National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology

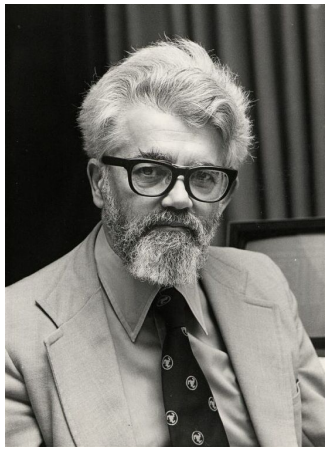
NASA Exoplanet Science Institute

Cloud Computing and Exoplanets

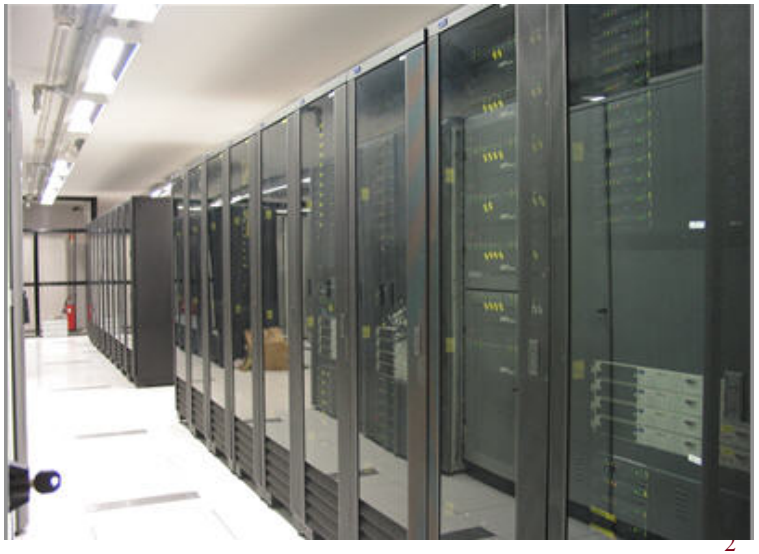
Bruce Berriman

May 23, 2016

What Is Cloud Computing?



- A new way of purchasing computer power and storage. **Pay only for what you use.**
- John McCarthy … "computation delivered as a public utility in…. the same way as water and power." (1963!)
- Uses commodity hardware
- Uses virtualization technologies.



Getting Started With Cloud Computing

All you need is a credit card!

This looks cheap!

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		

"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

… 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…

Annual Costs!

