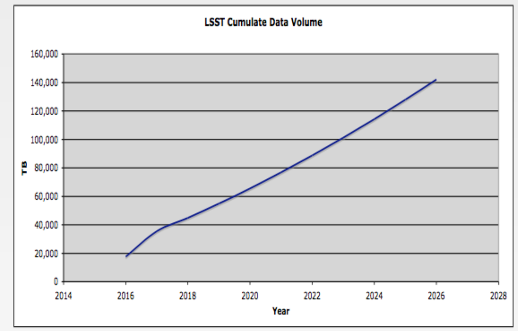
Approaches to Investigating Technology Solutions In Data Intensive Astronomy

G. Bruce Berriman
gbb@ipac.caltech.edu
NASA Exoplanet Science Institute,
Infrared Processing and Analysis Center, Caltech

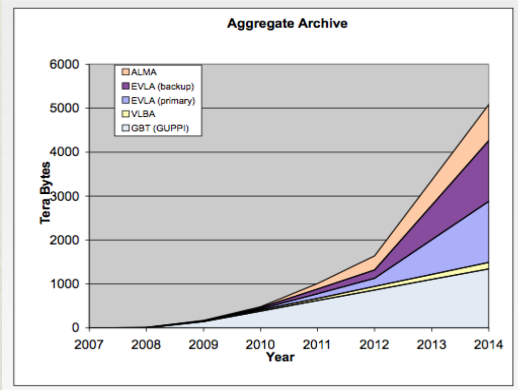
Innovations in Data Intensive Astronomy, May 3-5 2011.

Developing A New Business Model For Astronomical Computing

- ❖ Astronomy is already a data intensive science
  - Over 1 PB served electronically through data centers and archives.
  - Growing at 0.5 PB/yr, and accelerating.



- ❖ Astro2010 recognized that future research will demand high performance computing on massive, distributed data sets.
  - High Performance/Massive Parallelization: Scalability
  - Current model for managing data unsustainable: universities hitting "power wall"



- ❖ **Learn how to unleash the power of new technologies**
- ❖ **Learn how to write applications that take advantage of the technology**
- ❖ **Learn how to develop innovative data discovery and access mechanisms.**

Cloud Computing In A Nutshell

New model for purchasing resources:
pay only for what you use.

This looks cheap!

Amazon EC2 front page:

Region: US East (Virginia)

	Linux/UNIX Usage	Windows Usage
Standard On-Demand Instances		
Small (Default)	$0.085 per hour	$0.12 per hour
Large	$0.34 per hour	$0.48 per hour
Extra Large	$0.68 per hour	$0.96 per hour
Micro On-Demand Instances		
Micro	$0.02 per hour	$0.03 per hour
Hi-Memory On-Demand Instances		
Extra Large	$0.50 per hour	$0.62 per hour
Double Extra Large	$1.00 per hour	$1.24 per hour
Quadruple Extra Large	$2.00 per hour	$2.48 per hour
Hi-CPU On-Demand Instances		
Medium	$0.17 per hour	$0.29 per hour
Extra Large	$0.68 per hour	$1.16 per hour
Cluster Compute Instances		
Quadruple Extra Large	$1.60 per hour	N/A*
Cluster GPU Instances		
Quadruple Extra Large	$2.10 per hour	N/A*

* Windows® is not currently available for Cluster Compute or Cluster GPU Instances

http://aws.amazon.com/ec2/

Commercial Providers
Amazon.com EC2
AT&T Synaptic Hosting
GNi Dedicated Hosting
IBM Computing on Demand
Rackspace Cloud Servers
Savvis Open Cloud
ServePath GoGrid
Skytap Virtual Lab
3Tera
Unisys Secure
Verizon Computing
Zimory Gateway
Science Clouds
FutureGrid
NERSC Magellan
NASA Nebula

“Little sins add up …”

OS	EC2 Instace	Demand Type	Cost / Hr	Hours	Length	Total
Windows	HCPU Extra Large	OnDemand	$1.16	8,736	Year	$10,133.76
Windows	Extra Large	OnDemand	$0.96	8,736	Year	$8,386.56
Linux/UNIX	Extra Large	OnDemand	$0.68	8,736	Year	$5,940.48
Linux/UNIX	HCPU Extra Large	OnDemand	$0.68	8,736	Year	$5,940.48
Linux/UNIX	Large	OnDemand	$0.68	8,736	Year	$5,940.48
Windows	HCPU Extra Large	Reserved	$0.50	8,736	Year	$4,368.00
Windows	Large	OnDemand	$0.48	8,736	Year	$4,193.28
Windows	HCPU Medium	OnDemand	$0.29	8,736	Year	$2,533.44
Linux/UNIX	Extra Large	Reserved	$0.24	8,736	Year	$2,096.64
Linux/UNIX	HCPU Extra Large	Reserved	$0.24	8,736	Year	$2,096.64
Linux/UNIX	HCPU Medium	OnDemand	$0.17	8,736	Year	$1,485.12
Linux/UNIX	Large	Reserved	$0.12	8,736	Year	$1,048.32
Windows	Small	OnDemand	$0.12	8,736	Year	$1,048.32

Instance	Memory (MB)	Virtual Core	ECU	ECU per Core	Storage (GB)	I/O	Platform
Micro Instance	633	1	2	-			32/64 bit
Small Instance - default	1740.8	1	1	1	160	Moderate	32 bit
Large Instance	7680	2	4	2	850	High	64 bit
Extra Large Instance	15360	4	8	2	1690		64 bit
High-Memory Extra Large Instance	17510.4	2	6.5	3.25	420	Moderate	64 bit
High-Memory Double Extra Large Instance	35020.8	4	13	3.25	850	High	
High-Memory Quadruple Extra Large Instance	70041.6	8	26	3.25	1690		64 bit
High-CPU Extra Large Instance	7168	8	20	2.5	1690	High	64 bit
Cluster Compute Quadruple Extra Large Instance*	18432	8	33.5	4.1875	1690	Very Large	64 bit
Cluster GPU Quadruple Extra Large Instance **	18432	8	33.5	4.1875	1690	Very Large	64 bit

Annual Costs!

… and that’s not all. You pay for:

- Transferring data into the cloud
- Transferring them back out again
- Storage while you are processing (or sitting idle)
- Storage of the VM and your own software
- Special services: virtual private cloud…

See Manav Gupta’s blog post *http://manavg.wordpress.com/2010/12/01/amazon-ec2-costs-a-reality-check/*

How Useful Is Cloud Computing For Scientific Workflow Applications?

- ❖ Loosely-coupled parallel applications
  - ❖ Many domains: astronomy, biology, earth science, others
  - ❖ Potentially very large: 10K tasks common, >1M not uncommon
  - ❖ Potentially data-intensive: 10GB common, >1TB not uncommon
- ❖ Data communicated via files
  - ❖ Shared storage system, or network transfers required

1. Compare performance/cost of different resource configurations

2. Compare performance of grid and cloud

3. Characterize virtualization overhead

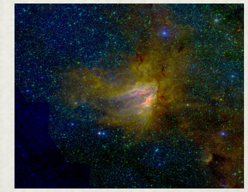
Scientific Workflow Applications on Amazon EC2. *G. Juve, et al.* ***arxiv.org/abs/1005.2718***
Data Sharing Options for Scientific Workflows on Amazon EC2*. G. Juve et al.*
arxiv.org/abs/1010.4822

