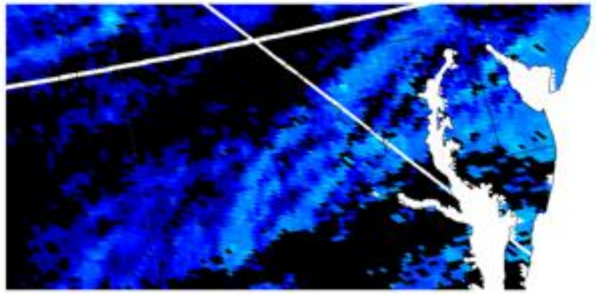


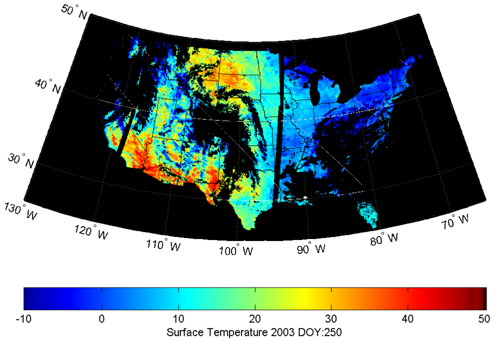
Reprojection Service

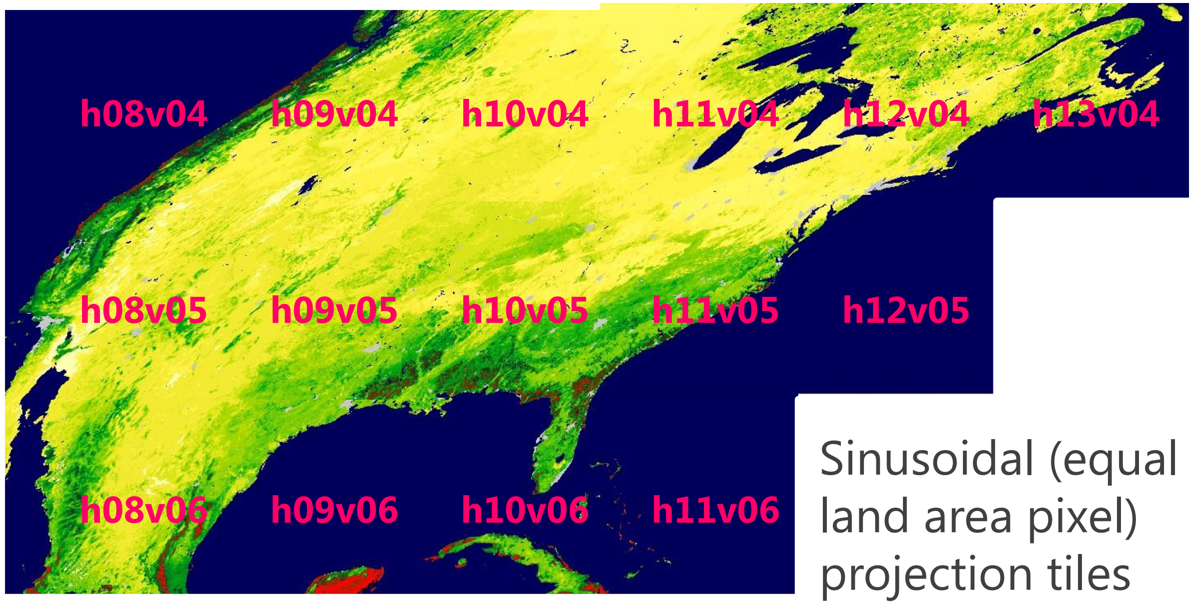


Why is Reprojection Tricky?

