

## HPC Clusters in the [almost]\* Infinite cloud

Brendan Bouffler (@boofla), #scico

WW Research & Technical Computing



# Scientific Computing



**Science** is one of the greatest areas of computation and can benefit from a democratization in cost and global accessibility that the cloud brings.

It's also where we think **Amazon** can make a **huge, really disruptive, impact** on the world by participating - which is, at the most basic level, **what we are about** as a company.

# AWS Research & Technical Computing Team



The **Research & Technical Computing** team is a global group of scientists and specialists from Amazon Web Services.

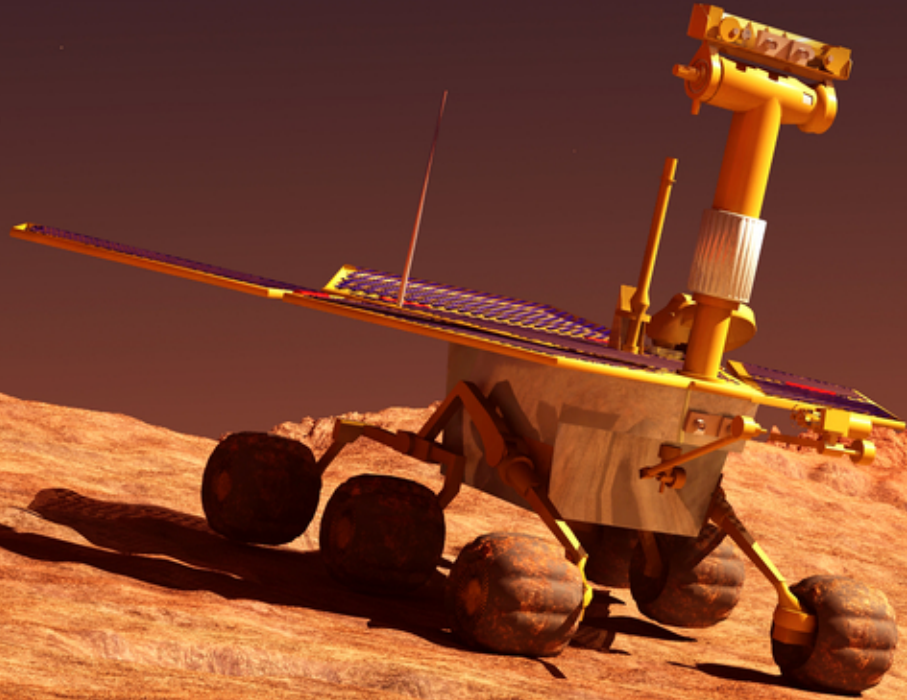
We're responsible for making the sure the cloud continually innovates in ways that **benefit the global community** of researchers from whom we draw our inspiration.

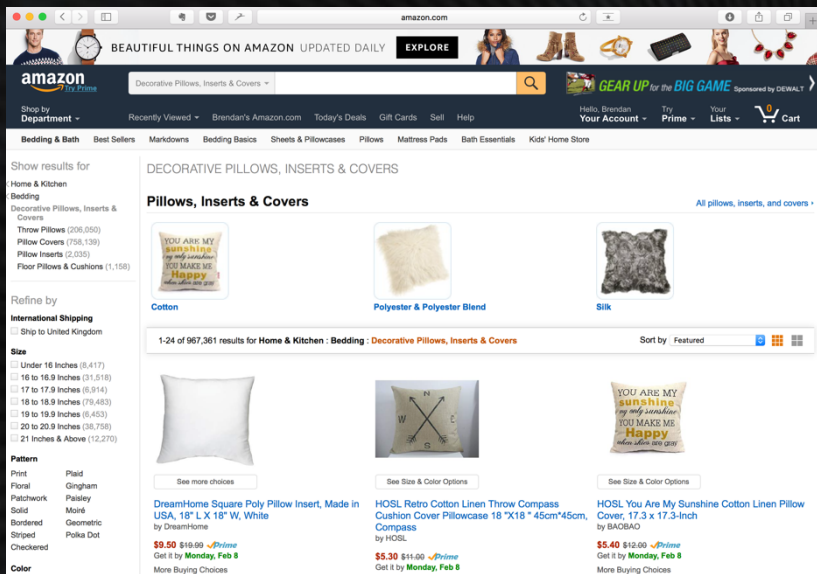
Our aim is to bring the revolutionary benefits of **agility and extreme scale** to this community so we can all keep making the discoveries that will change the world and impact the lives of everyone on our planet.

We have team members from **physics, astronomy,** aeronautical engineering, and **genomics** and all have extensive experience in research and high performance computing. **We even have a rocket scientist.**



**Disrupting science, wherever it's happening.**





“... the online book and decorative pillow seller Amazon.com swooped in and, in 2006, launched its own computer rental system—the future Amazon Web Services. The once-fledgling service has since turned cloud computing into a mainstream phenomenon ...”

Source: Bloomberg Business - April 22, 2015

# 2006 2016

amazon.com



\$7B retail business  
10,000 employees

A whole lot of  
servers

We deploy the  
equivalent of a  
top500  
supercomputer  
every day.