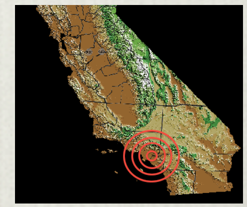
The Applications

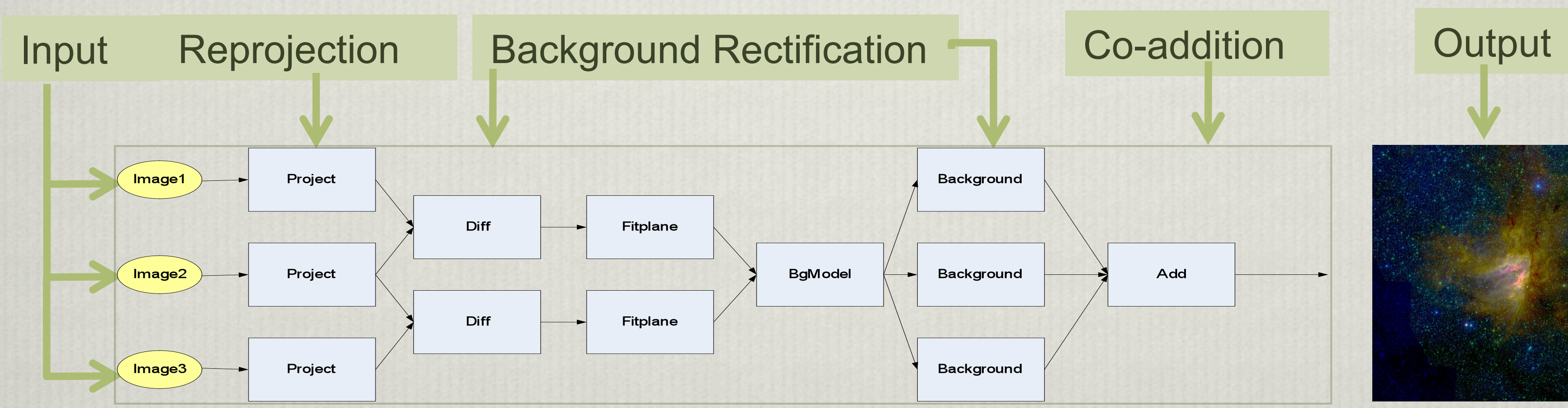
Montage **(http://montage.ipac.caltech.edu)** creates science-grade image mosaics from multiple input images.

Broadband calculates seismograms from simulated earthquakes.

Epigenome maps short DNA segments collected with gene sequencing machines to a reference genome.







Montage Workflow

Characteristics of Workflows

Workflow Specifications for this Study

Application	Workflow	# Tasks	Input	Output
Montage	8 deg. sq. mosaic of M16, 2MASS K-band	10,429	4.2 GB	7.9 GB
Broadband	4 earthquake sources, 5 sites	320	6 GB	160 MB
Epigenome	Maps DNA sequences to ref. chromosome 21	81	1.8 GB	300 MB

Resource Usage of the Three Workflow Applications

Application	I/O	Memory	CPU
Montage	High	Low	Low
Broadband	Medium	High	Medium
Epigenome	Low	Medium	High

Computing Resources

Type	Arch	CPU	Cores	Memory	Network	Storage	Price
Amazon EC2							
m1.small	32-bit	2.0-2.6 GHz Opteron	1-2	1.7 GB	1-Gbps Ethernet	Local	$0.10/hr
m1.large	64-bit	2.0-2.6 GHz Opteron	2	7.5 GB	1-Gbps Ethernet	Local	$0.40/hr
m1.xlarge	64-bit	2.0-2.6 GHz Opteron	4	15 GB	1-Gbps Ethernet	Local	$0.80/hr
c1.medium	32-bit	2.33-2.66 GHz Xeon	2	1.7 GB	1-Gbps Ethernet	Local	$0.20/hr
c1.xlarge	64-bit	2.0-2.66 GHz Xeon	8	7.5 GB	1-Gbps Ethernet	Local	$0.80/hr
Abe							
abe.local	64-bit	2.33 GHz Xeon	8	8 GB	10-Gbps InfiniBand	Local	…
abe.lustre	64-bit	2.33 GHz Xeon	8	8 GB	10-Gbps InfiniBand	Lustre	…

Processors and OS

- Amazon offers wide selection of processors.
- Ran Linux Red Hat Enterprise with VMWare
- *c1.xlarge* and *abe.local* are equivalent – estimate overhead due to virtualization
- *abe.lustre* and *abe.local* differ only in file system

Networks and File Systems

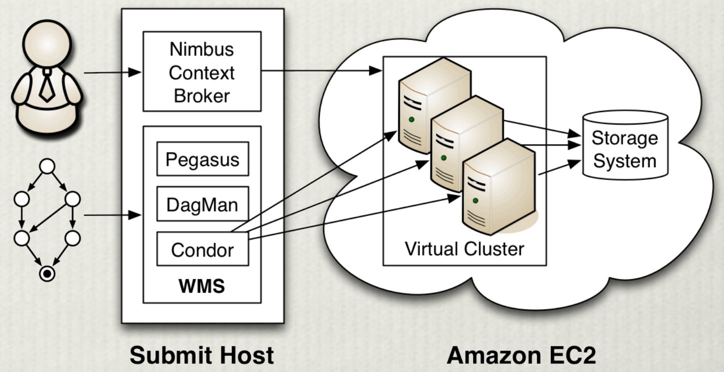
- HPC systems use high-performance network and parallel file systems
- Amazon EC2 uses commodity hardware

➔Ran all processes on single, multi-core nodes. Used local and parallel file system on Abe.

Execution Environment

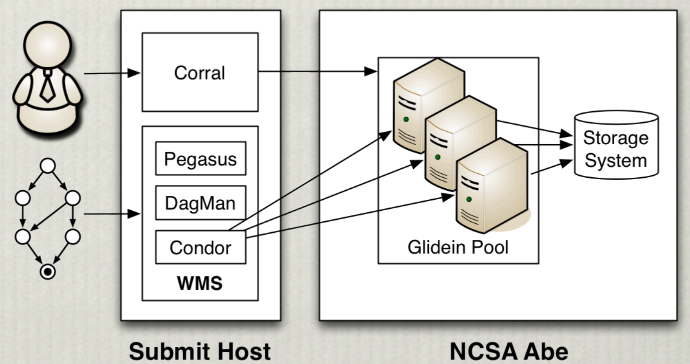
- Amazon provides the resources.
- End- user must configure and manage them

