Microsoft® Research

FacultySummit 2011

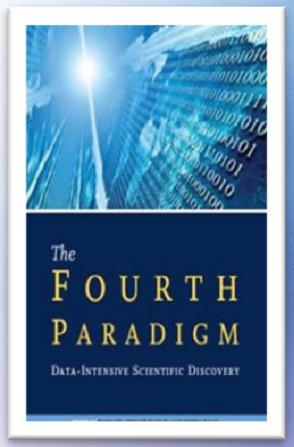
Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS

Scaling Science in the Cloud: From Satellite to Science Variables at the Global Scale with MODISAzure

Catharine van Ingen
Partner Architect, Microsoft Research
Presented by Harold Javid

The Data Flood: Science and the 4th Paradigm

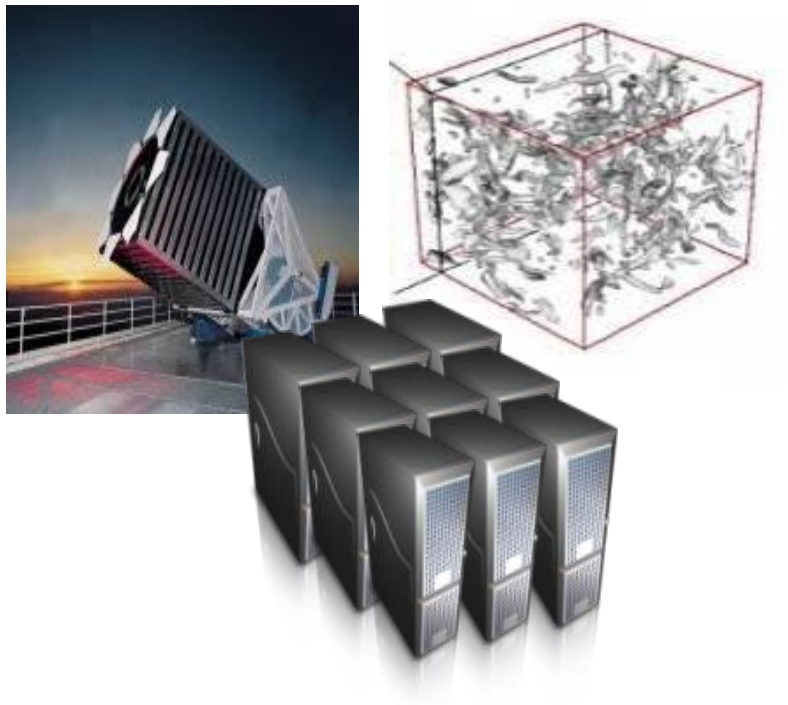
Thoughts without content are empty,
intuitions without concepts are blind.
Immanuel Kant



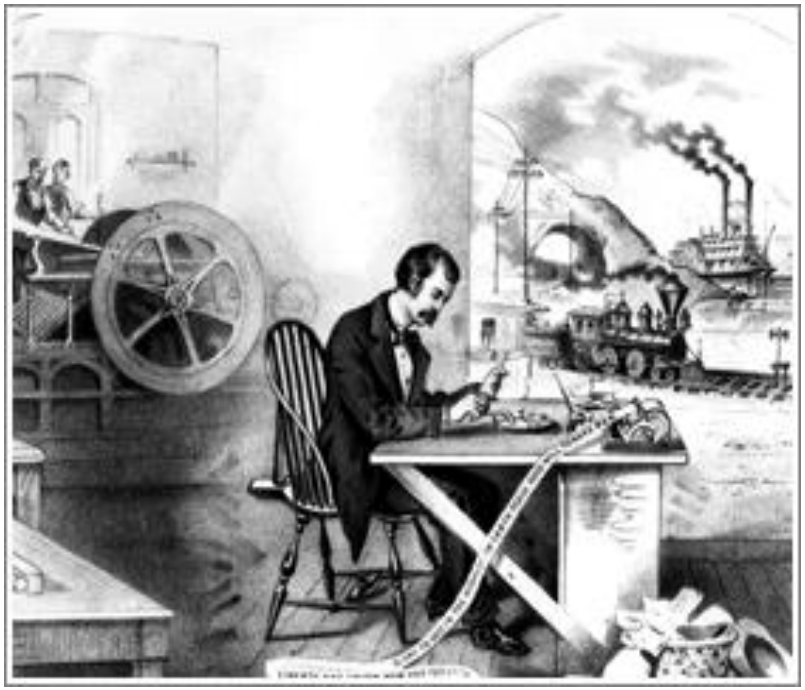
From Jim Gray, 2007 Emergence of a Fourth Paradigm

- 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

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



From George Djorgovski, LATAM Summit 2010



Information technology revolution is historically unprecedented - in its impact it is like the industrial revolution and the invention of printing combined

It is transforming science and scholarship as much as any other field of the modern human endeavor, as they become data-rich, and computationally enabled

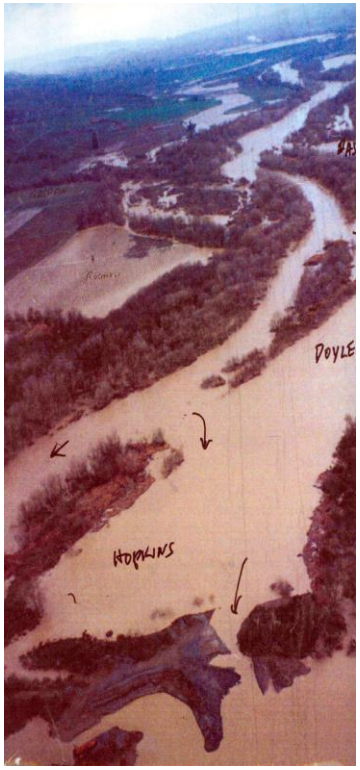
Through e-Science, we are developing a new scientific methodology for the 21st century