# Global AWS Regions



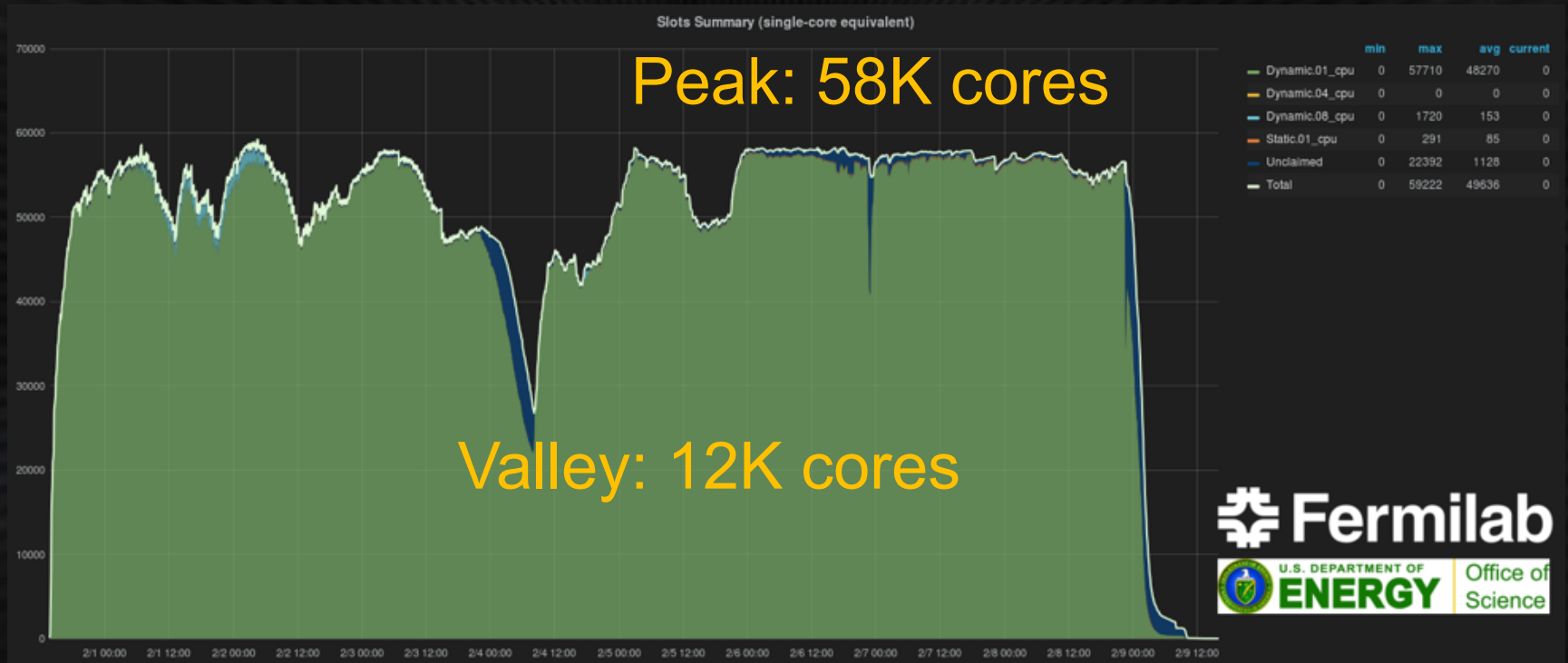
- Region & Number of Availability Zones
- New Region Coming Soon

**AWS Region** = A cluster of Availability Zones  
**Availability Zone** = A cluster of data centers

All regions are sovereign, meaning your data never leaves that location unless you cause it to.



# Agility is...Paying Only for IT You Use



# Science means Collaboration





# Collaboration is easier in the cloud



More time spent **computing the data** than **moving the data**.



# Public Data Sets

## 1000 Genomes Project and AWS

The 1000 Genomes Project is an international research effort coordinated by a consortium of 75 companies and organizations to establish the most detailed catalogue of human genetic variation. The project has grown to 2001

can now of more remainii

The data The data

Access

AWS is m centralize AWS ser, organizat public da or analysi

All 200 TE

You can t and PHP.

Analys

Research work with available

## TCGA on AWS

The [Cancer Genome Atlas \(TCGA\)](#) is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to accelerate our understanding of the molecular basis of cancer. TCGA-funded researchers across the United States have produced a corpus of raw and processed genomic, transcriptomic, and epigenomic data from thousands of cancer patients.

These data are now freely available on AWS via the National Cancer Institute's [Cancer Genomics Cloud pilot](#) to credentialed researchers subject to NIH data sharing policies. As the [NIH Trusted Partner](#) for this project, Seven Bridges Genomics is responsible for authorizing access to the data.

The Cancer Genome Atlas is one of the world's largest collections of cancer genome data available. Making the data available on a cloud platform greatly lowers the barrier to entry for researchers that are seeking to work with these data to create better models of disease, and ultimately develop new treatments for cancer. Qualified researchers can use the data on-demand without worrying about download time or storage costs.

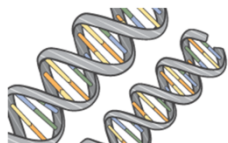
For more information, please visit <http://www.cancergenomicscloud.org/>. If you have any questions, please email [cgc@sbgenomics.com](mailto:cgc@sbgenomics.com).

### Accessing the Data

While the data are hosted within Amazon S3, access is currently only possible through the National Cancer Institute's [Cancer Genomics Cloud Pilot](#). Researchers wishing to access the TCGA controlled data must be registered within that system, and also be listed on an approved TCGA Data Access Request.

For more information on gaining accessing to these data, visit: <http://www.cancergenomicscloud.org/controlled-access-data> or <http://docs.cancergenomicscloud.org/>.

### Tools and Tutorials



### Project Updates

If you are interested in using the TCGA data or learning more about this project, please fill out the form below.