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

Characteristics of Workflows

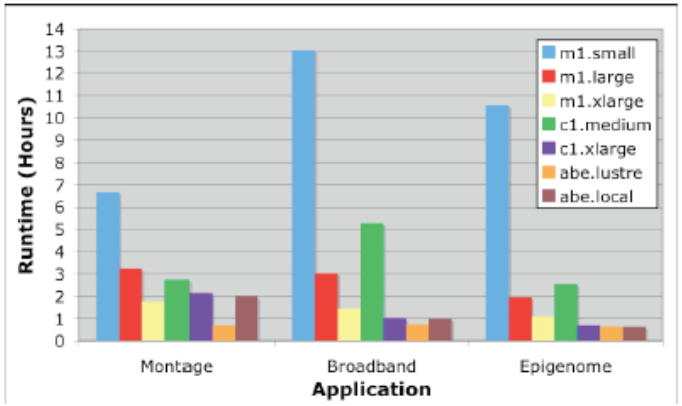
Workflow Specifications for this Study

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

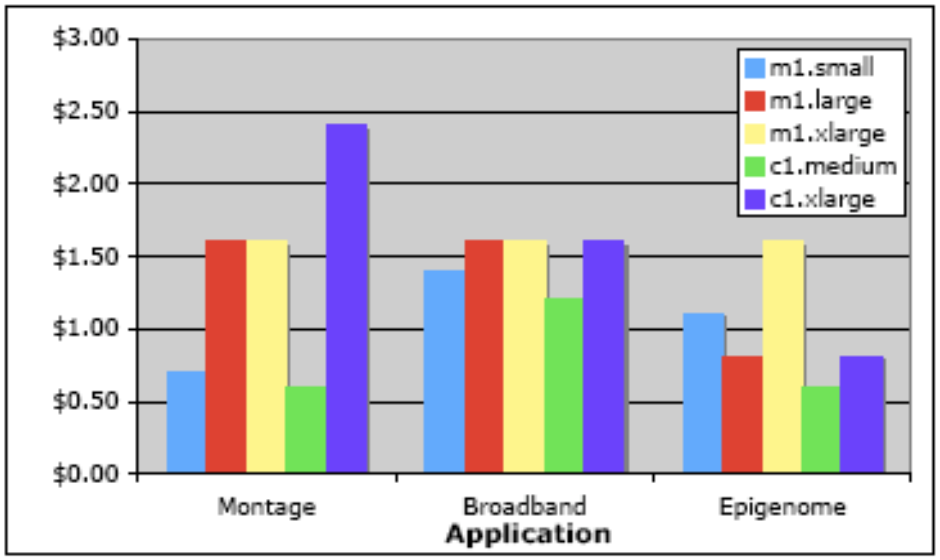
How Much Was The Processing?



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

Broadband and Epigenome:

- Choose the most powerful machines.



Montage:

- Trade-off between performance and cost.
- **Most powerful processor *c1.xlarge* offers 3x the performance of *m1.small* – but at 4.5x the cost.**
- Most cost-effective processor is *c1.medium* – 20% performance loss over *m1.small*, but 5x lower cost.

NASA Exoplanet Science Insti

Storage Costs

Data Storage Charges

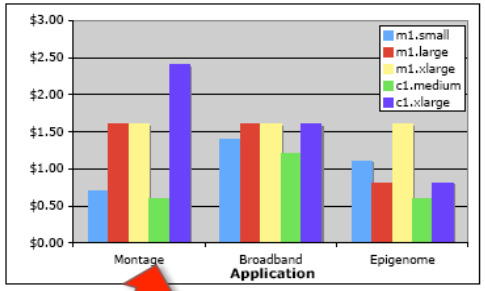
- Amazon charges for storing Virtual Machines (VM) and users applications in local disk
- It also charges for storing data in network-attached Elastic Block Storage (EBS).

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

When Should I Use The Cloud?

- **Recommended best practice: Perform a cost-benefit analysis to identify the most cost-effective processing and data storage strategy.**
- Amazon offers the best value
  - For **compute-** and **memory-bound** applications with predictable processing times.
  - For one-time bulk-processing tasks, providing excess capacity under load, and running test-beds.
- Amazon offers worst value
  - For mass storage

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology

"Working With Exoplanet Light Curves"

NASA Exoplanet Science Institute

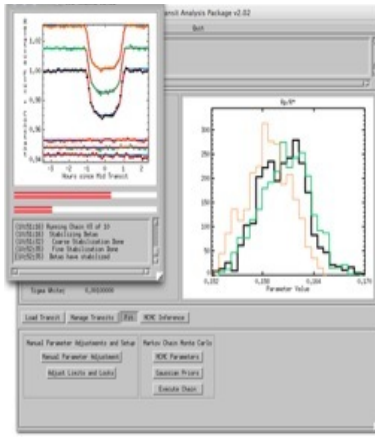
July 22-27, 2012. Pasadena, CA.

http://nexsci.caltech.edu/workshop/2012

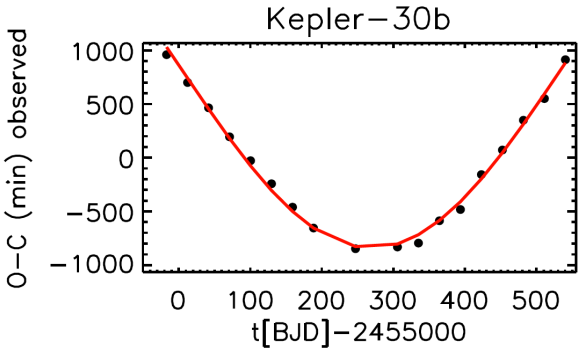
•Interactive Activities include **hands-on data sessions** such as working with Kepler data.