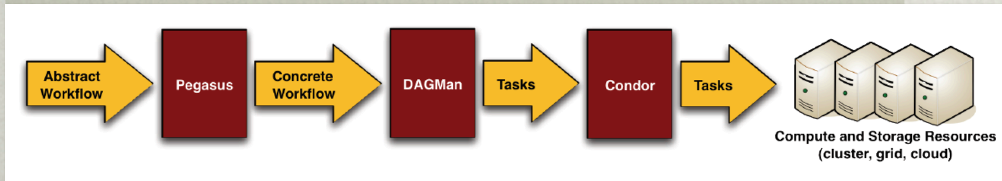
Amazon EC2



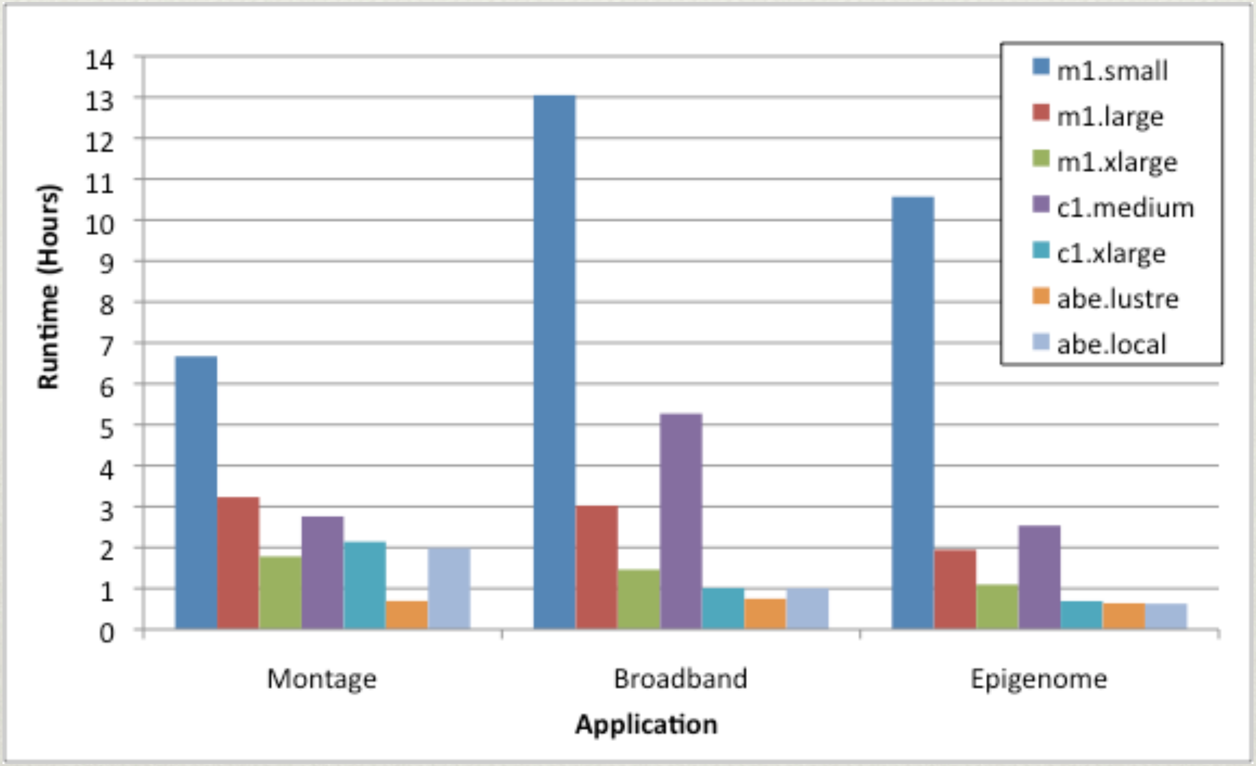
- **Pegasus** – workflow planner
  - Maps tasks and data from abstract descriptions to executable resources
  - Performance optimizer
- **DAGMan** – workflow engine
  - Tracks dependencies, releases tasks, retries tasks
- **Condor** – task manager; schedules and dispatches tasks (and data) to resources

NCSA Abe - high-performance cluster.



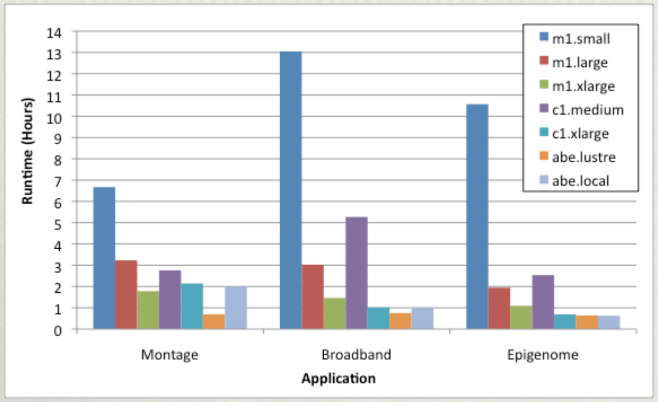


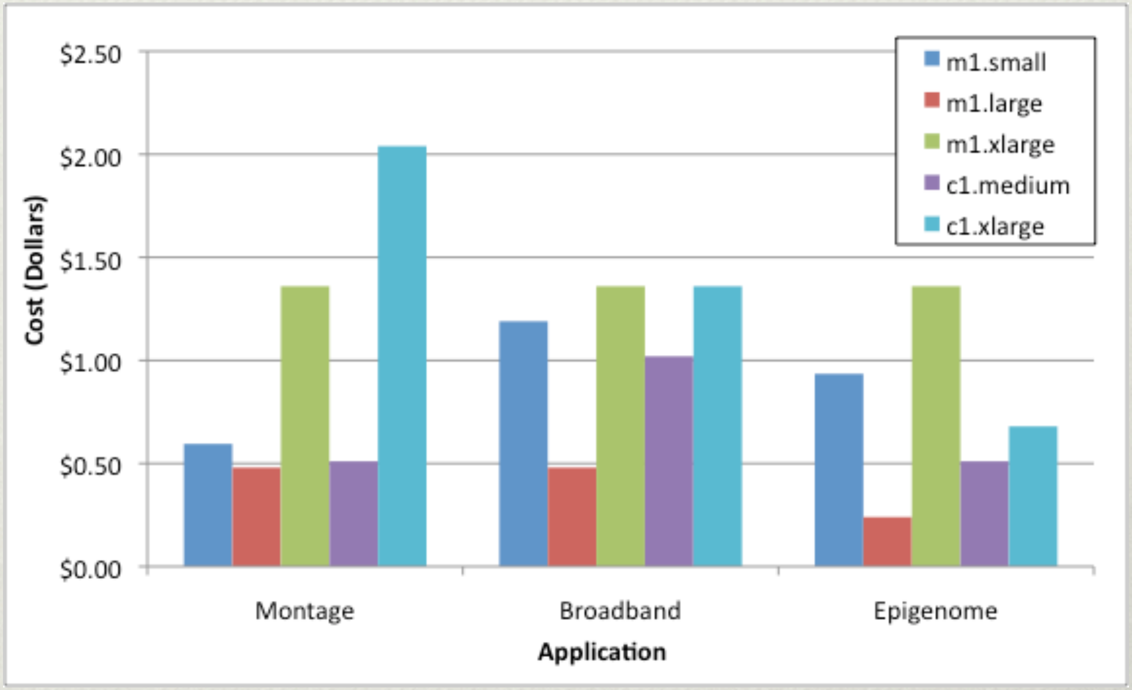
Performance Results



- Virtualization Overhead <10%
- Large differences in performance between the resources and between the applications
- The parallel file system on *abe.lustre* offers a big performance advantage of x3 for Montage

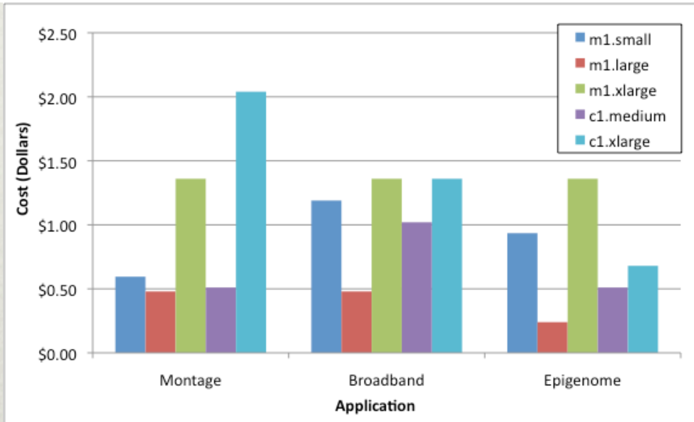
How Much Did It Cost?





Instance	Cost $/hr
m1.small	0.10
m1.large	0.40
m1.xlarge	0.80
c1.medium	0.20
c1.xlarge	0.80

- Clear trade-off between performance and cost.
- **Most powerful processor *c1.xlarge* offers 3x the performance of *m1.small* – but at 4x the cost.**
- Most cost-effective processor for Montage is *c1.medium* – 20% performance loss over *m1.small*, but 5x lower cost.



Data Transfer Costs

Operation	Cost $/GB
Transfer In	0.10
Transfer Out	0.17

Transfer Rates

- ❖ Amazon charges different rates for transferring data into the cloud and back out again.
- ❖ Transfer-out costs are the higher of the two.