First Name\*

Last Name\*

Email Address\*

Job Role \*

Telephone\*

## NASA NEX

NASA NEX is a collaboration and analytical platform that combines state-of-the-art supercomputing, Earth system modeling, workflow management and NASA remote-sensing data. Through NEX, users can explore and analyze projects ar

Three NAS simulations data set, p Terra and / record from land.

Accessin

AWS is makir

- Simple HT

- AWS Com

- Amazon E

- Amazon E

The data is h

Available

Downscaled

charge

the U.S

invent

AWS h

Amaz

## Landsat on AWS

Landsat 8 data is available for anyone to use via Amazon S3. All Landsat 8 scenes from 2015 are available along with a selection of cloud-free scenes from 2013 and 2014. All new Landsat 8 scenes are made available each day, often within hours of production.

The [Landsat](#) program is a joint effort of the [U.S. Geological Survey](#) and [NASA](#). First launched in 1972, the Landsat series of satellites has produced the longest, continuous record of Earth's land surface as

seen fr

remote

launch

perform

system

satellite

archiviz

data h

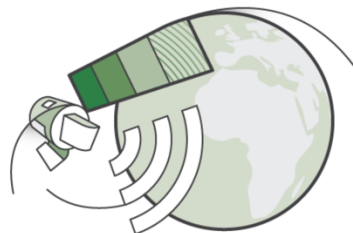
charge

the U.S

invent

AWS h

Amaz



## Sentinel-2 on AWS

Sentinel-2 data is available for anyone to use via Amazon S3.

About the data

Data structure

Browse through data

Accessing the Data

Featured uses

Contact us

Sentinel-2 data is available for anyone via Amazon S3, either over Internet or within AWS. All Sentinel-2 scenes are made available, often within hours of production.

Earth observation data provided by the [Sentinel-2](#) satellites are revolutionizing the market of space applications. Free, full and open access to data with very short revisit times, high spatial resolution, and good spectral resolution can benefit several sectors - agriculture, environmental and land-change monitoring, natural disaster response, insurance and others.

The Sentinel-2 mission is a land monitoring constellation of two satellites (Sentinel-2A was launched on 23 June 2015 and Sentinel-2B will follow in the second half of 2016) that provide high resolution optical imagery and provide continuity for the current SPOT and Landsat missions. The mission will provide a global coverage of the Earth's land surface every 10 days with one satellite (and 5 days with 2 satellites), making the data of great use in on-going studies.

Sentinel-2 delivers high-resolution optical images for land monitoring, emergency response and security



# Saving People

ESA, Planet Labs, Copernicus data from ESA's Sentinels & Zooniverse's network of citizen scientists

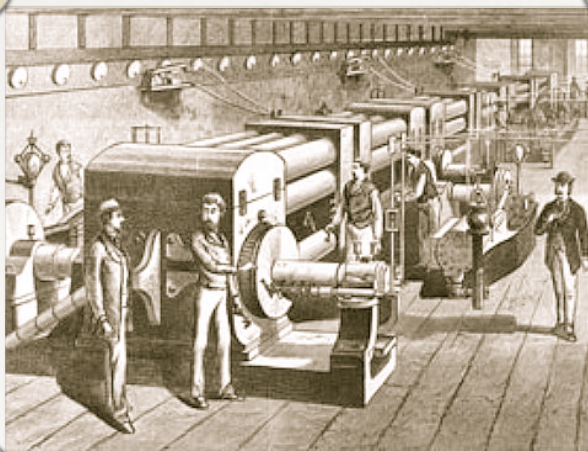
The first live public test for an effort dubbed the Planetary Response Network (PRN)

The screenshot shows a web browser displaying a news article on the Nature.com website. The article title is "Citizen scientists aid Ecuador earthquake relief" with a sub-headline "Effort to identify damaged areas combines crowdsourcing with machine-learning algorithms." The author is Mark Zastrow, and the date is 03 May 2016. A large satellite image of a forested area with a road network is visible. To the right, there is a sidebar with a "Hack the hackers" section and a "Sign up for FREE today" banner.

<http://www.nature.com/news/citizen-scientists-aid-ecuador-earthquake-relief-1.19861>

“Within 2 hours of the Ecuador test project going live with a first set of 1,300 images, each photo had been checked at least 20 times. “It was one of the fastest responses I’ve seen,” says Brooke Simmons, an astronomer at the University of California, San Diego, who leads the image processing. Steven Reece, who heads the Oxford team’s machine-learning effort, says that results — a “heat map” of damage with possible road blockages — were ready in another two hours.”





Pearl Street  
Power Station





Cray supercomputer  
28 Sept 1993

Cray Supercomputer



# < Please insert revolution >

## Linus Torvalds

From Wikipedia, the free encyclopedia

↻ → **A** This article **may be expanded with text translated from the corresponding article in** [show]  
**Finnish.** *(November 2014)* Click [show] for important translation instructions.