Environmental Data Comes in Many Forms

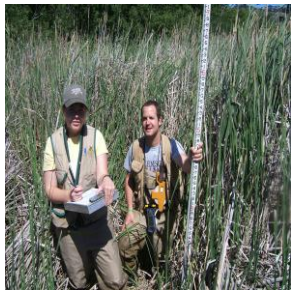
Manual Measurement

Automated Measurement

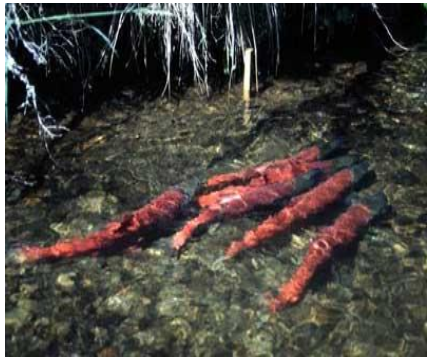
Sample Collection



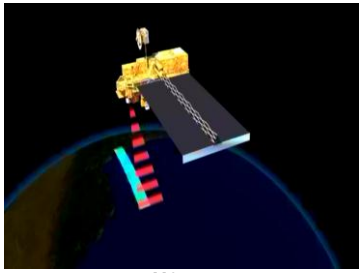
Historical Photographs



Typing

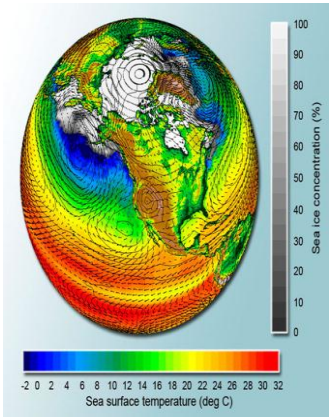


Counting



Satellite

Aircraft Surveys

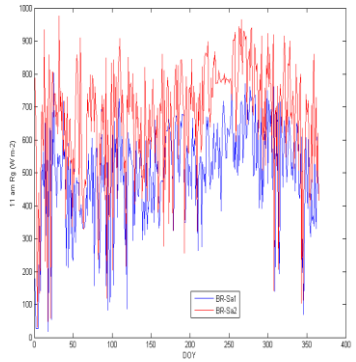


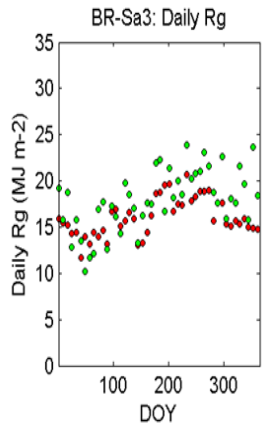
Model Output

Relatively Ubiquitous Motes

Ever Increasing Distance from Observation

- Deriving science variables from sensor output often active research in its own right
  - Handling day/night or 3-d effects is challenging
- Observational data has spikes, drift, and gaps
  - Correcting these must be done with knowledge of the science as well as the instrument
- Systematic and random errors introduced by such transformations often understood only when data are used for analysis.
- Some data users ignore all of these concerns while others pay a lot of attention.

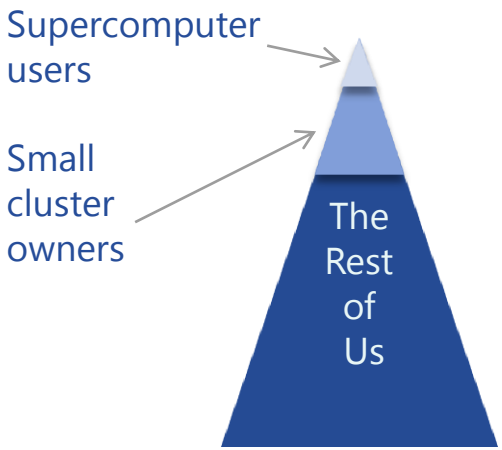




Dataset creation takes work and specialized knowledge. Data reuse amortizes that and improves overall quality.

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:
  - obtain 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

MODISAzure: Estimating Water Balance in the Cloud

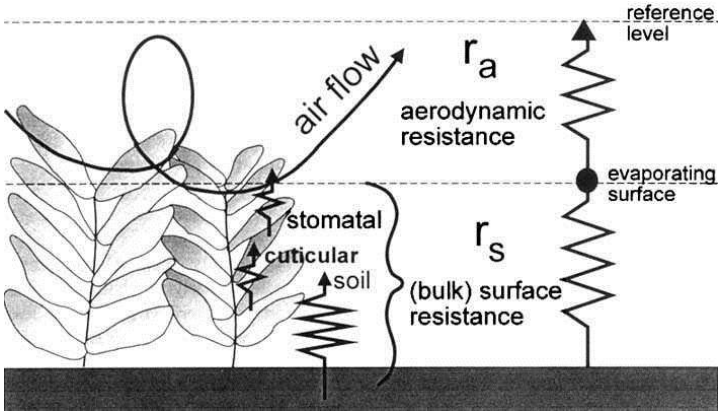
You never miss the water 'til the well has run dry
Irish Proverb

Computing Water Balance (ET) from First Principles

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

Penman-Monteith (1964)

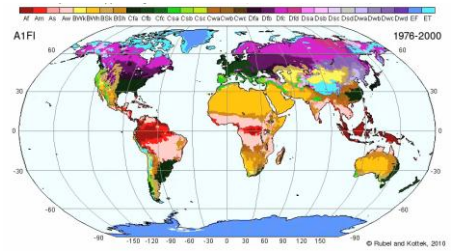
ET = Water volume evapotranspired ($m^3\ s^{-1}\ m^{-2}$)
Δ = Rate of change of saturation specific humidity with air temp.($Pa\ K^{-1}$)
λ_v = Latent heat of vaporization (J/g)
R_n = Net radiation ($W\ m^{-2}$)
c_p = Specific heat capacity of air ($J\ kg^{-1}\ K^{-1}$)
ρ_a = dry air density ($kg\ m^{-3}$)
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 ($\gamma \approx 66\ Pa\ K^{-1}$)



Estimating resistance/conductivity across a catchment can be tricky

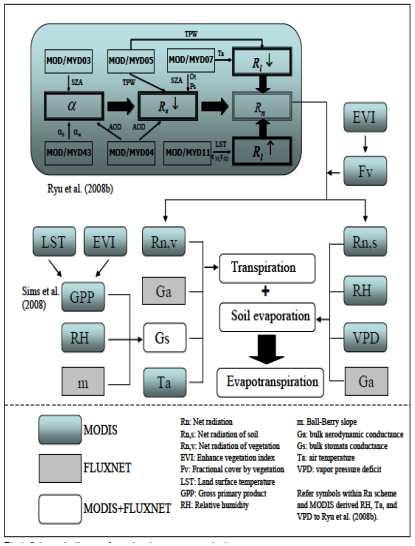
- Lots of inputs : big reduction
- Some of the inputs are not so simple
- Many have categorical dependencies

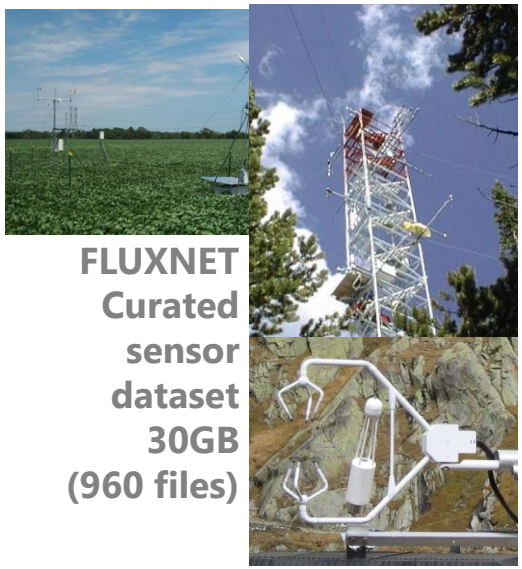
Estimating ET from Imagery, Sensors and Field Data



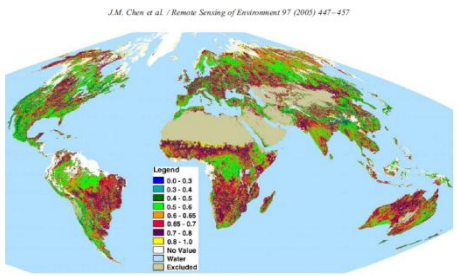
Climate classification
~1MB (1file)

Not just a simple matrix computation due to dry region leaf/air temperatures differences, snow cover, leaf area fill, temporal up-scaling, gap fill, biome conductance lookup, C3/C4 plants, etc. etc.