Application	Input (GB)	Output (GB)	Logs (MB)
Montage	4.2	7.9	40
Broadband	4.1	0.16	5.5
Epigenome	1.8	0.3	3.3

Application	Input	Output	Logs	Total
Montage	$0.42	$1.32	<$0.01	$1.75
Broadband	$0.40	$0.03	<$0.01	$0.43
Epigenome	$0.18	$0.05	<0.01	$0.23

Transfer Costs

- ❖ For Montage, the **cost to transfer data out of the cloud is higher** than monthly storage and processing costs.
- ❖ For Broadband and Epigenome, **processing incurs the biggest costs**.

Data Storage Charges

Storage Costs

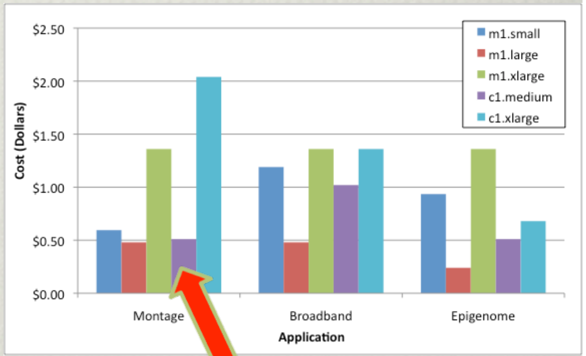
- Amazon charges for storing Virtual Machines (VM) and user's applications in local disk
- It also charges for storing data in persistent network-attached Elastic Block Storage (EBS).

Storage Rates

Item	Charges $
Storage of VM's in local Disk (S3)	0.15/GB-Month
Storage of data in EBS disk	0.10/GB-Month

Storage Volumes

Application	Input (GB)	Output (GB)	Logs (MB)
Montage	4.2	7.9	40
Broadband	4.1	0.16	5.5
Epigenome	1.8	0.3	3.3



Storage Costs

Application	Data ($)	VM ($)	Monthly Cost ($)
Montage	$0.95	$0.12	$1.07
Broadband	$0.02	$0.10	$0.12
Epigenome	$0.22	$0.10	$0.32

Montage Storage Costs Exceed Most Cost-Effective Processor Costs

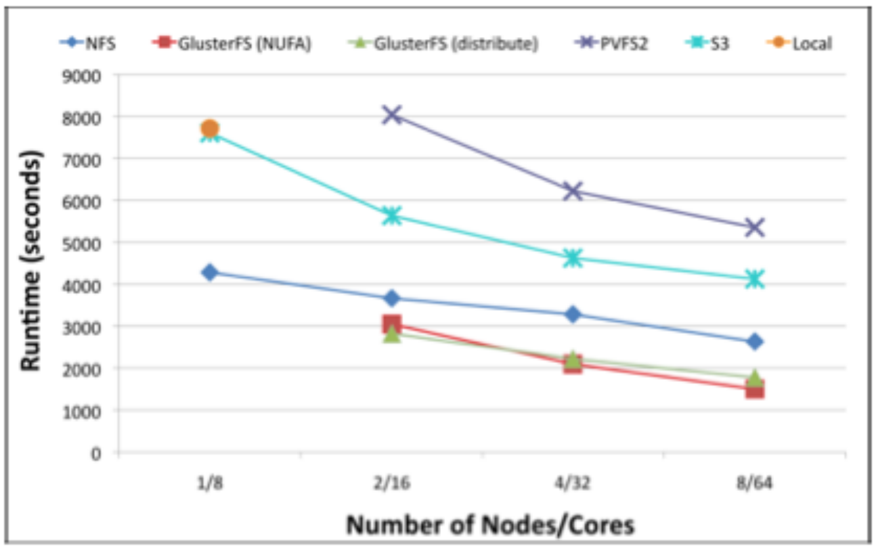
The bottom line for Montage

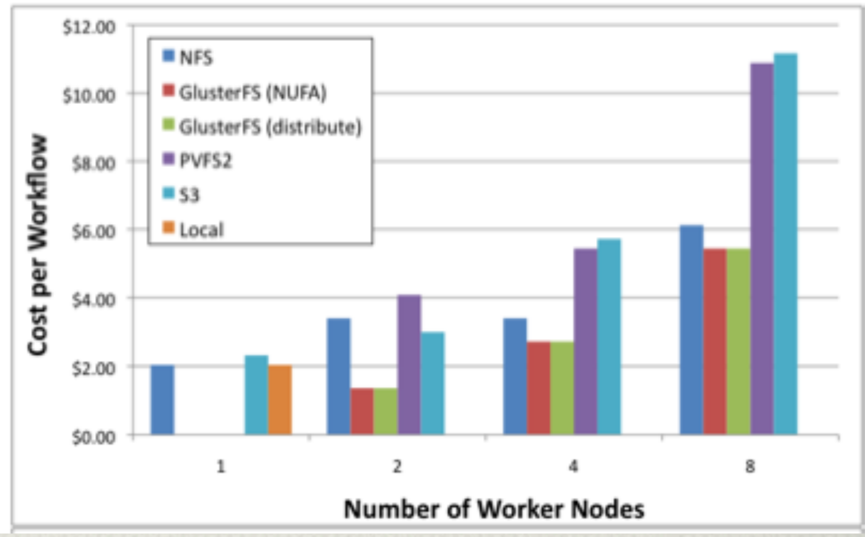
Item	Best Value	Best Performance
	c1.medium	*c1.xlarge*
Transfer Data In	$ 0.42	$ 0.42
Processing	$ 0.55	$ 2.45
Storage/month	$ 1.07	$ 1.07
Transfer Out	$ 1.32	$ 1.32
Totals	$ 3.36	$ 5.26

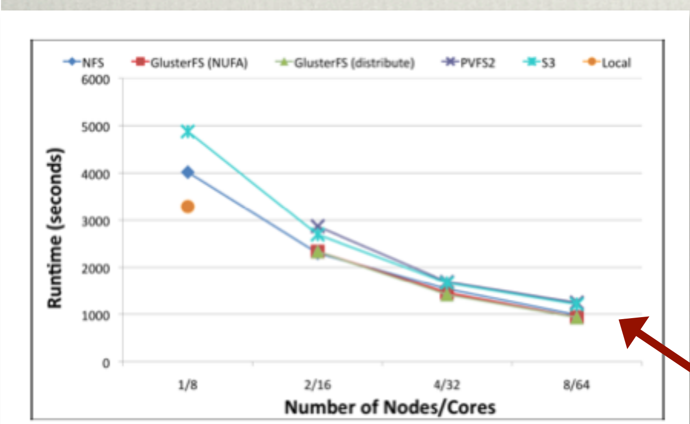
4.5x the processor cost for 20% better performance

Just To Keep It Interesting …

Running the Montage Workflow With Different File Storage Systems







Cf. Epigenome

Cost and performance vary widely with different types of file storage dependence on how storage architecture handles lots of small files

Cost-Effective Mosaic Service

Local Option

Item	Cost ($)
12 TB RAID 5 disk farm and enclosure (3 yr support)	12,000
Dell 2650 Xeon quad–core processor, 1 TB staging area	5,000
Power, cooling and administration	6,000
Total 3-year Cost	**23,000**
Cost per mosaic	**0.64**

-2MASS image data set
- 1,000 x 4 square degree mosaics/month

Amazon EBS Option

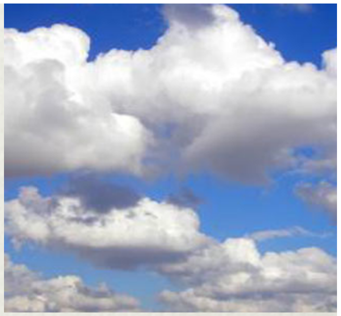
Item	Price ($)
Input Transfer (10 TB)	1,024.00
Output Transfer (24 TB)	3,691.41
Storage (10 TB)	36,864.00
Input I/O (10 TB)	24.58
Output I/O (27 TB)	67.50
Compute (c1.medium)	4,467.60
Total	**46,139.08**
Cost per mosaic	**1.28**

Amazon S3 Options

Item	Price ($)
Input Transfer (10 TB)	1,024.00
Output Transfer (24 TB)	3,691.41
PUT Ops (5.24 M)	52.43
GET Ops (14.4 M)	14.40
Compute (c1.medium)	4,467.60
Normal Storage (10 TB)	55,296.00
Total w/ Normal Storage	**64,545.84**
Cost per mosaic (Normal)	**1.79**
Reduced Redundancy Storage (10TB)	36,864.00
Total w/ RR Storage	**46,113.84**
Cost per mosaic (RR)	**1.28**

Amazon cost is 2X local!