**Linus Benedict Torvalds** (/ˈliːnəsˈtɔːrvɔldz/<sup>[5]</sup> Swedish: [ˈlin.əs ˈtur.valds] (listen<sup>ⓘ</sup>); born December 28, 1969) is a Finnish-American<sup>[2][6]</sup> **software engineer**, who is the creator and, for a long time, principal developer, of the **Linux kernel**, which became the kernel for operating systems (and many **distributions** of each) such as **GNU** and years later **Android** and **Chrome OS**. He also created the distributed revision control system **git**. He was honored, along with **Shinya Yamanaka**, with the 2012 **Millennium Technology Prize** by the **Technology Academy Finland** "in recognition of his creation of a new open source operating system for computers leading to the widely used Linux kernel".<sup>[7]</sup> He is also the recipient of the 2014 **IEEE Computer Society Computer Pioneer Award**.<sup>[8]</sup>

### Contents [hide]

- Biography
  - Early years
  - Linux
- The Linux/Linux connection
- Authority and trademark
- Personal life
- Awards and Achievements
- Media recognition
- See also
- References
  - Footnotes
  - Bibliography
  - Further reading
- External links

Linus Torvalds



Torvalds at LinuxCon Europe 2014

**Born** Linus Benedict Torvalds  
December 28, 1969 (age 46)  
Helsinki, Finland

**Residence** Dunthorpe, Oregon, United States<sup>[1]</sup>

**Nationality** Finnish, American (naturalized in 2010)<sup>[2]</sup>



Cray supercomputer  
28 Sept 1993

Cray Supercomputer





Beowulf Cluster



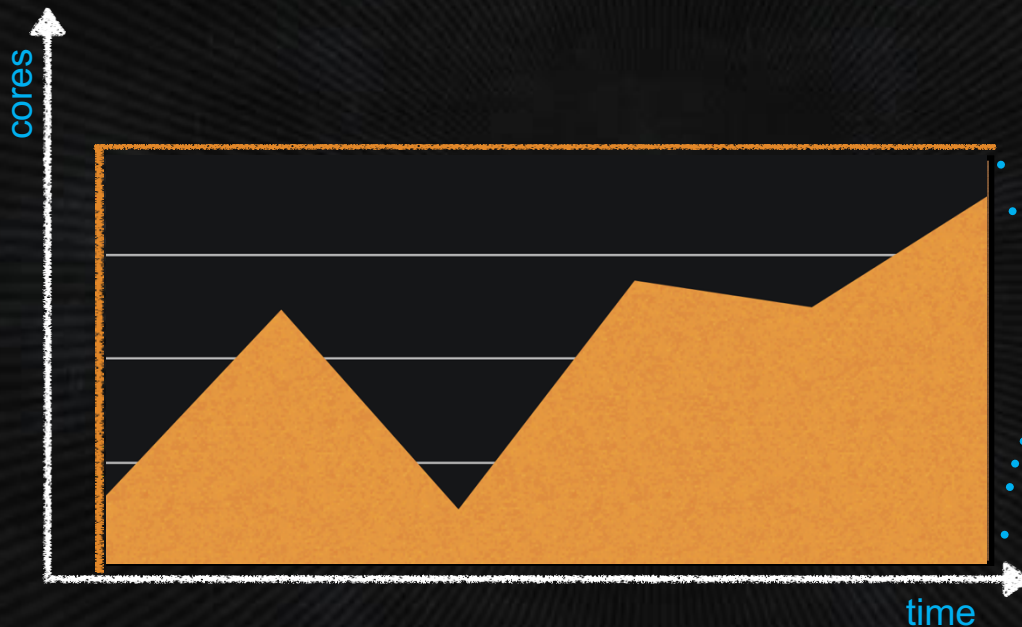


# The spherical model of owning a supercomputer



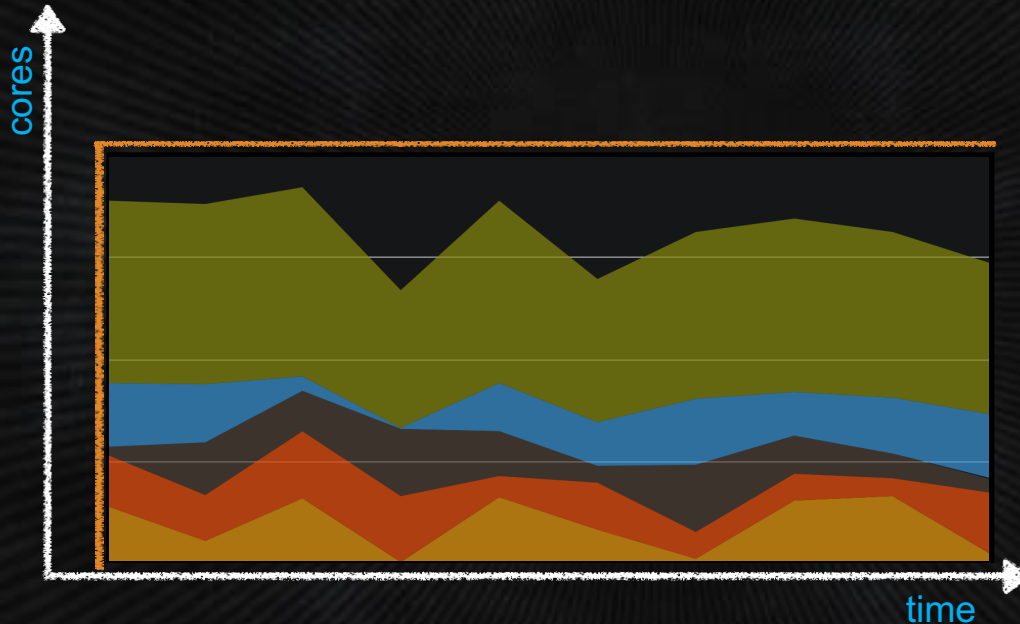
$$\begin{aligned} \$ &= A \int_{t_0}^T CPU(t).dt \\ &= \max(CPU) \cdot T_{3Yrs} \end{aligned}$$

# Empirical data



$$\begin{aligned} \$ &= A \int_{t_0}^T CPU(t).dt \\ &= \max(CPU) \cdot T_{3Yrs} \end{aligned}$$