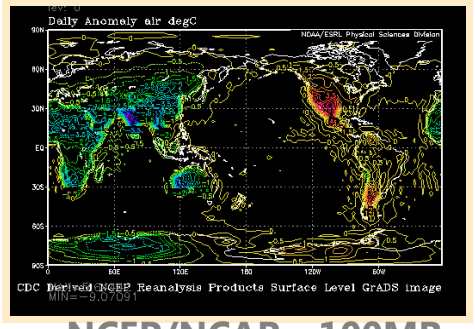




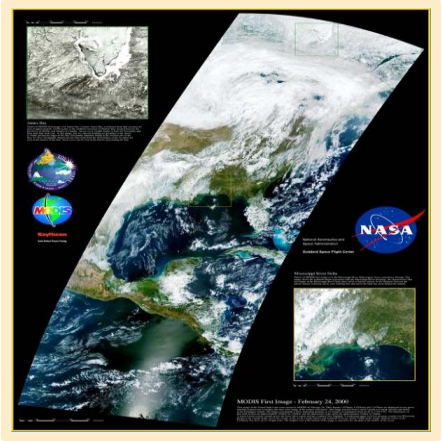
FLUXNET
Curated
sensor
dataset
30GB
(960 files)



Vegetative clumping
~5MB (1file)



NCEP/NCAR ~100MB
(4K files)



NASA MODIS imagery archives
5 TB (600K files) for 10 US years

FLUXNET
curated field dataset
2 KB (1 file)

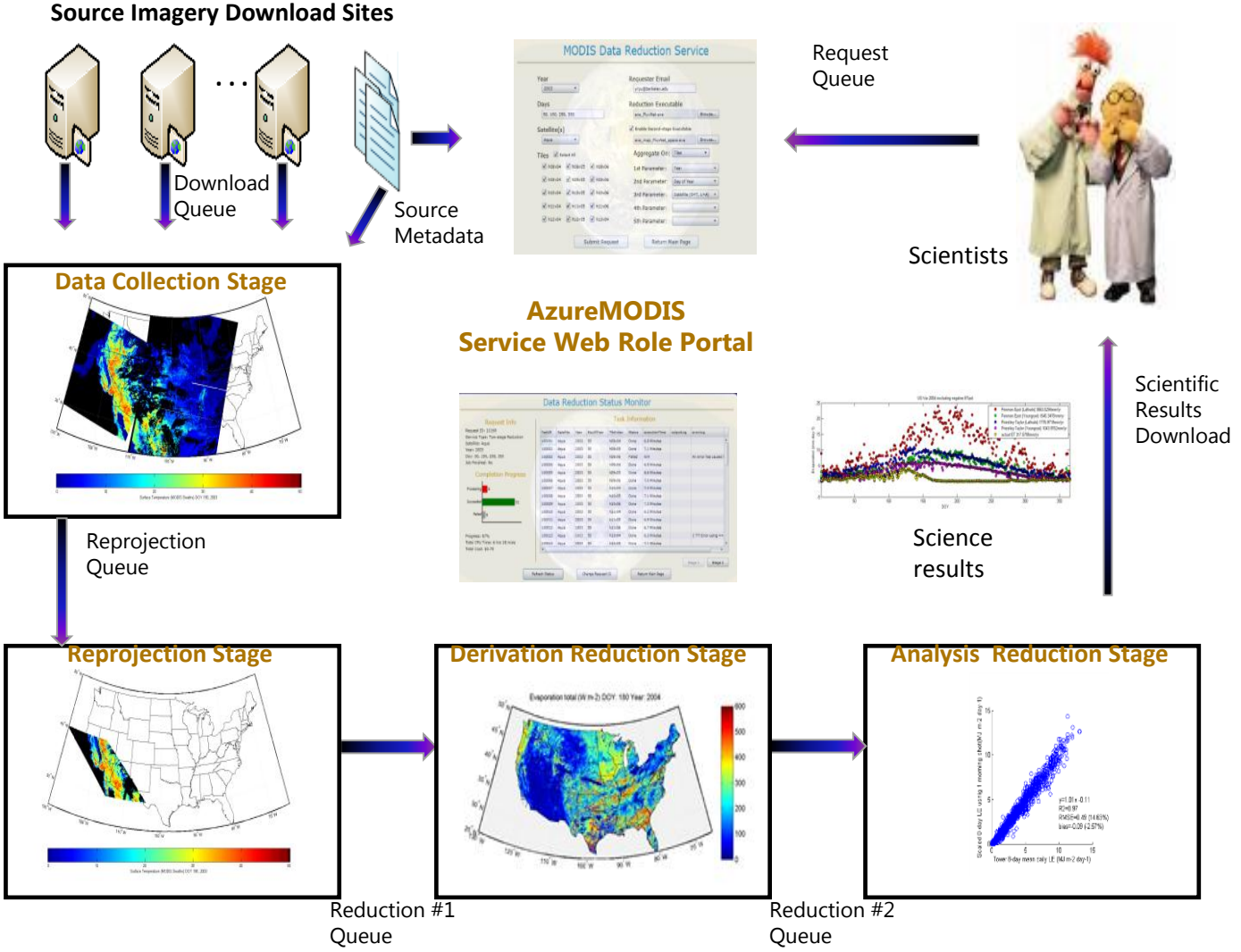
MODISAzure: Four Stage Image Processing Pipeline

Data collection **(map)** 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 **(map)** stage
- Converts source tile(s) to intermediate result sinusoidal tiles
- Simple nearest neighbor or spline algorithms

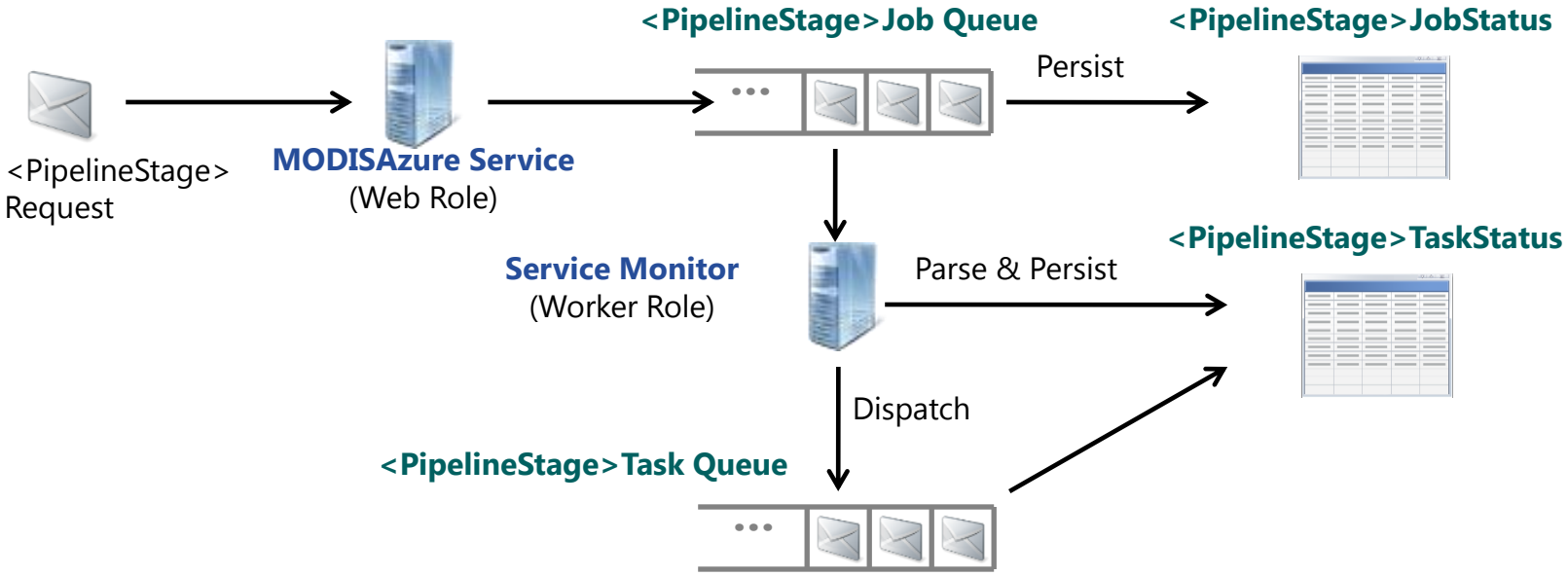
Derivation **reduction** stage
- First stage visible to scientist
- Computes ET in our initial use