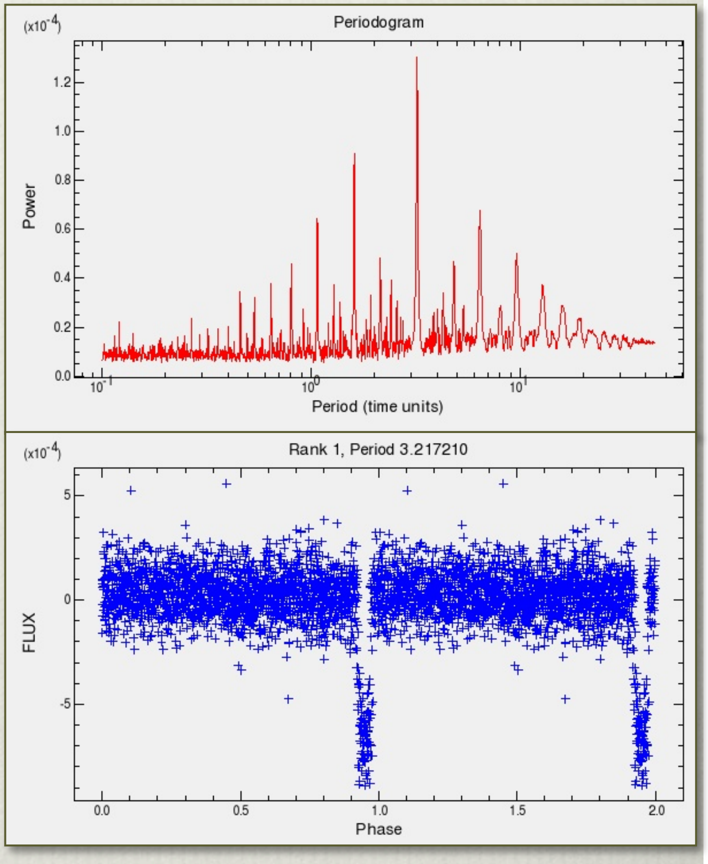
When Should I Use The Cloud?



- **The answer is….it depends on your application and use case.**
- **Recommended best practice: Perform a cost-benefit analysis to identify the most cost-effective processing and data storage strategy. Tools to support this would be beneficial.**
- Amazon offers the best value
  - For compute- and memory-bound applications.
  - For one-time bulk-processing tasks, providing excess capacity under load, and running test-beds.
- Parallel file systems and high-speed networks offer the best performance for I/O-bound applications.
- Mass storage is **very** expensive on Amazon EC2

Periodograms and the Search for Exoplanets

- ❖ What is a periodogram?
  - ❖ Calculates the significance of different frequencies in time-series data to identify periodic signals.
  - ❖ Powerful tool in the search for **exoplanets**
- ❖ NStED Periodogram tool
  - ❖ Computes periodograms using 3 algorithms: Box Least Squares, Lomb-Scargle, Plavchan
  - ❖ Fast, portable implementation in C
  - ❖ Easily scalable: each frequency sampled independently of all other frequencies
  - ❖ Implemented a NStED on 128-node cluster.



The Application of Cloud Computing to Astronomy: A Study of Cost and Performance. *Berriman et al. 2010.* ***http://arxiv.org/abs/1006.4860***

http://nsted.ipac.caltech.edu/periodogram/cgi-bin/Periodogram/nph-simpleupload

Kepler Periodogram Atlas

- Compute periodogram atlas for public Kepler dataset
  - ~200K light curves X 3 algorithms X 3 parameter sets
  - Each parameter set was a different "Run", 3 runs total
  - Use 128 prrocessor cores in parallel

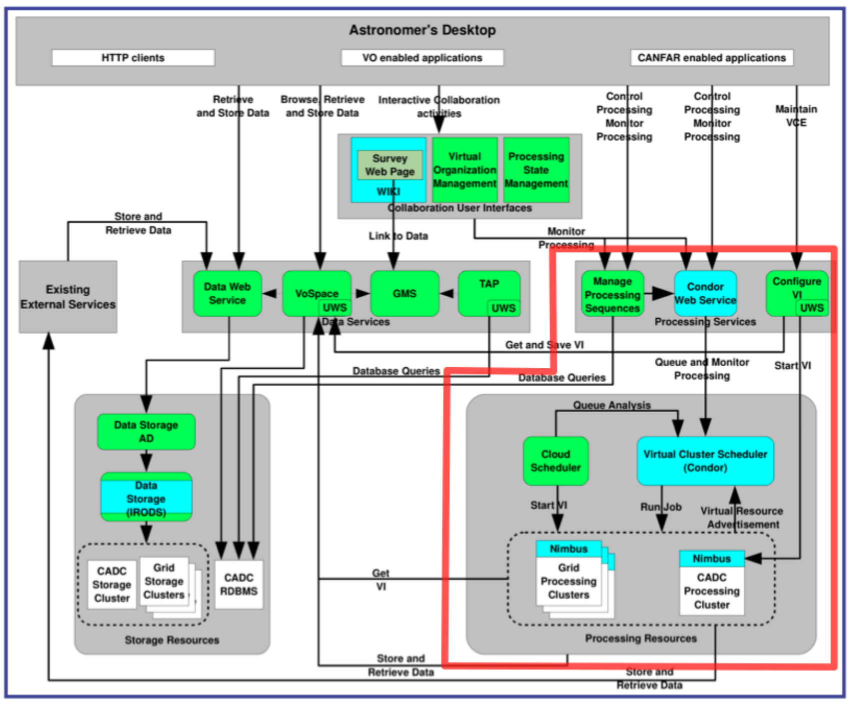
		Run 1 (EC2)	Run 2 (EC2)	Run 3 (TeraGrid)
Runtimes	Tasks	631992	631992	631992
	Mean Task Runtime	7.44 sec	6.34 sec	285 sec
	Jobs	25401	25401	25401
	Mean Job Runtime	3.08 min	2.62 min	118 min
	Total CPU Time	1304 hr	1113 hr	50019 hr
	Total Wall Time	16.5 hr	26.8 hr	448 hr
Inputs	Input Files	210664	210664	210664
	Mean Input Size	0.084 MB	0.084 MB	0.084 MB
	Total Input Size	17.3 GB	17.3 GB	17.3 GB
Outputs	Output Files	1263984	1263984	1263984
	Mean Output Size	0.171 MB	0.124 MB	5.019 MB
	Total Output Size	105.3 GB	76.52 GB	3097.87 GB
Cost	Compute Cost	$179.52	$94.61	$4,874.24
	Output Cost	$15.80	$11.48	$464.68
	Total Cost	**$195.32**	**$106.08**	**$5,338.92**

Compute is ~10X Transfer

Estimated cost

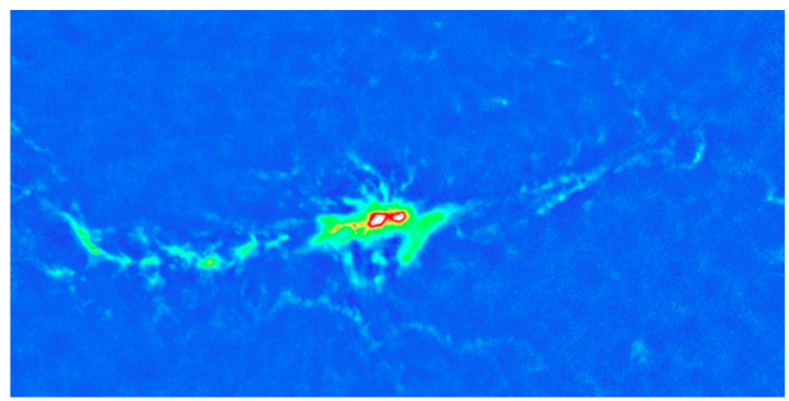
Should We All Move To The Cloud?