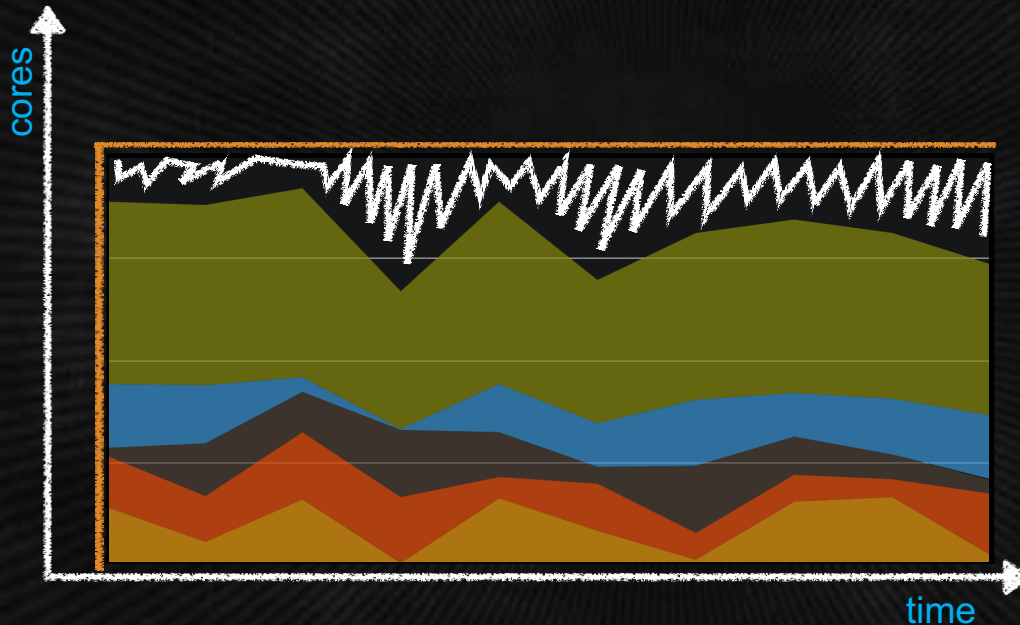
# Meeeeelions of uncorrelated workloads



## Collective action

When everyone comes together in the cloud to share the resource, and **only pays for what they use**, the efficiency is huge.

# Spot Market



## Spot Market

Our ultimate space filler.

Spot Instances allow you to name your own price for spare AWS EC2 computing capacity.

Great for workloads that aren't time sensitive, and especially popular in research (hint: **it's really cheap**).

# Spot Market Behavior

## Spot Bid Advisor

The Spot Bid Advisor analyzes Spot price history to help you determine a bid price that suits your needs.

You should weigh your application's tolerance for interruption and your cost saving goals when selecting a Spot instance and bid price.

The lower your frequency of being outbid, the longer your Spot instances are likely to run without interruption.

Spot Bid Advisor

Region: EU (Ireland) OS: Linux/UNIX Bid Price: 50% On-Demand

Instance type filter:  
vCPU (min): 8 Memory GiB (min): 0  Instance types supported by EMR

Instance Type	vCPU	Memory GiB	Savings over On-Demand*	Frequency of being outbid (month)	Frequency of being outbid (week)
m4.2xlarge	8	32	86%	Low	Low
c3.8xlarge	32	60	81%	Low	Low
c1.xlarge	8	7	87%	Low	Low
m2.4xlarge	8	68.4	92%	Low	Low
cr1.8xlarge	32	244	91%	Low	Low
hi1.4xlarge	16	60.5	95%	Low	Low
m3.2xlarge	8	30	84%	Medium	High
m4.4xlarge	16	64	86%	Medium	High
m4.10xlarge	40	160	87%	Medium	Medium
c4.2xlarge	8	15	84%	Medium	Medium
c4.xlarge	16	30	82%	Medium	Low
c4.8xlarge	36	60	82%	Medium	Low
c3.2xlarge	8	15	81%	Medium	Medium

## Bid Price & Savings

Your bid price affects your ranking when it comes to acquiring resources in the SPOT market, and is the maximum price you will pay.

But frequently you'll pay a lot less.