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



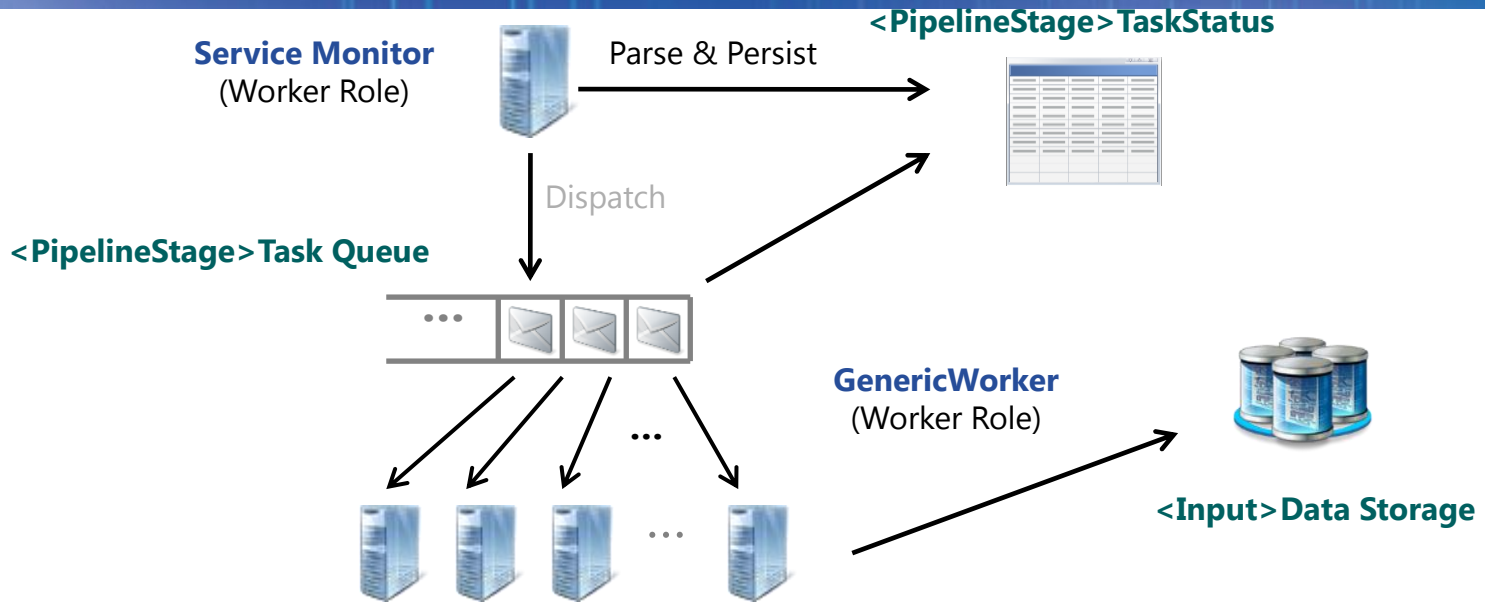
http://research.microsoft.com/en-us/projects/azure/azuremodis.aspx

MODISAzure: Architectural Big Picture (1/2)



- ModisAzure Service is the Web Role front door
  - Receives all user requests
  - Queues request to appropriate Download, Reprojection, or Reduction Job Queue
- Service Monitor is a dedicated Worker Role
  - Parses all job requests into tasks – recoverable units of work
  - Execution status of all jobs and tasks persisted in Tables

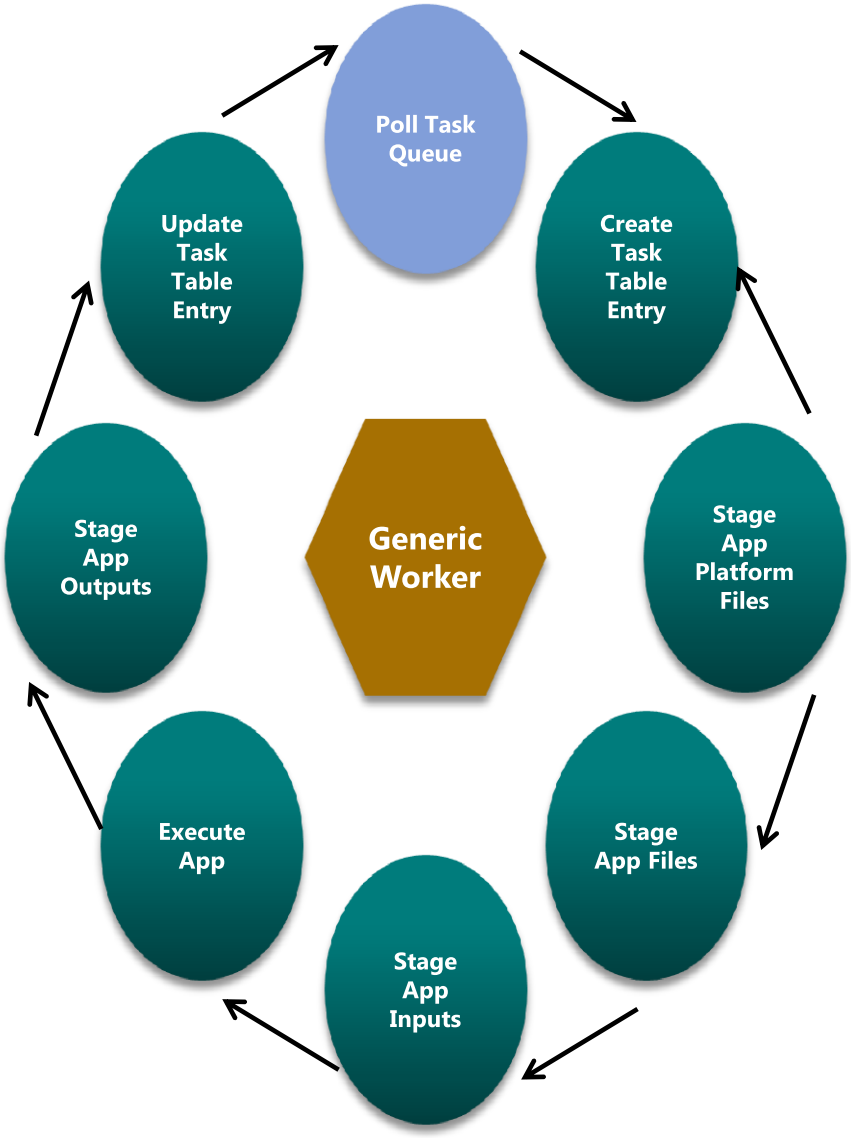
MODISAzure: Architectural Big Picture (2/2)



- All work actually done by a GenericWorker Worker Role
  - Dequeues tasks created by the Service Monitor
  - Retries failed tasks 3 times
  - Maintains all task status
  - Sandboxes science or other executable
  - Obtains all storage from/to Azure blob storage to/from local Azure Worker instance files