"The Canadian Advanced Network For Astronomical Research (CANFAR) is an operational system for the delivery, processing, storage, analysis, and distribution of very large astronomical datasets. The goal of CANFAR is to support large Canadian astronomy projects."



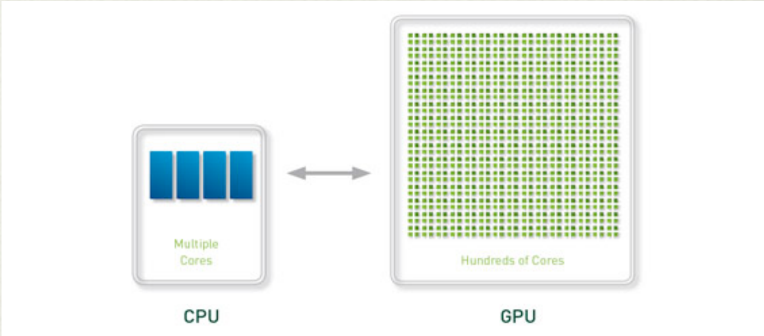
Project Name	Project Lead	Institution	Telescope
SCUBA-2 Cosmology Legacy Survey	Mark Halpern	UBC	JCMT
SCUBA-2 All-Sky Survey	Douglas Scott	UBC	JCMT
Next Generation Virgo Survey	Laura Ferrarese	NRC-HIA	CFHT
Pan-Andromeda Archaeological Survey	Harvey Richer	UBC	CFHT
Time Variable Sky	Chris Pritchet	UVic	CFHT
Canada-France Ecliptic Plane Survey Simulator	JJ Kavelaars	NRC-HIA	CFHT
Shapes and Photometric Redshifts for Large Surveys	Ludo Van Waerbeke	UBC	CFHT

Projects currently using the CANFAR processing system



Dust emission from Orion Molecular Cloud at 850um processed using CANFAR (JCMT SCUBA-2 map thanks to Ed Chapin and Mark Halpern, UBC)

GPU's In Astronomy



- GPU invented to accelerate building of images in a frame buffer as an output on a display device.

- Consist of many floating point processor cores

- Highly parallel structure makes them attractive for processing huge blocks of data in parallel.

- In early days, apps had look like video apps, but there are now frameworks to support application development: CUDA, Open GL

What Types of Applications Do We Run on GPU's?

Barsdell, Barnes and Fluke (2010) have analyzed astronomy algorithms to understand which types are best suited to running on GPU's. (*arxiv.org/abs/1007.1660*)

Field	High efficiency	Moderate efficiency
Simulation	• Direct N-body • Fixed-resolution mesh simulations • Semi-analytic modelling • Gravitational lensing ray-shooting • Other Monte-Carlo methods	• Tree-code N-body and SPH • Halo-finding • Adaptive mesh refinement
Data reduction	• Radio-telescope signal correlation • General image processing • Flat-fielding etc. • Source extraction • Convolution and deconvolution	• Pulsar signal processing • Stacking/mosaicing • CLEAN algorithm • Gridding visibilities and single-dish data
Data analysis	• Machine learning • Fitting/optimisation • Numerical integration • Volume rendering	• Selection via criteria-matching

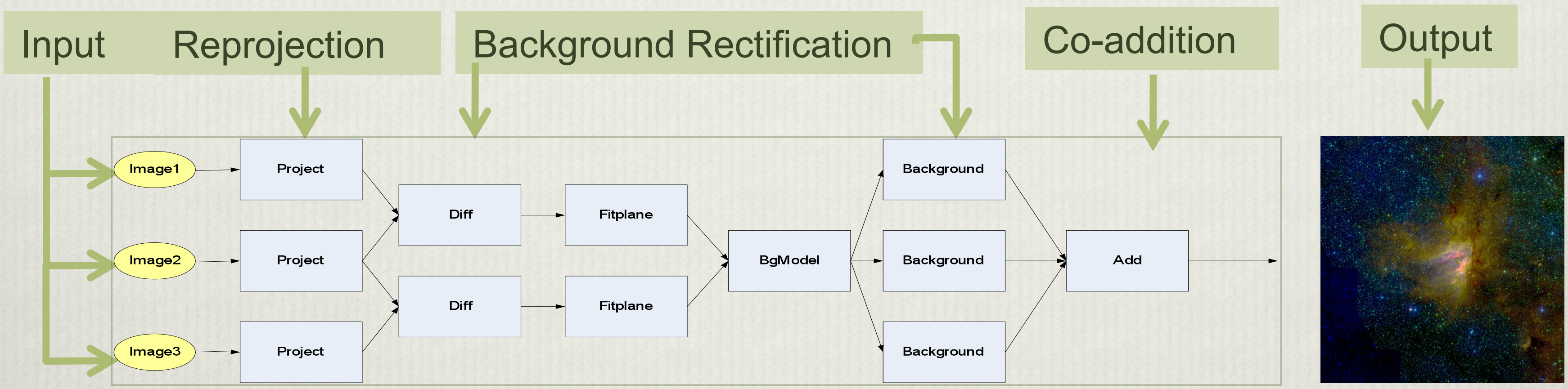
"CPU's handle complexity, GPU's handle concurrency"

- ❖ Can be parallelized into many fine-grained elements.
- ❖ Neighboring threads access similar locations in memory.
- ❖ Minimize neighboring threads that execute different instructions.
- ❖ Have high arithmetic intensity
- ❖ Avoid host-device memory transfers

"Critical Decisions For Early Adopters"

- Title of a paper by Fluke et al (2010) on Astrophysical Supercomputing with GPU's. (*arxiv.org/abs/1008.4623*)
- Suggest brute-force parallelization may be highly competitive with algorithmic complexity.
  - Development times can be reduced with brute-force approach.
- GPU's support single precision calculations, but astronomy often needs double precision.
- Need to understand architecture to get speed-ups of x100
  - Speeds quoted are for graphics-like calculations
- Code profiling will very likely help code optimization

What Have We Learned About "Next Generation" Code?

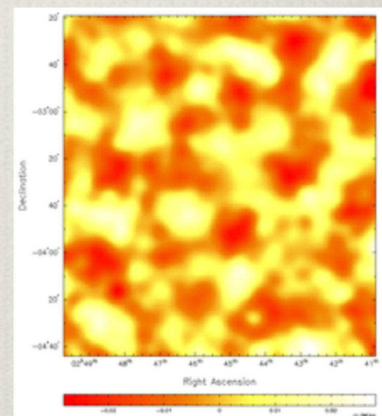


Montage Workflow

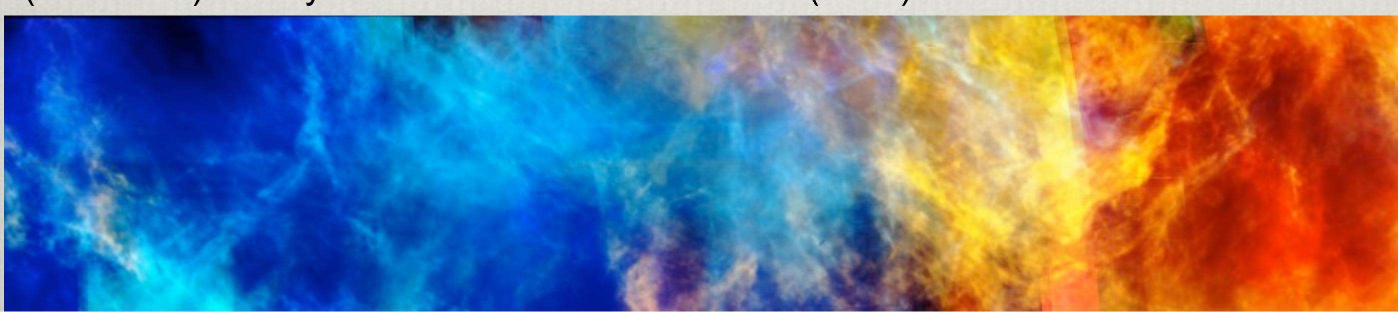
- Downloaded 5,000 times with wide applicability in astronomy and computer science.
- Simple to build.
- Written in ANSI-C for performance and portability.
- Portable to all flavors of *nix
- Developed as a component-based toolkit for flexibility.
- Environment agnostic
- Naturally "data parallel"
- Technology Agnostic: Supports tools such as Pegasus, MPI, .. Same code runs on all platforms.