# Spot Market Behavior

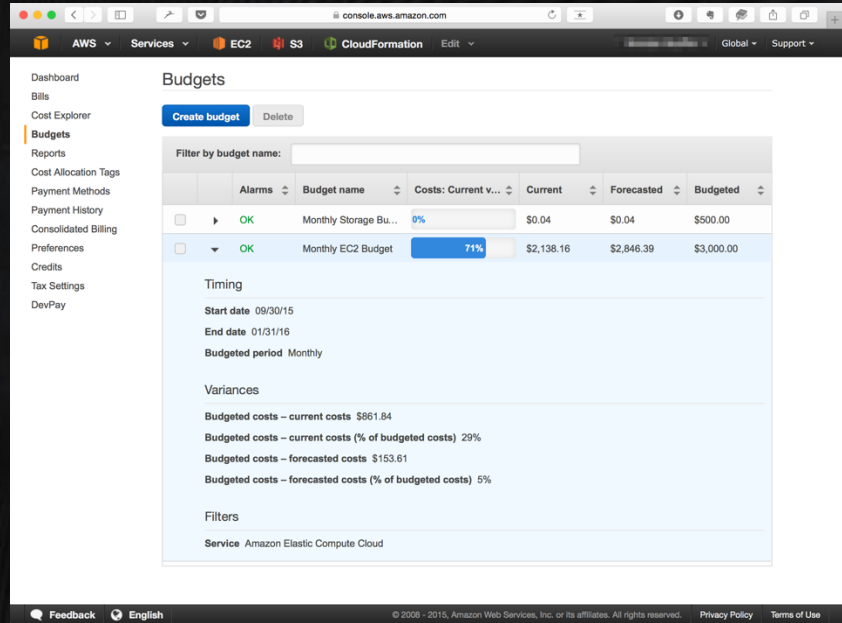
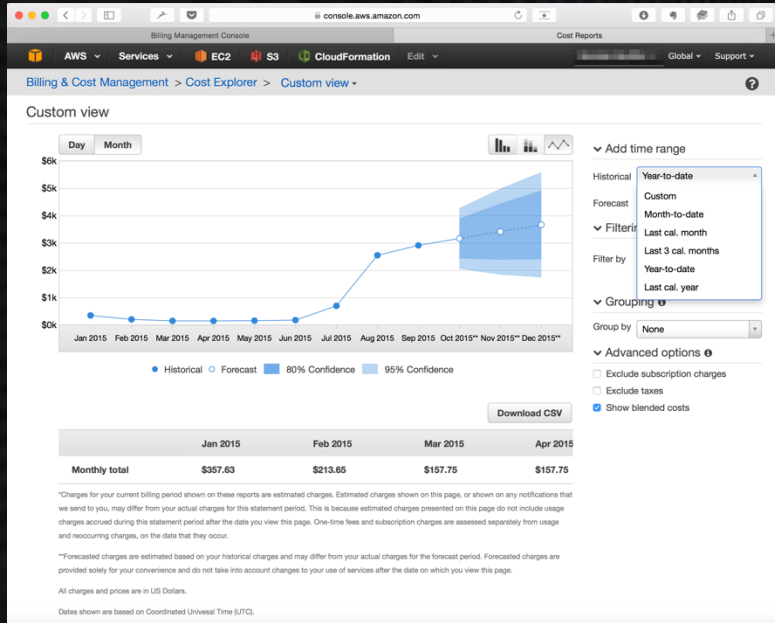
```
bash-3.2$ python get_spot_duration.py \  
--region us-east-1 \  
--product-description 'Linux/UNIX' \  
--bids c3.large:0.05,c3.xlarge:0.105,c3.2xlarge:0.21,c3.4xlarge:0.42,c3.8xlarge:0.84
```

Duration	Instance Type	Availability Zone
168.0	c3.8xlarge	us-east-1a
168.0	c3.8xlarge	us-east-1d
168.0	c3.8xlarge	us-east-1e
168.0	c3.4xlarge	us-east-1b
168.0	c3.4xlarge	us-east-1d
168.0	c3.4xlarge	us-east-1e
168.0	c3.xlarge	us-east-1d
168.0	c3.xlarge	us-east-1e
168.0	c3.large	us-east-1b
168.0	c3.large	us-east-1d
168.0	c3.large	us-east-1e
168.0	c3.2xlarge	us-east-1b
168.0	c3.2xlarge	us-east-1e
117.7	c3.large	us-east-1a
36.1	c3.2xlarge	us-east-1d
34.5	c3.4xlarge	us-east-1a
23.0	c3.xlarge	us-east-1b
21.9	c3.2xlarge	us-east-1a
17.3	c3.8xlarge	us-east-1b
0.1	c3.xlarge	us-east-1a

## Spot Bid Advisor

As usual with AWS, anything you can do with the web console, you can do with an API or command line.

# Cost Control & Budgeting



# Galaxies in the Cloud

**CHILES** will produce the first HI deep field, to be carried out with the VLA in B array and covering a redshift range from  $z=0$  to  $z=0.45$ . The field is centered at the COSMOS field. It will produce neutral hydrogen images of at least 300 galaxies spread over the entire redshift range.

The team at **ICRAR in Australia** have been able to implement the entire processing pipeline in the cloud for around \$2,000 per month by exploiting the SPOT market, which means the **\$1.75M** they otherwise needed to spend on an HPC cluster can be spent on way cooler things that impact their research ... like astronomers.



**CAASTRO**  
THE CENTER OF EXCELLENCE  
FOR ALIEN ASTROPHYSICS

## Computing efforts


**Single Machine**  
Big desktop: 48 Gb RAM

**Conventional Cluster (pleiades)**  
5 nodes each node has 2x Intel Xeon X5650  
2.66GHz CPUs (6 cores / 12 HTs)  
with 64-192 GB of RAM

**Super computer (MAGNUS)**  
Cray XC40 - 24 cores per node

Good for testing  
Would take ~year to finish

Enough computing power,  
however disk access  
limitations



PRISMC 2015, 16-18 March, New Jersey ARL114 Popping

**CAASTRO**  
THE CENTER OF EXCELLENCE  
FOR ALIEN ASTROPHYSICS

## Alternative (AWS)

**amazon web services**

	On demand	Spot Price
r3.4xlarge	\$1.68	\$0.20
r3.2xlarge	\$0.840	\$0.09
m3.xlarge	\$0.392	\$0.04
m3.medium	\$0.098	\$0.01



Works!  
costs so far : ~\$2000

**Spot Instance Pricing History**

Product: Linux/UNIX Instance type: r3.4xlarge Date range: 1 week Availability zone: All zones



PRISMC 2015, 16-18 March, New Jersey ARL114 Popping






# Breakthrough discoveries in the Cloud

The **CHILES project** astronomers have detected radio emissions from hydrogen in a galaxy more than **5 billion light years away**, **shattering the previous record** by almost twice. This has important implications for our understanding of how galaxies have evolved over time.

The team at **ICRAR in Western Australia** estimates that the amount of compute capacity required to shift and crunch this data **would have made this work infeasible**.