Inside A Generic Worker

- Manages application sandbox
  - Ensures all application binaries such as the MatLab runtime are installed for “known” application types
  - Stages all input blobs from Azure storage to local files
  - Passes any marshalled inputs to uploaded application binary
  - Stages all output blobs to Azure storage from local files
  - Preserves any marshalled outputs to the appropriate Task table
- Simplifies desktop development and cloud deployment



Storage Management

Source

- Original source image download
  - Can be deleted when all dependent reprojections complete

- Reduction results
  - Older results can be aged out over time
  - A zip file blob is created for each job to simplify download

Reduction Storage

- Reprojection results
  - May include the same target tile at different spatial resolution

Reprojection Storage

Metadata Storage

- Metadata includes geospatial lookup, known application library binaries, etc
  - Necessary for service function
  - Never directly accessed by scientist code

Storage separated by usage to simplify management policies

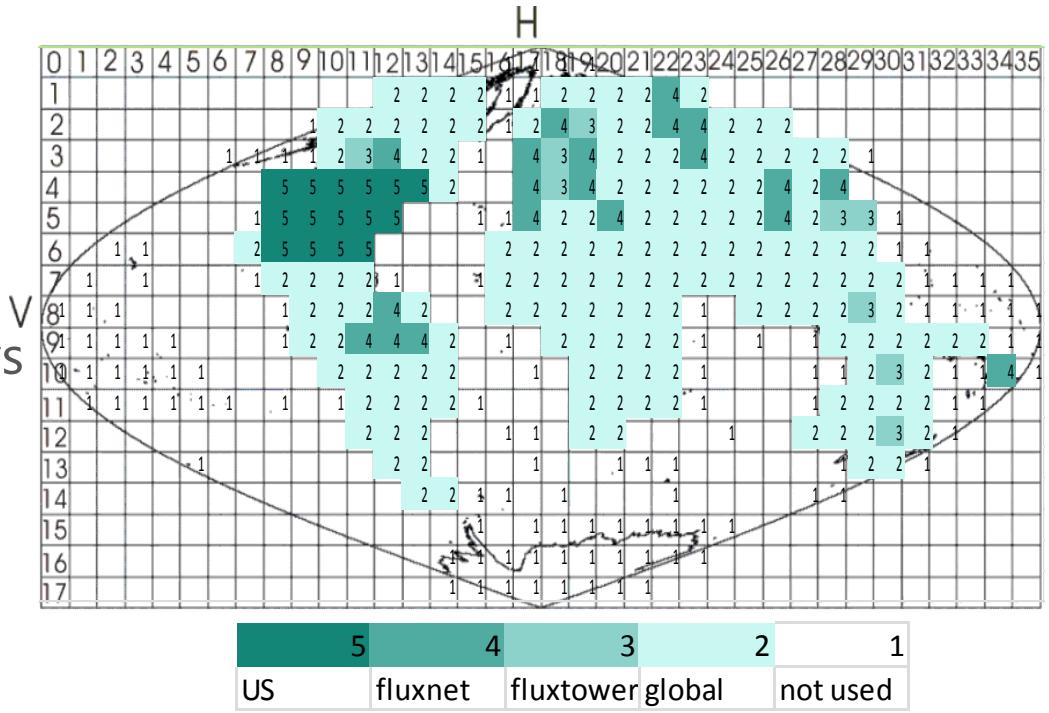
Pipeline Stage Priorities and Interactions

- 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

Sizing the 3 year MODISAzure Global Computation

- 194 sinusoidal cells, each covers 1.2x1.2 KM or 11M 5 KM pixels
- 1.06 M reprojected tiles and 40.5K source sinusoidal tiles
- 14 TB (>10 M files) downloaded from NASA ftp
- Not all files are downloaded or reprojected at first (3 rapid retries) attempt or actually available due to satellite outage, polar winter, missing tiles, etc. etc.

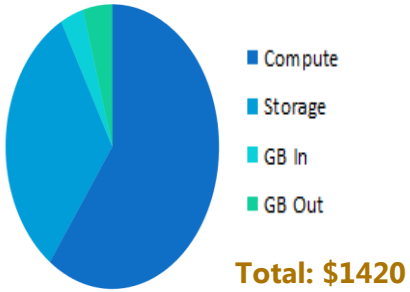
- 55 NASA download days
- 150K reprojection compute hours
- 940 TB moved across Azure fabric
- 1 month result download days (est) KM or 11M 5 KM pixels



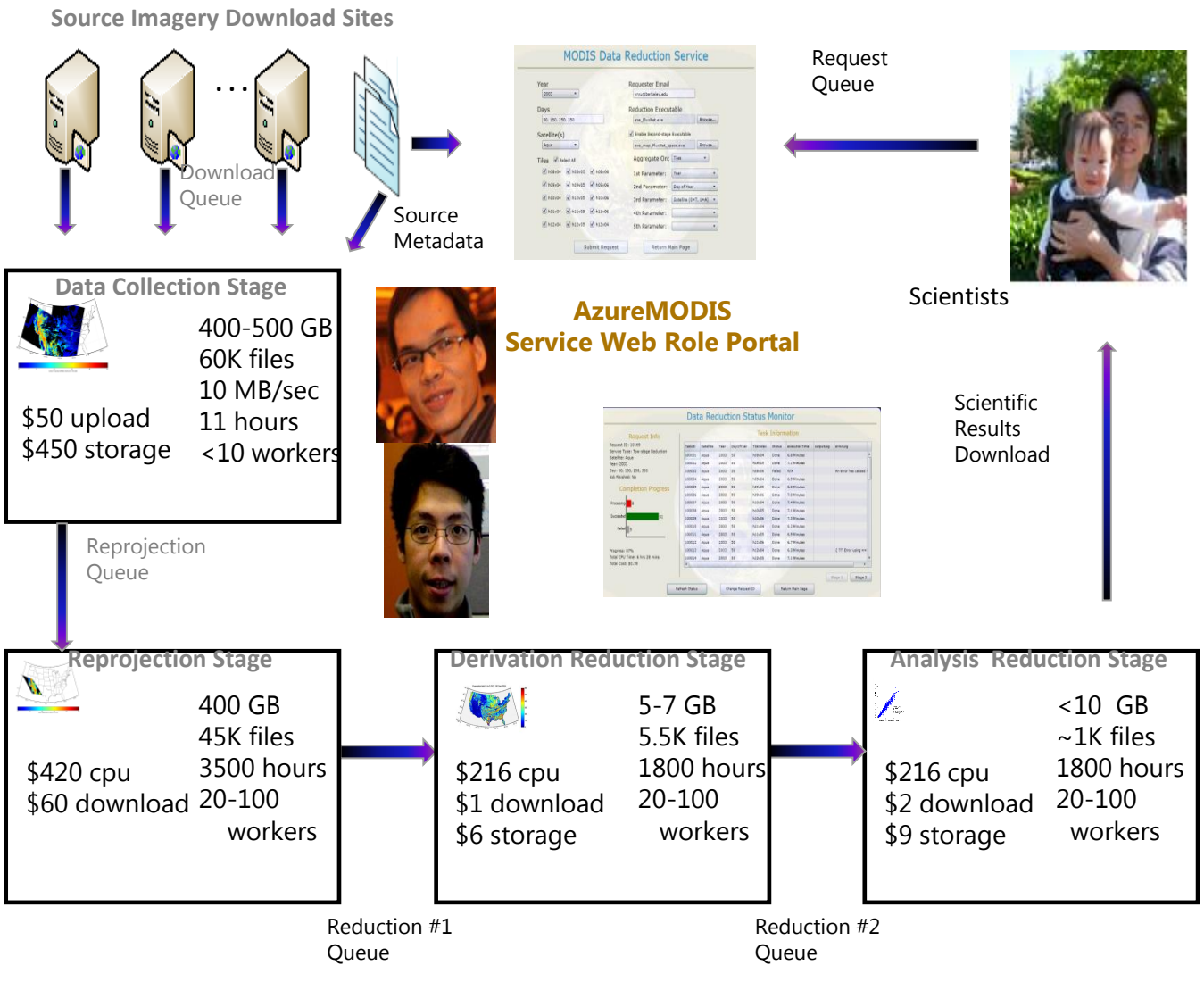
15 seconds on the Cray Jaguar (1.75 PFLOPs), but only if we could get the PB in !