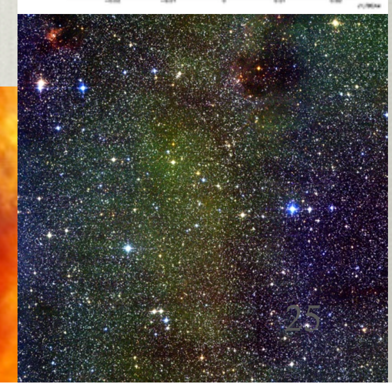
Applications of Montage: Science Analysis

- Desktop research tool – astronomers now sharing their scripts
- Incorporation into pipelines to generate products or perform QA.
  - Spitzer Space Telescope Legacy teams
  - Cosmic Background Imager
  - ALFALFA
  - BOLOCAM



1,500-square-degree-equal-area Aitoff projection mosaic, of HI observed with (ALFALFA) survey near the North Galactic Pole (NGP). *Dr Brian Kent*





Applications of Montage: Computational Infrastructure

- Task scheduling in distributed environments (performance focused)
- Designing job schedulers for the grid
- Designing fault tolerance techniques for job schedulers
- Exploring issues of data provenance in scientific workflows
- Exploring the cost of scientific applications running on Clouds
- Developing high-performance workflow restructuring techniques
- Developing application performance frameworks
- Developing workflow orchestration techniques

List kindly provided by Dr. Ewa Deelman

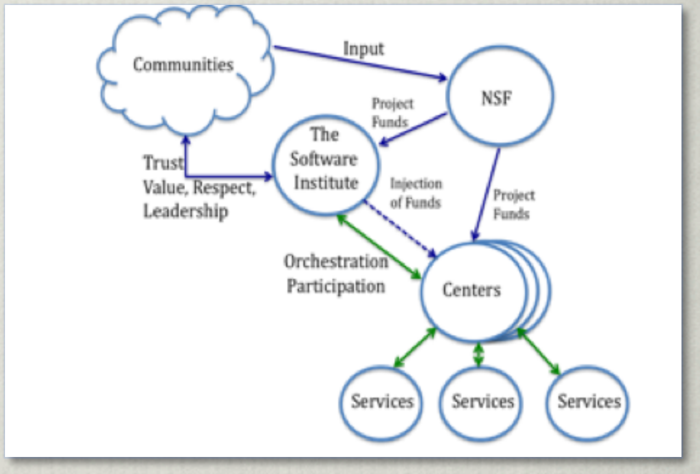
What Are The Next Steps?

- Greater recognition of the role of software engineering
  - Provide career-paths for IT professionals.
  - Next generation software skills should be a mandatory part of graduate education.
- An on-line journal devoted to computational techniques in astronomy.
- Share computational knowledge from different fields and take advantage of it.

A U.S. Software Sustainability Institute: A Brain Trust For Software

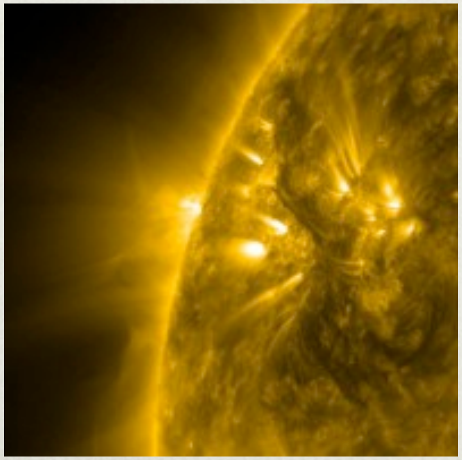
"A US Software Infrastructure Institute that provides a national center of excellence for community based software architecture, design and production; expertise and services in support of software life cycle practices; marketing, documentation and networking services; and transformative workforce development activities."

Report from the *Workshops on Distributed Computing, Multidisciplinary Science, and the NSF's Scientific Software Innovation Institutes Program* Miron Livny, Ian Foster, Ruth Pordes, Scott Koranda, JP Navarro. August 2011



U.K. Software Sustainability Institute

http://www.software.ac.uk

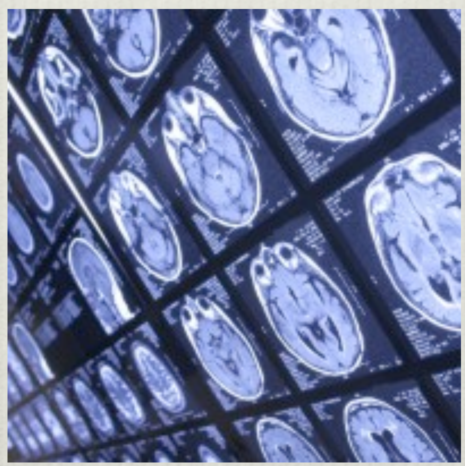


Nuclear Fusion - Culham Centre for Fusion Energy

Pharmacology - DMACRYS

Climate change - Enhancing Community Integrated Assessment

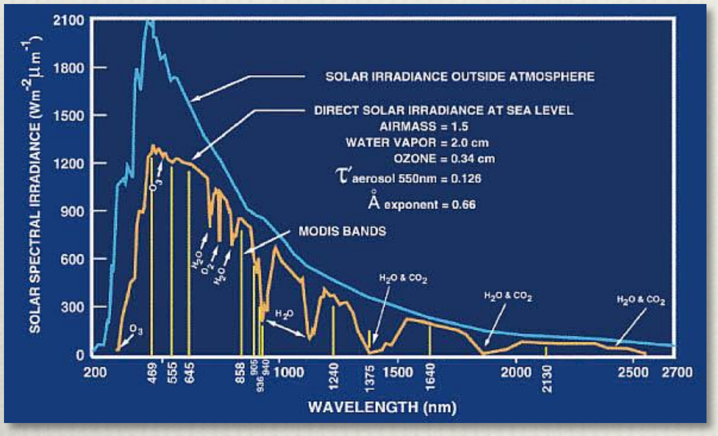
Geospatial Information - Geospatial transformations with OGSA-DAI



Scottish Brain Imaging Research Centre

Keeping up to date with research

The Moderate Resolution Imaging Spectroradiometer (MODIS)