- It's not just nearest neighbor vs aggregating spline and nadir vs oblique pixels



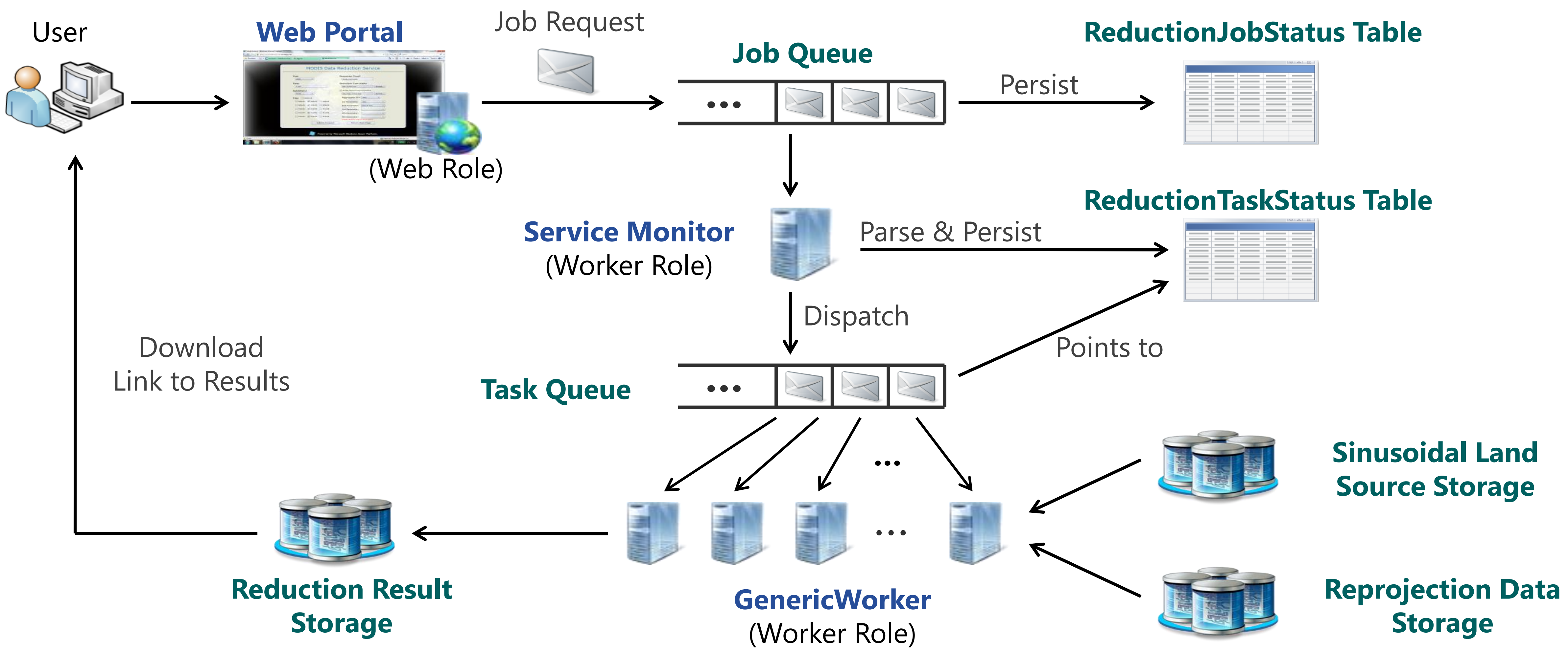




Sinusoidal (equal land area pixel) projection tiles across the US

- Black pixels have no data
  - Non-US land surface masked
  - Vertical bands are gaps between swath tiles; these can be filled by spatial spline or other fit
  - Clouds cause gaps in surface measurement; these can be filled by temporal fit or model result leveraging variables in other products
- White lines have no data
  - Unable to find nearest neighbor at edges of sinusoidal tiles; either due to quality+gap or programming algorithm bug
- Processing only the layers of interest makes dramatic savings in compute and storage

Reduction Service (Single Stage Only)