By using **AWS**, they were able to quickly and cheaply build their new pipelines, and then scale them as massive amounts of data arrived from their instruments.



International Centre for  
Radio Astronomy Research

**Embargoed Press Release**

Under embargo until:

June 1<sup>st</sup>: 8pm EST (New York), Midnight GMT (London)  
June 2<sup>nd</sup>: 2am CEST (Amsterdam), 8am AWST (Perth), 10am AEDT (Melbourne & Sydney)

**Astronomers smash cosmic records to see hydrogen in distant galaxy**

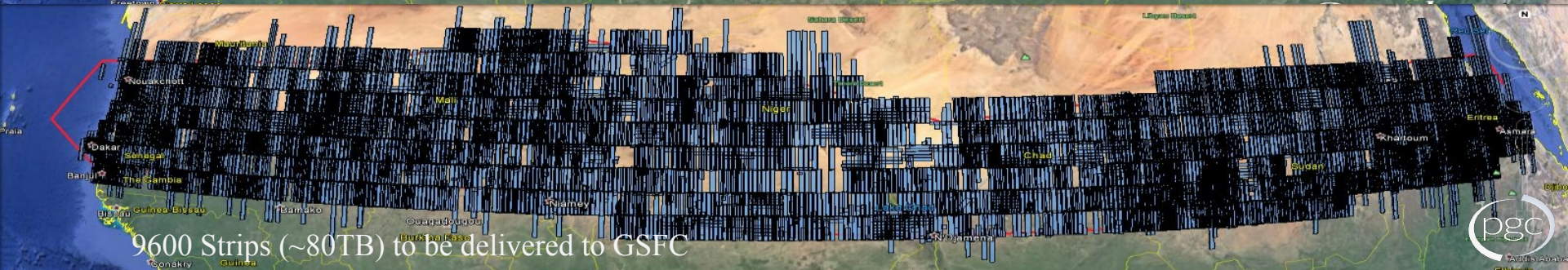
An international team of scientists has pushed the limits of radio astronomy to detect a faint signal emitted by hydrogen gas in a galaxy more than five billion light years away—almost double the previous record.

Using the Very Large Array of the National Radio Astronomy Observatory in the US, the team observed radio emission from hydrogen in a distant galaxy and found that it would have contained billions of young, massive stars surrounded by clouds of hydrogen gas.

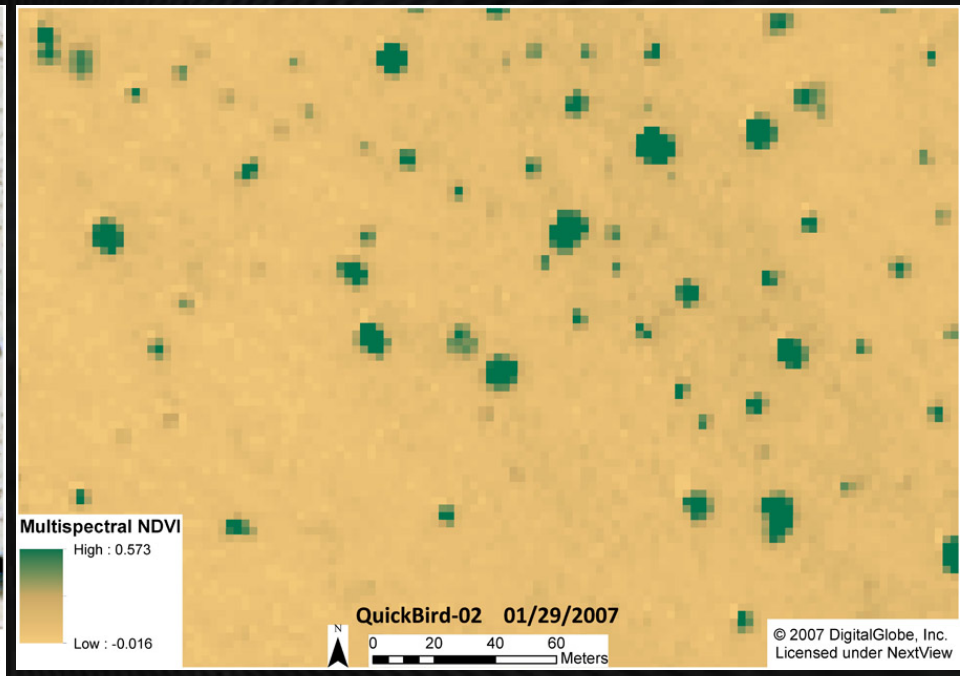
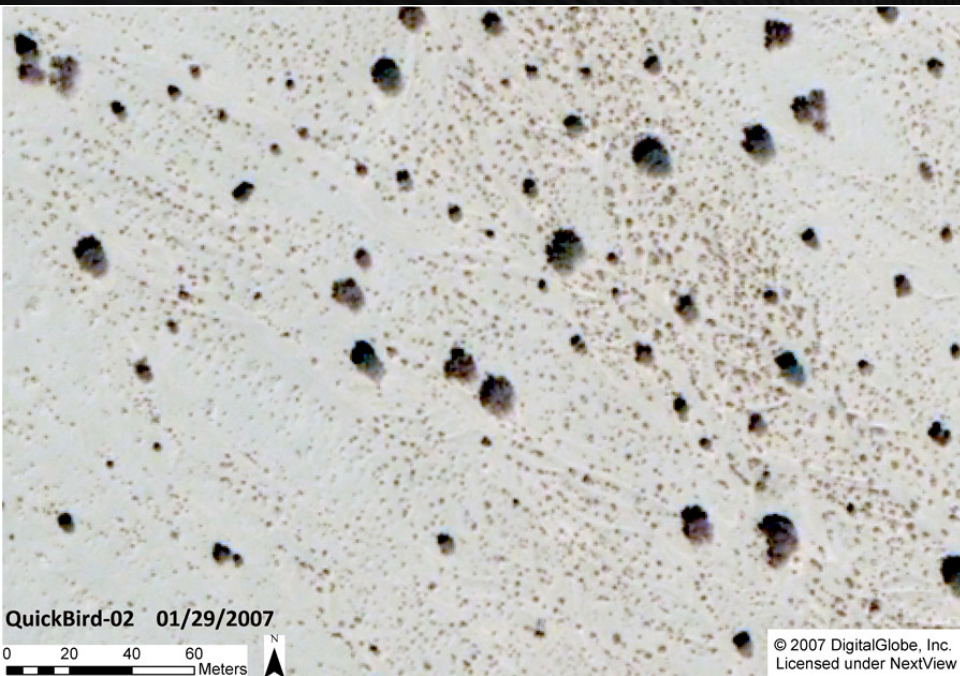
As the most abundant element in the Universe and the raw fuel for creating stars, hydrogen is used by radio astronomers to detect and understand the makeup of other galaxies.