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

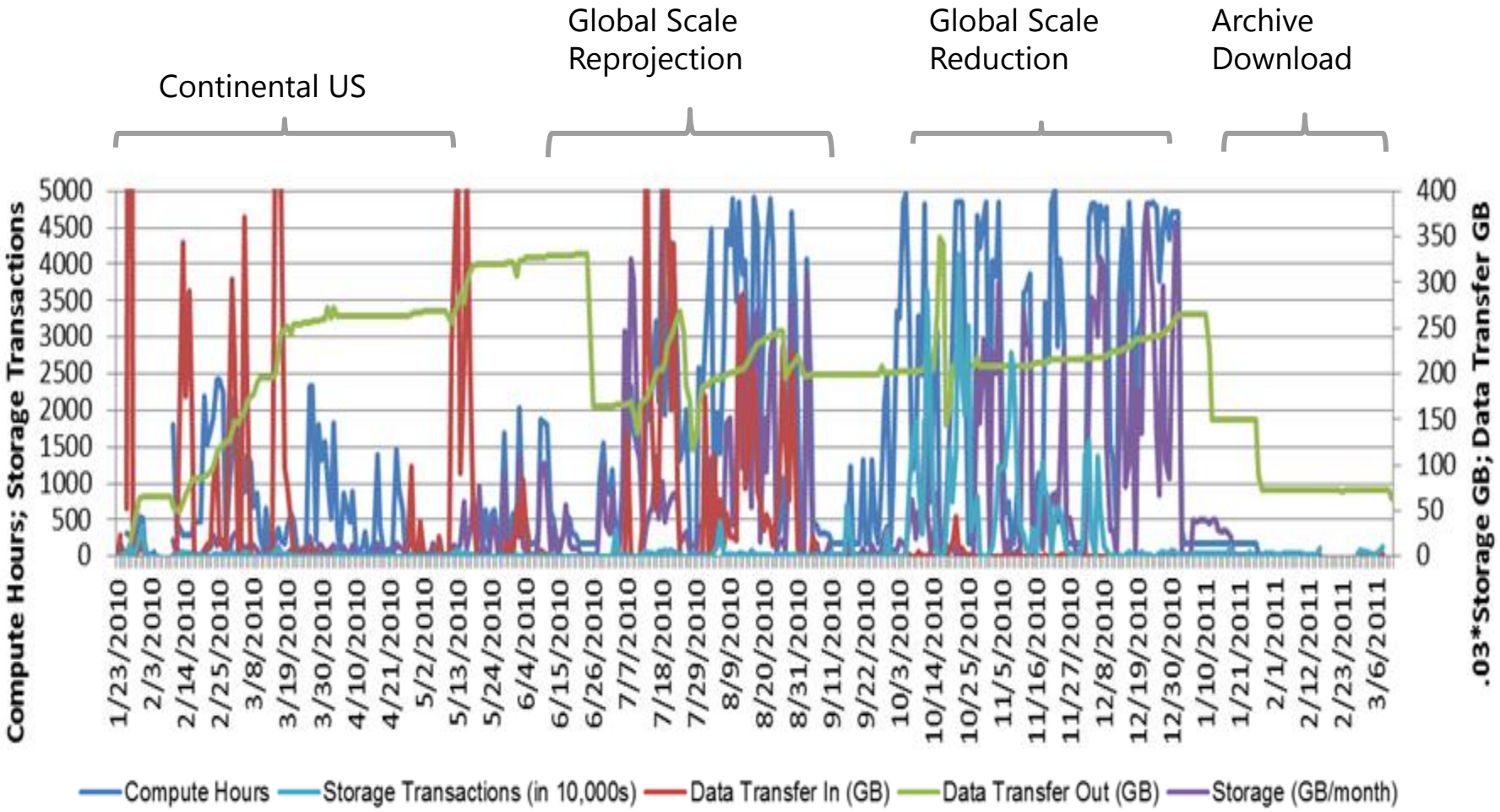


The MODISAzure "ity" Experience

Why, I'd like nothing better than to achieve some bold adventure, worthy of our trip.
Aristophanes

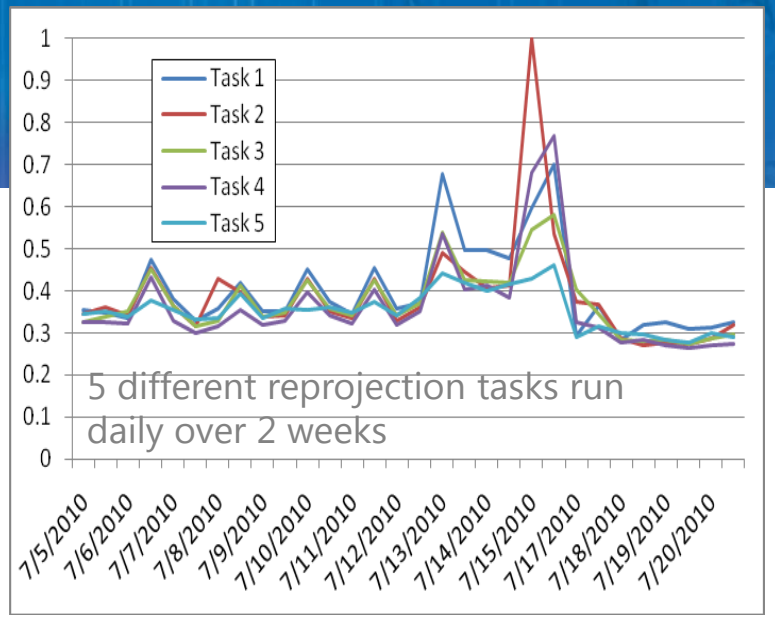
Agility

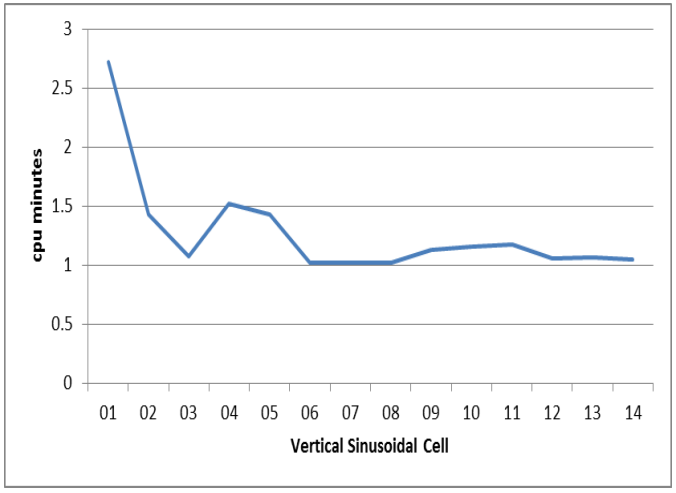
- The computation changed over time while Azure just scaled