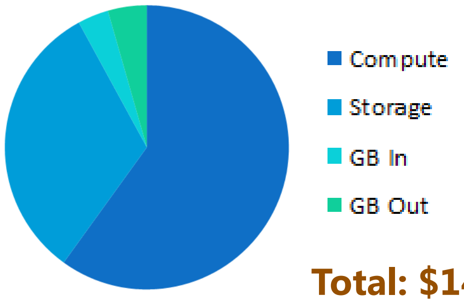


Pipeline Stage Priorities and Interactions

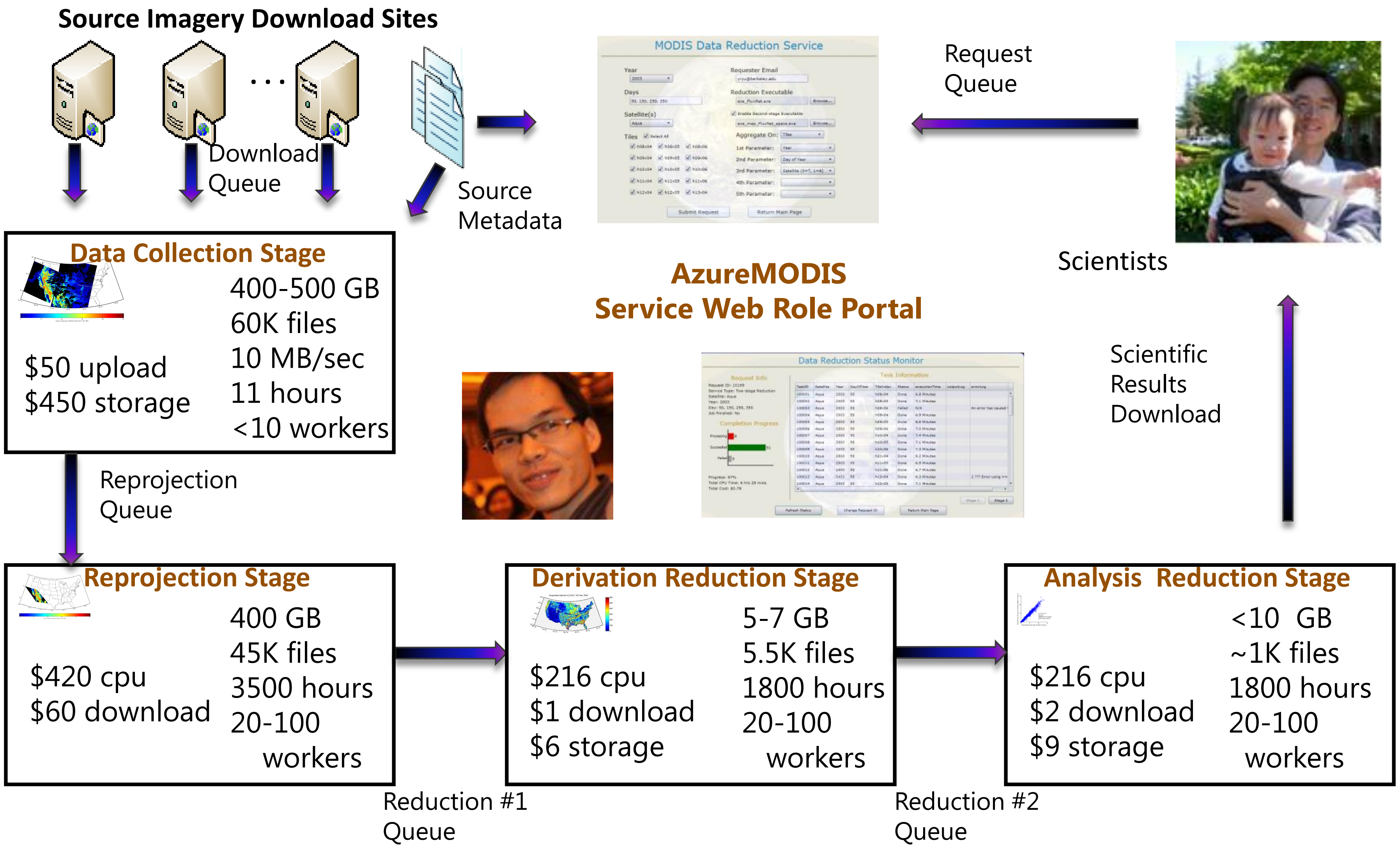
- The Web Portal Role, Service Monitor Role and 5 Generic Worker Roles are deployed at most times
  - 5 Generic Workers are sufficient for reduction algorithm testing and development ($20/day)
  - Early results returned to scientist while deploying up to 93 additional Generic Workers; such a deployment typically takes 45 minutes
  - Deployment taken down when long periods of idle time are known
  - Heuristic for scaling number of Generic Workers up and down
- Download stage runs in the deep background in all deployed generic worker roles
  - IO, not CPU bound so no competition
- Reduction tasks that have available inputs run preferentially to Reprojection tasks
  - Expedites interactive science result generation
  - If no available inputs and a backlog of reprojection tasks, number of Generic Workers scale up naturally until backlog addressed and reduction can continue
  - Second stage reduction runs only after all first stage reductions have completed