Applications Run In Different Environments

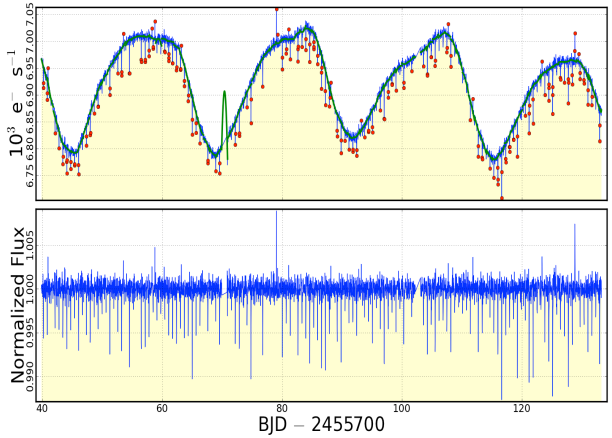
NASA Exoplanet Science Institute



- Transit Analysis Package
- **IDL 8.1 and 8.2.**
- Uses Markov Chain Monte Carlo techniques



- Transit Timing Variations
- **Java GUI**

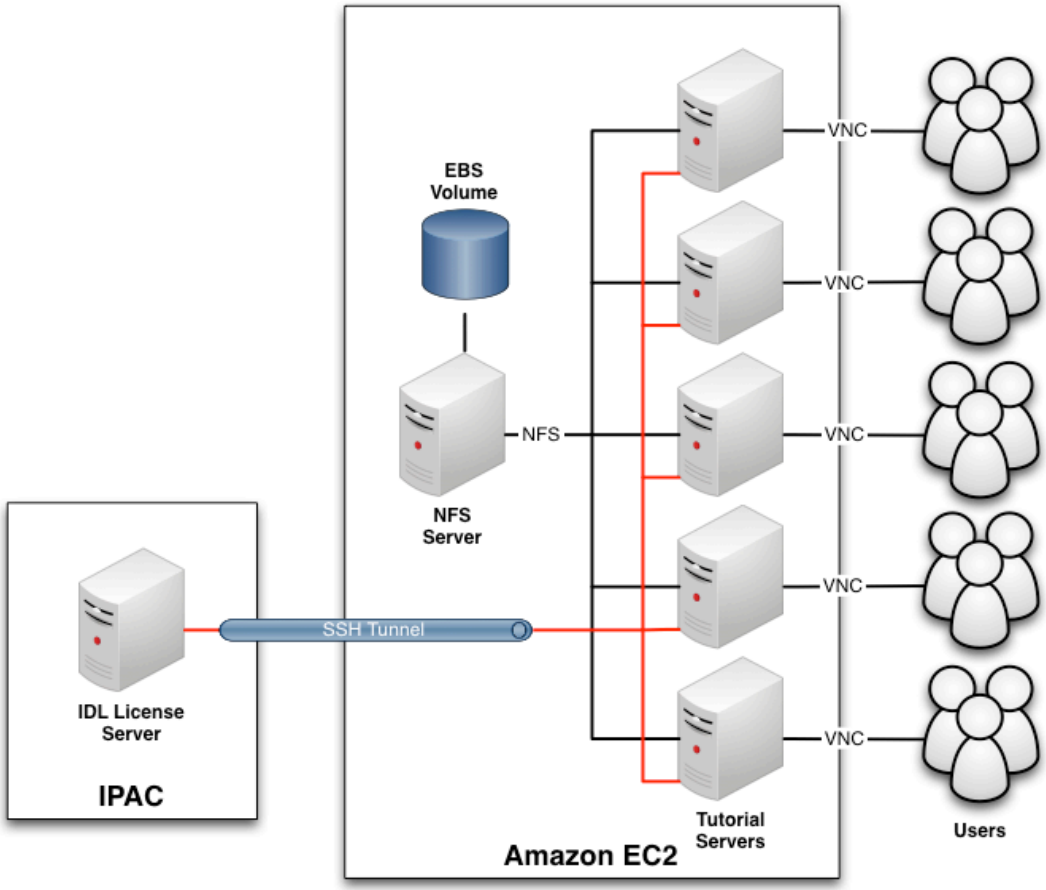


- PyKE
- **PyRAF** toolkit
- Toolkit for analyzing Kepler Data

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology

Simple System Design

NASA Exoplanet Science Institute



The Cost, Had We Paid For It …

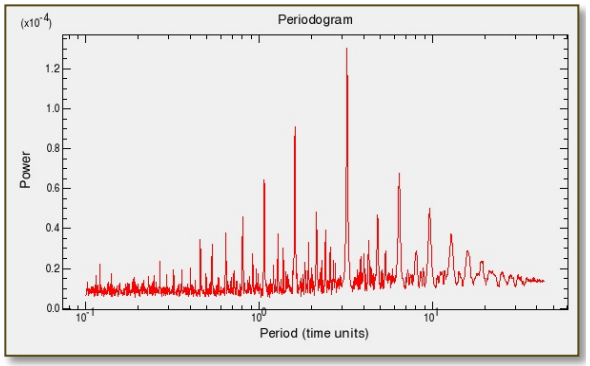
$2,876

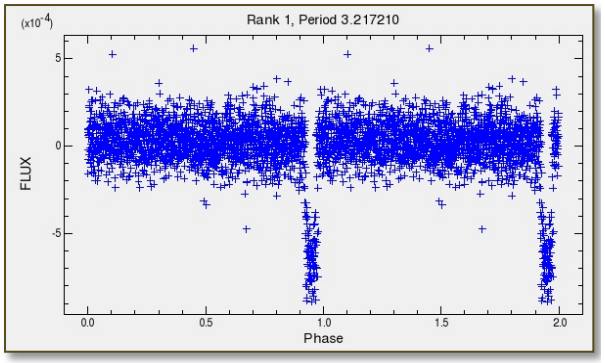
Set-up, Testing, Running the sessions

Resource	Consumed	Cost ($)
VM Instances	4,159 hours	2,738
EBS Storage	1.25 TB	126
I/O requests	12 million	1
Snapshot data storage	22 GB	3
Elastic IP addresses	604 hours	3
Data Transfer	55 GB	5
Total		**2,876**

Digging Out Exoplanets with Periodograms

- A **periodogram** calculates the significance of different frequencies in time-series data to identify periodic signals.
- *NASA Star and Exoplanet Database* Periodogram tool
  - Fast, portable implementation in C
  - Easily scalable: each frequency sampled independently of all other frequencies
- Calculations are slow: 1 hour for 100,000-200,000 points typical of Kepler light curves.
- How can we process the entire data set? Candidate for the cloud: "high-burst," processor-bound, easily parallelizable.