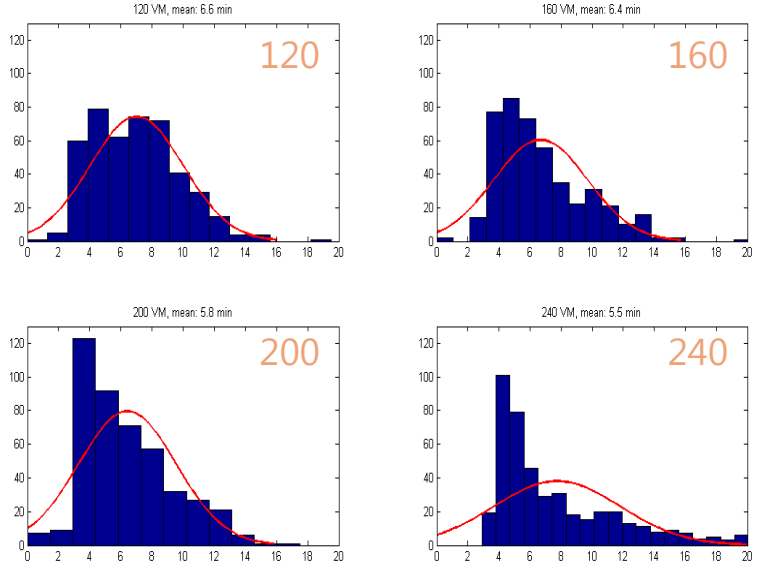
Predictability

- Performance varies over time: rerunning the same task gives different timings on different days
- Performance varies over space: satellites are over the poles more often





Average reprojection time (after algorithm improvements!) as a function of longitude

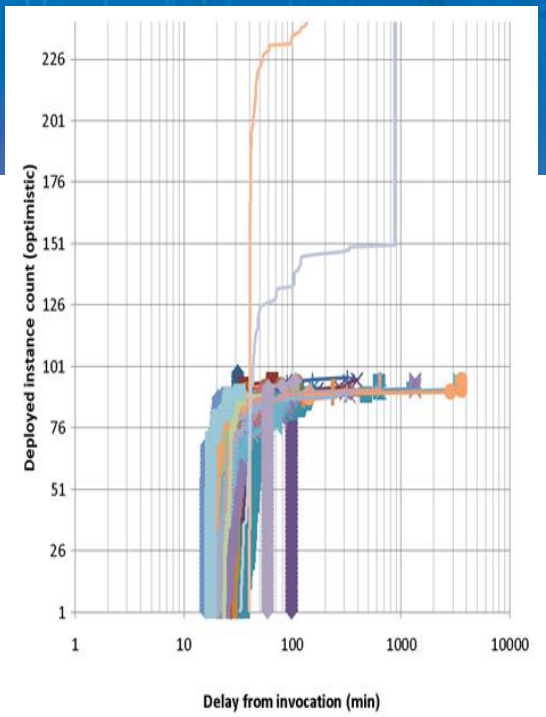


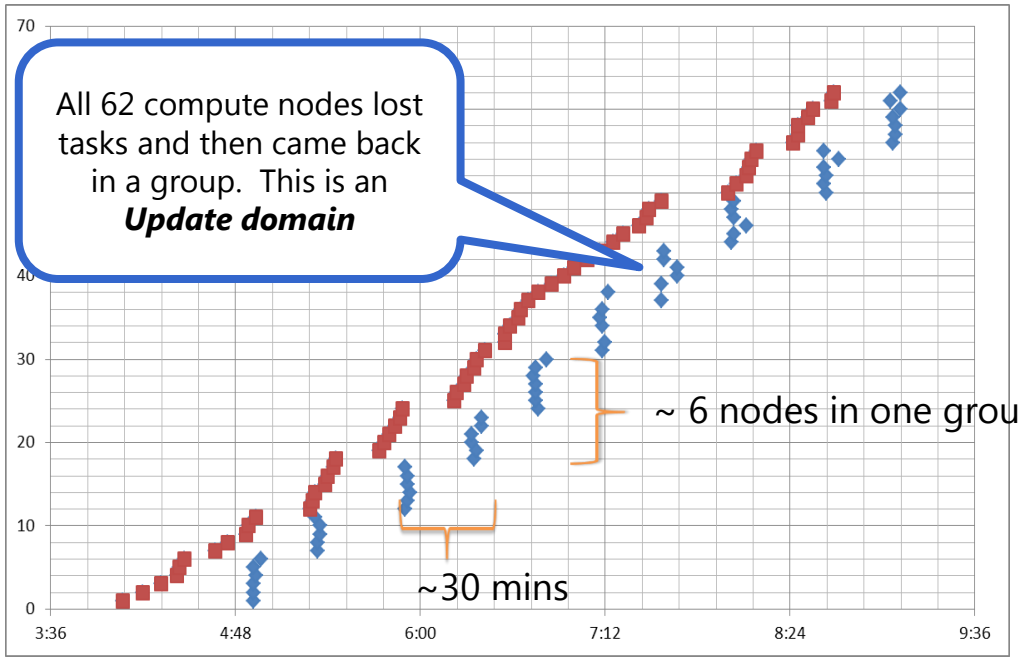
The same reduction task run on different numbers of VMs

Reliability

- Even with 99.999% reliability, bad things happen
  - 1-2 % of MODISAzure tasks fail but succeed on retry

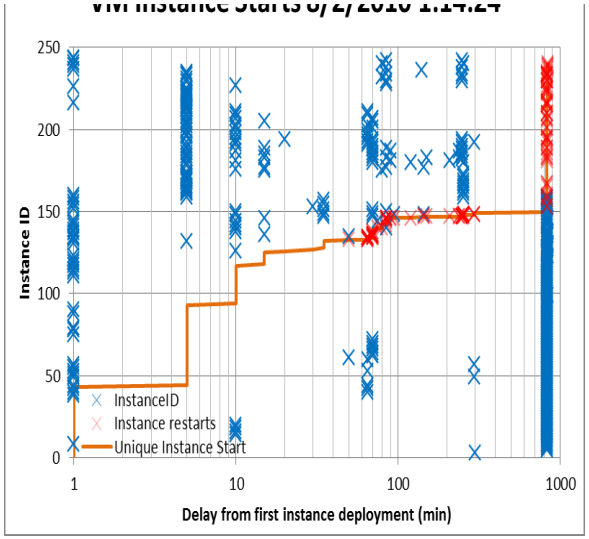
Observed VM starts for 76-100 VMs





From AzureBlast
http://research.microsoft.com/en-us/people/barga/faculty_summit_2010.pdf

Worst case attempt to start 250 VMs



Maintainability

- Some “Early Adopter” artifacts
  - Generic worker sandbox
  - “dir” for blobs : need to have a parsable list, not just browse and many tools simply could not scale beyond O(50K) blobs
  - “downloader” for blobs : smaller blobs are dwarfed by REST open/close.

  - Slow upload (FEDEX disk is still “in plan”; IN2 connections helped download tremendously)
- Can we move catalog and other tracking to SQL Azure for better scaling?
  - Current tracking database is 140 GB
  - Partitions naturally, but would mean $300/mo (external) charges.