Costs for 1 US Year ET Computation

- Computational costs driven by data scale and need to run reduction multiple times
- Storage costs driven by data scale and 6 month project duration
- Small with respect to the people costs even at graduate student rates !

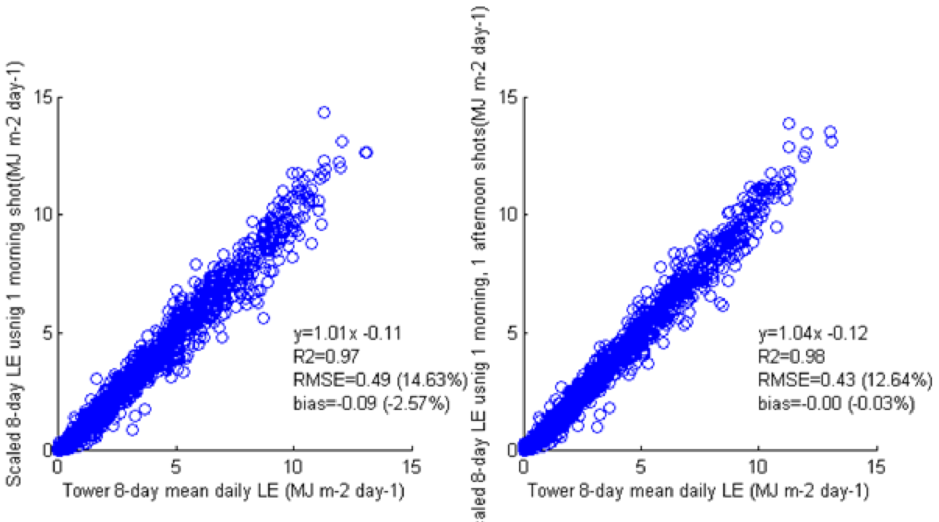


Total: $1420



Current Status (5/6/2010)

- 10 US year results encouraging
  - Still some work to be done when forest floor is snow covered
- 1 FluxTower year now under investigation
  - 1 FluxTower year ~ 4 US years
  - Adds significant biomes such as tropical rain forests and tundras
  - Added comparison with similar European sites
- Global calculation with 5 KM pixels under consideration
  - 1 global year ~ 1 US year



Comparison of using only morning observation with using both morning and afternoon observation. Plotted is ET expressed as LE computed Scaled 8-day average vs Flux tower 8-day mean daily.