National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology

Kepler Periodograms

NASA Exoplanet Science Institute

Compute periodogram atlas for public Kepler dataset

- Use 128 processor cores in parallel on Amazon EC2 and TeraGrid
- ~210K light curves X 3 algorithms

Run	Algorithm	Optimization
1 (EC1)	Lomb-Scargle	Sinusoids
2 (EC1)	Box-Least Squares	Box
3 (TG)	Plavchan	Unrestricted

		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

Periodograms on Academic Clouds

Site	CPU	RAM (SW)	Walltime	Cum. Dur.	Speed-Up
Magellan	8 x 2.6 GHz	19 GB	5.2 h	226.6 h	43.6
Amazon	8 x 2.3 GHz	7 GB	7.2 h	295.8 h	41.1
FutureGrid	8 x 2.5 GHz	29 GB	5.7 h	248.0 h	43.5

- 33 K periodograms with Plavchan algorithm
- Given 48 physical cores
  - Speed-up ≈ 43 considered *good*
  - AWS cost ≈ $31:
    - 7.2 h x 6 x c1.large ≈ $29
    - 1.8 GB in + 9.9 GB out ≈ $2
- Results encouraging.

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology

NASA Exoplanet Archive Periodogram

NASA Exoplanet Science Institute

NASA Exoplanet Science Institute

- The periodogram is one of the most utilized tools at the NASA Exoplanet Archive (~10,000 calls/month)
- Currently runs on a 8+ year old cluster
- Ideal task for cloud computing: CPU intensive, predictable run times
- Status: Have a code version that runs on Amazon cloud, working on job management