Conclusion

Adventure is just bad planning.
Roald Amundsen

The Data are Coming ! The Cloud is Here !

- We have much work ahead mapping science requirements to the new evolving cloud infrastructures.
  - Science computations are becoming much more diverse.
  - Cloud computing is just beginning.
- Azure means doing some things differently and leveraging new capabilities.
  - Virtualized computing resources often are black box resources
  - New capabilities still emerging
- We need research to develop best practices for scaling up!
  - “Rare” events become more common and consume time
  - What’s common? What’s specific to the science domain or computation?

Cloud Computing Learnings

- Clouds are the largest scale computer centers ever constructed and have the potential to be important to both large and small scale science problems.
- Clouds suitable for "loosely coupled" data parallel applications, but tightly coupled low-latency applications perform poorly on clouds today.
- Clouds exploit economies of scale, healthy commercial competition, and an active research community.

Chicago, IL

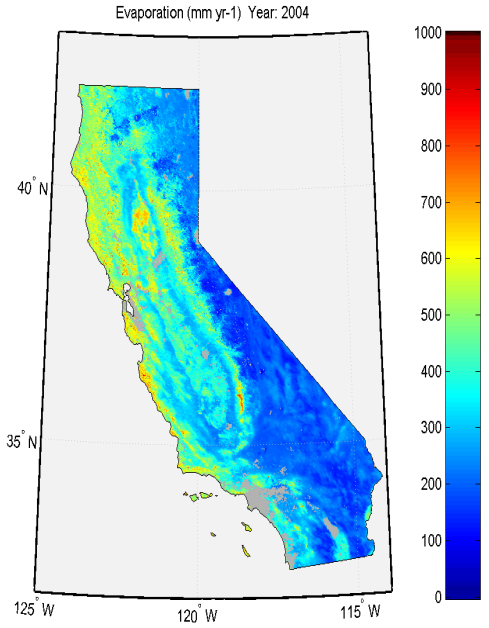
Dublin, Ireland

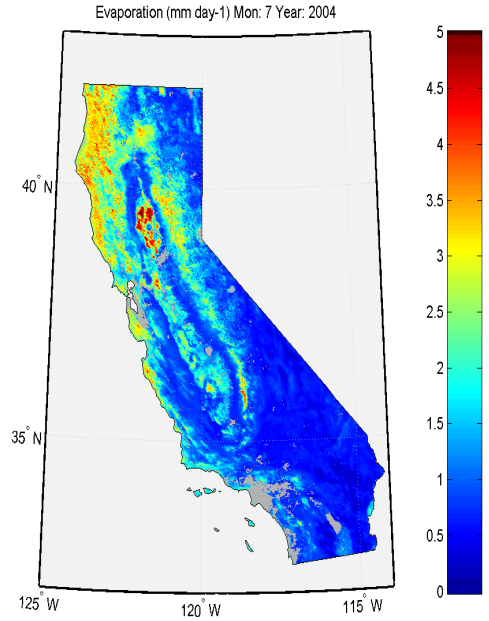
Generation 4 DCs

Science computations are becoming more diverse. We have much work ahead mapping those new needs to evolving cloud infrastructures.

Azure Learnings

- 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
- Azure is a rapidly moving target and unlike the Grid
  - We've seen many API changes and new services over the last year
- At scale, understanding even a 0.01% failure rate is time consuming
  - Bake in the faults for scaling and resilience
  - Bake in end:end reconciliation of sources and results

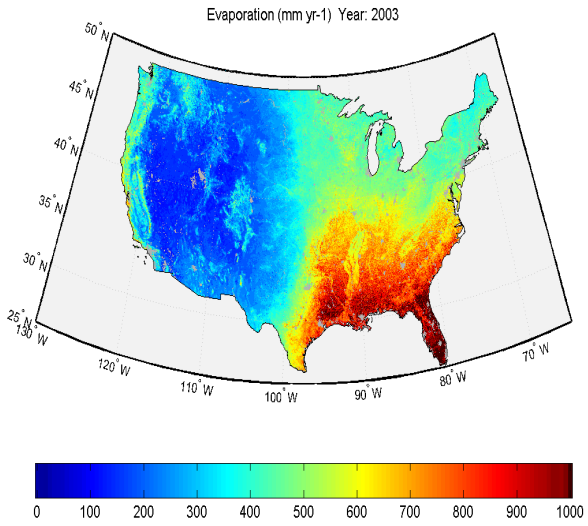


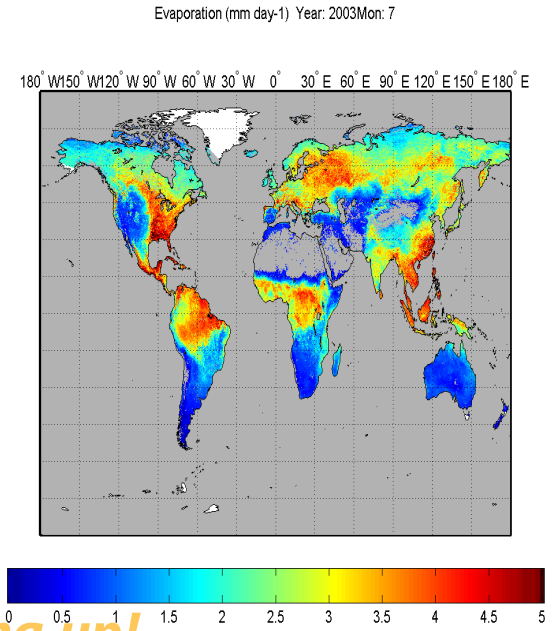


Azure means doing some things differently and leveraging new capabilities.

eScience Learnings

- 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
- Developing concrete plans for capacity planning prior to having results in hand is tricky
  - Precedents break down when scaling up 100x or more
  - Don't forget to include sensitivity and error analyses requirements





We need research to develop best practices for scaling up!

Acknowledgements

Microsoft Research
- Dan Reed
- Tony Hey
- Dennis Gannon
- David Heckerman
- Nelson Araujo
- Dan Fay
- Jared Jackson
- Wei Liu
- Jaliya Ekanayake
- Simon Mercer
- Yogesh Simmhan
- Michael Zyskowski

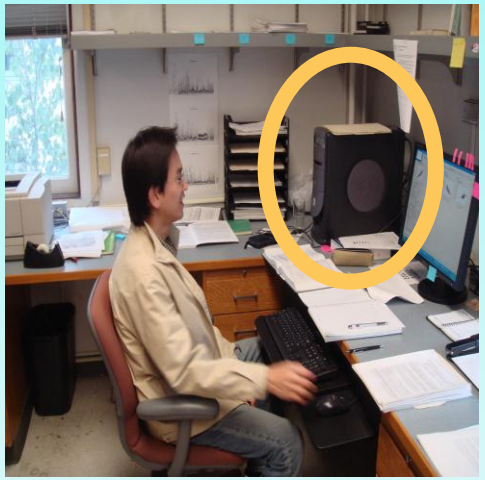
Berkeley Water Center, University of California, Berkeley, Lawrence Berkeley Laboratory
- Deb Agarwal
- Dennis Baldocchi
- Jim Hunt
- Monte Goode
- Susan Hubbard
- Keith Jackson
- Rebecca Leonardson (student)
- Carolyn Remick

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

Indiana University
- You-Wei Cheah (student)

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

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



Youngryel was lonely with 1 PC

http://azurescope.cloudapp.net/

http://www.fluxdata.org