Br-SP1
Sao Paulo Cerrado

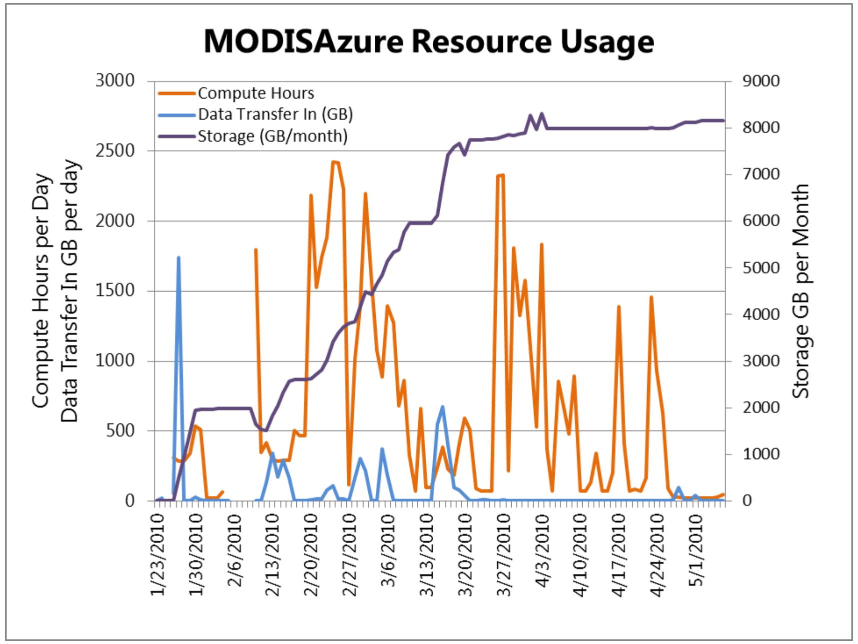
Br-Ma2
Manaus - ZF2 K34

Summary

I can see clearly now, the rain has gone. I can see all obstacles in my way.
Johnny Nash

Learnings

- Lowering the barriers to use remote sensing data can enable science
  - NASA makes the data accessible, not science ready
  - At AGU 2009, we learned that a cloud service that just made on-demand jpg mosaics would help tremendously
- Science and algorithm debugging benefit from the same infrastructure as both need to scale up and down
  - Debugging an algorithm on the desktop isn't enough – you have to debug in the cloud too
  - Whenever running at scale in the cloud, you must reduce down to the desktop to understand the results
- Putting all your eggs in the cloud basket means watching that basket
  - Cloud scale resources often mean you still manage small numbers of resources: 100 instances over 24 hours = $288 even if idle
  - Where is the long term archive for any results ?
- Azure is a rapidly moving target and unlike the Grid
  - Commercial cloud backed by large commercial development team
  - Bake in the faults for scaling and resilience



Acknowledgements

Berkeley Water Center, University of California, Berkeley, Lawrence Berkeley Laboratory
- Jim Hunt
- Dennis Baldocchi
- Deb Agarwal
- Monte Goode
- Keith Jackson
- Rebecca Leonardson (student)
- Carolyn Remick
- Susan Hubbard

University of Virginia
- Marty Humphrey
- Norm Beekwilder
- Jie Li (student)

San Diego Supercomputing Center
- Ilya Zavlavsky
- David Valentine
- Matt Rodriguez (student)
- Tom Whitenack

CUAHSI
- David Maidment
- David Tarboton
- Rick Hooper
- Jon Goodman

RENCI
- John McGee
- Oleg Kapeljushnik (student)

Fluxnet Collaboration
- Dennis Baldocchi
- Rodrigo Vargas (postdoc)
- Youngryel Ryu (student)
- Dario Papale (CarboEurope)
- Markus Reichstein (CarboEurope)
- Hank Margolis (Fluxnet-Canada)
- Alan Barr (Fluxnet-Canada)
- Bob Cook
- Susan Holladay
- Dorothea Frank

Ameriflux Collaboration
- Beverly Law
- Tara Hudiburg (student)
- Gretchen Miller (student)
- Andrea Scheutz (student)
- Christoph Thomas
- Hongyan Luo (postdoc)
- Lucie Ploude (student)
- Andrew Richardson
- Mattias Falk
- Tom Boden

North American Carbon Program
- Kevin Schaefer
- Peter Thornton

University of Queensland
- Jane Hunter

University of Indiana
- You-Wei Cheah (student)

Microsoft Research
- Yogesh Simmhan
- Roger Barga
- Dennis Gannon
- Jared Jackson
- Nelson Araujo
- Wei Liu
- Tony Hey
- Dan Fay

EXTREME COMPUTING GROUP Microsoft Defining the future.

http://azurescope.cloudapp.net/

http://www.fluxdata.org