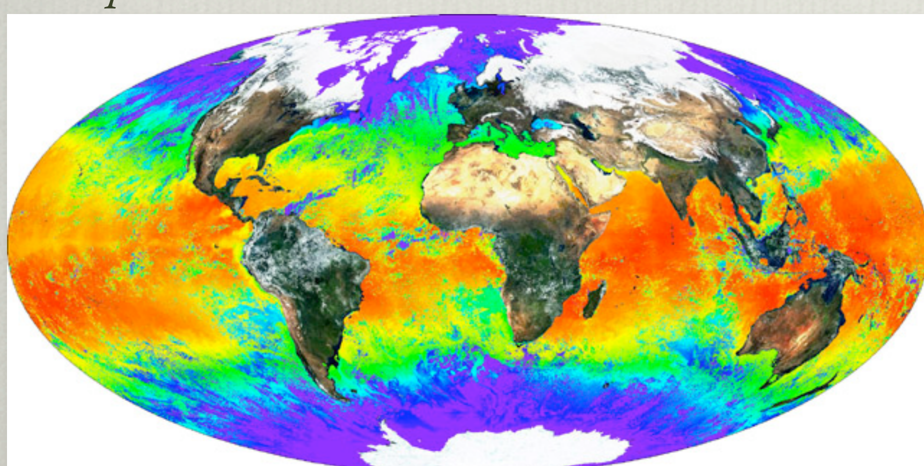


- ❖ Science products created by aggregating calibrated products in various bands
- ❖ Calibrated data kept for 30-60 days (size) and so:
- ❖ MODIS maintains a *virtual archive* of the provenance of the data and processing history that enables reproduction of any science product

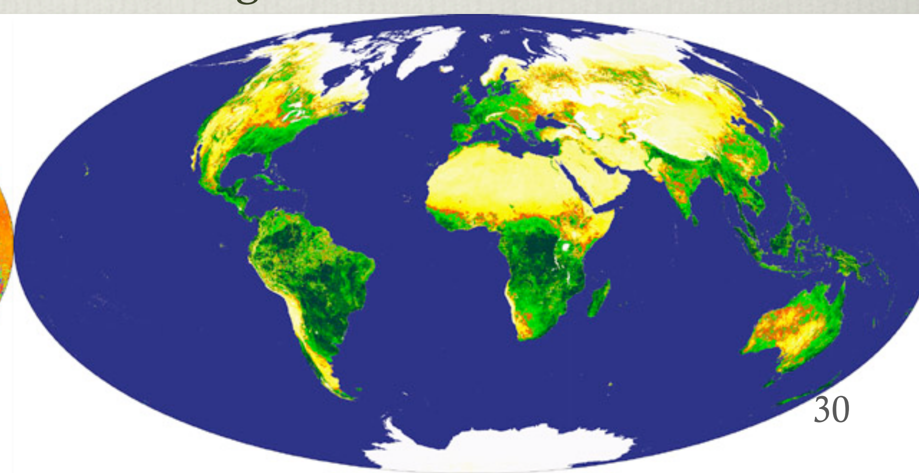
Scans Earth every 2 days in 36 bands

Application of Cloud Computing to the Creation of Image Mosaics and Management of Their Provenance, Berriman et al. arxiv.org/abs/1006.4860

Global Surface Reflectance and Sea Surface Temperature



Global Vegetation Index

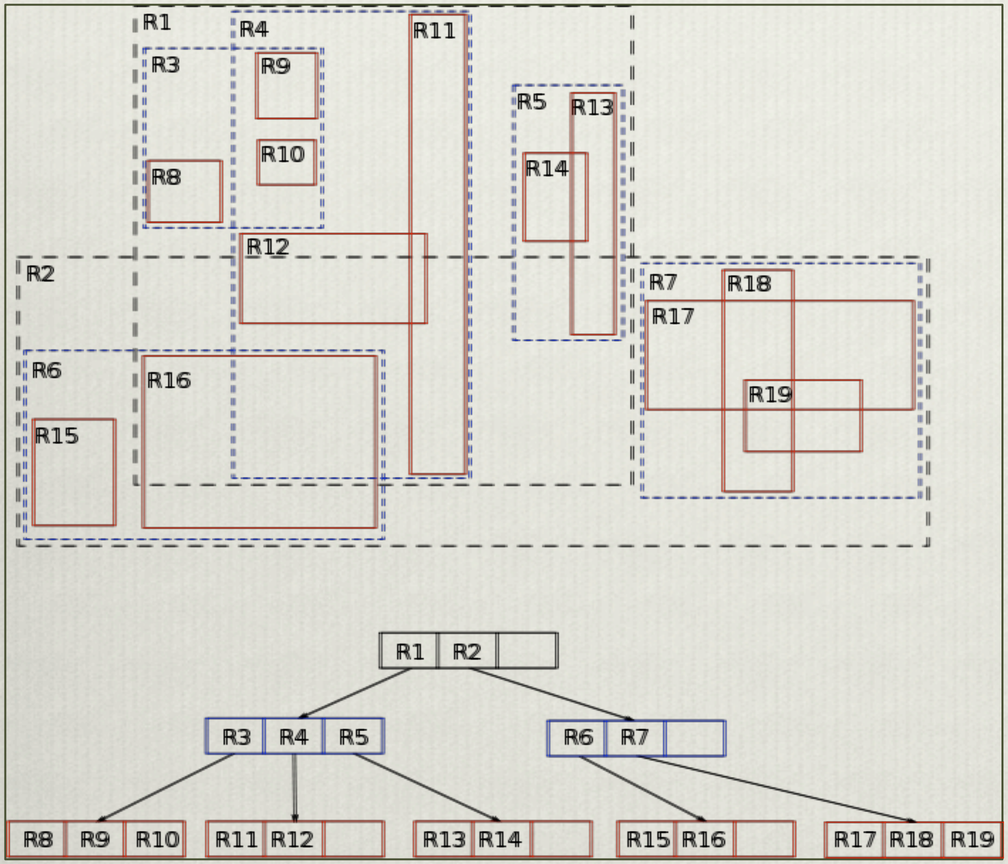


What Are The Next Steps?

- The VAO can play a big role in providing sharable, scalable software for the community.
- From the VAO's Expected Outcomes:
  - "The VAO's services and libraries, developed to respond to **the growing scale and complexity of modern data sets**, will be indispensable tools for astronomers integrating data sets and creating new data sets."
  - "The VAO will collaborate and cooperate with missions, observatories and new projects, who will be able to routinely integrate VAO libraries into their processing environments to simplify and accelerate the development and dissemination of new data products."

\- *VAO Program Execution Plan, version 1.1 (Nov 2010)*

VAO Inventory: R-tree Indexing



- Fast searches over very large and distributed data sets
- Performance scales as log(N)
- Performance gain of x1000 over table scan
- Used in Spitzer and WISE image archives

- Memory-mapped files
- Parallelization / cluster processing
- REST-based web services

Segment of virtual memory is assigned a byte for byte correlation with part of a file.

Where Can I Learn More?

- **Scientific Workflow Applications on Amazon EC2.** G. Juve et al. Cloud Computing Workshop in Conjunction with e-Science 2009 (Oxford, UK). **http://arxiv.org/abs/1005.2718**
- **Data Sharing Options for Scientific Workflows on Amazon EC2**, G. Juve et al. Proceedings of Supercomputing 10 (SC10), 2010. **http://arxiv.org/abs/1010.4822**
- **The Application of Cloud Computing to the Creation of Image Mosaics and Management of Their Provenance**, G. B. Berriman, et al. SPIE Conference 7740: *Software and Cyberinfrastructure for Astronomy*. 2010. **http://arxiv.org/abs/1006.4860**
- **The Application of Cloud Computing to Astronomy: A Study of Cost and Performance.** G. B. Berriman et al. 2010. Proceedings of "e-Science in Astronomy" Workshop. Brisbane. **http://arxiv.org/abs/1006.4860**
- **Astrophysical Supercomputing with GPUs: Critical Decisions for Early Adopters.** Fluke et al. 2011. PASA Submitted. **http://arxiv.org/abs/1008.4623.**
- **Analysing Astronomy Algorithms for GPUs and Beyond.** Barsdell, Barnes and Fluke. 2010. Submitted to MNRAS. **http://arxiv.org/abs/1007.1660**
- Bruce Berriman's blog, "Astronomy Computing Today," at **http://astrocompute.wordpress.com**