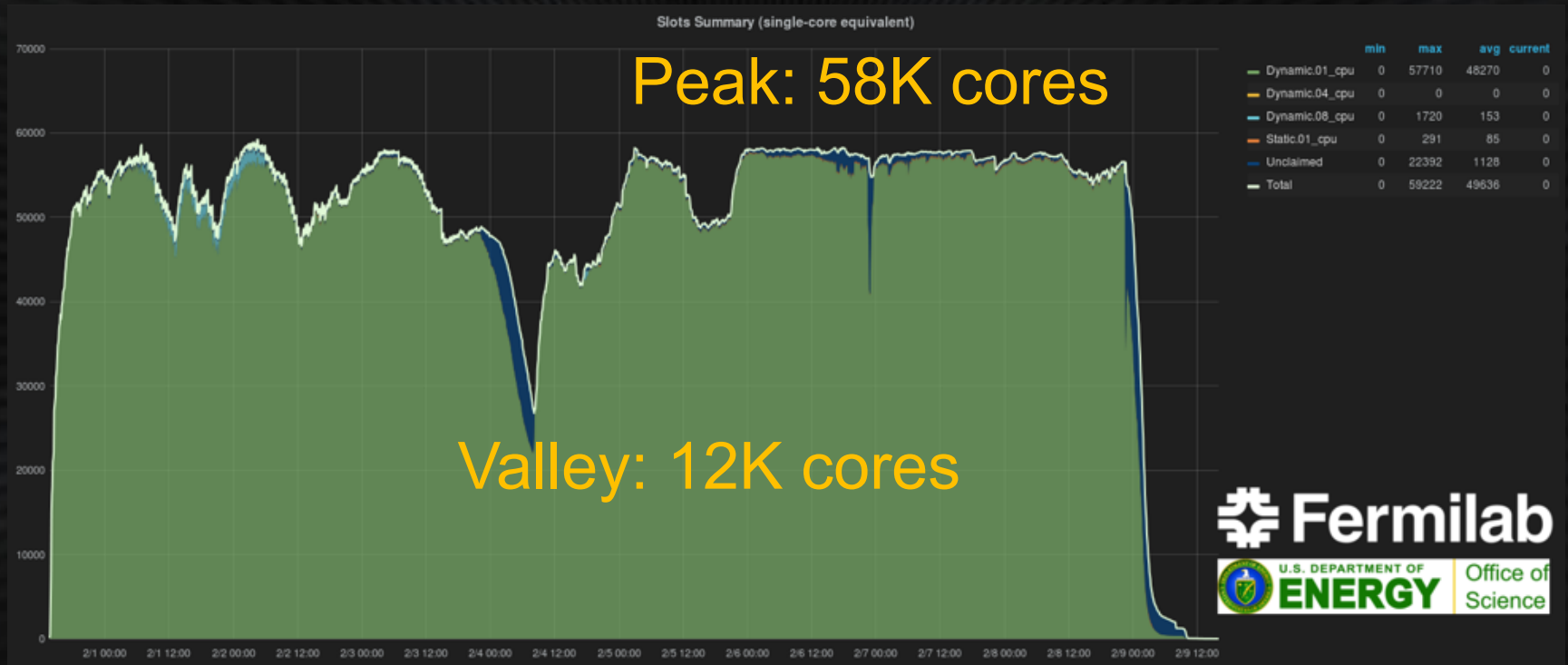
# Existing Sub-Saharan Arid and Semi-arid Sub-meter Commercial Imagery



# Panchromatic & Multi-Spectral Mapping at the 40cm - 50cm Scale



# Agility is...Paying Only for IT You Use



# Time travel for job queues



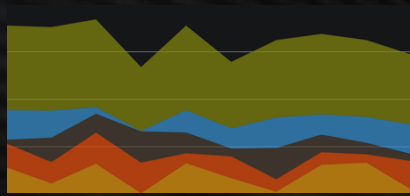
Wall clock time: ~1 hour



Wall clock time: ~1 week

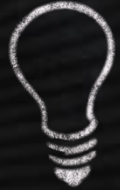
Cost: the same

# Choices



When you only pay for what you use ...

- If you're only able to use your compute, say, 30% of the time, you only pay for that time.



... you have options.

1

## Go Large

- Do 3x the science, or consume 3x the data.

2

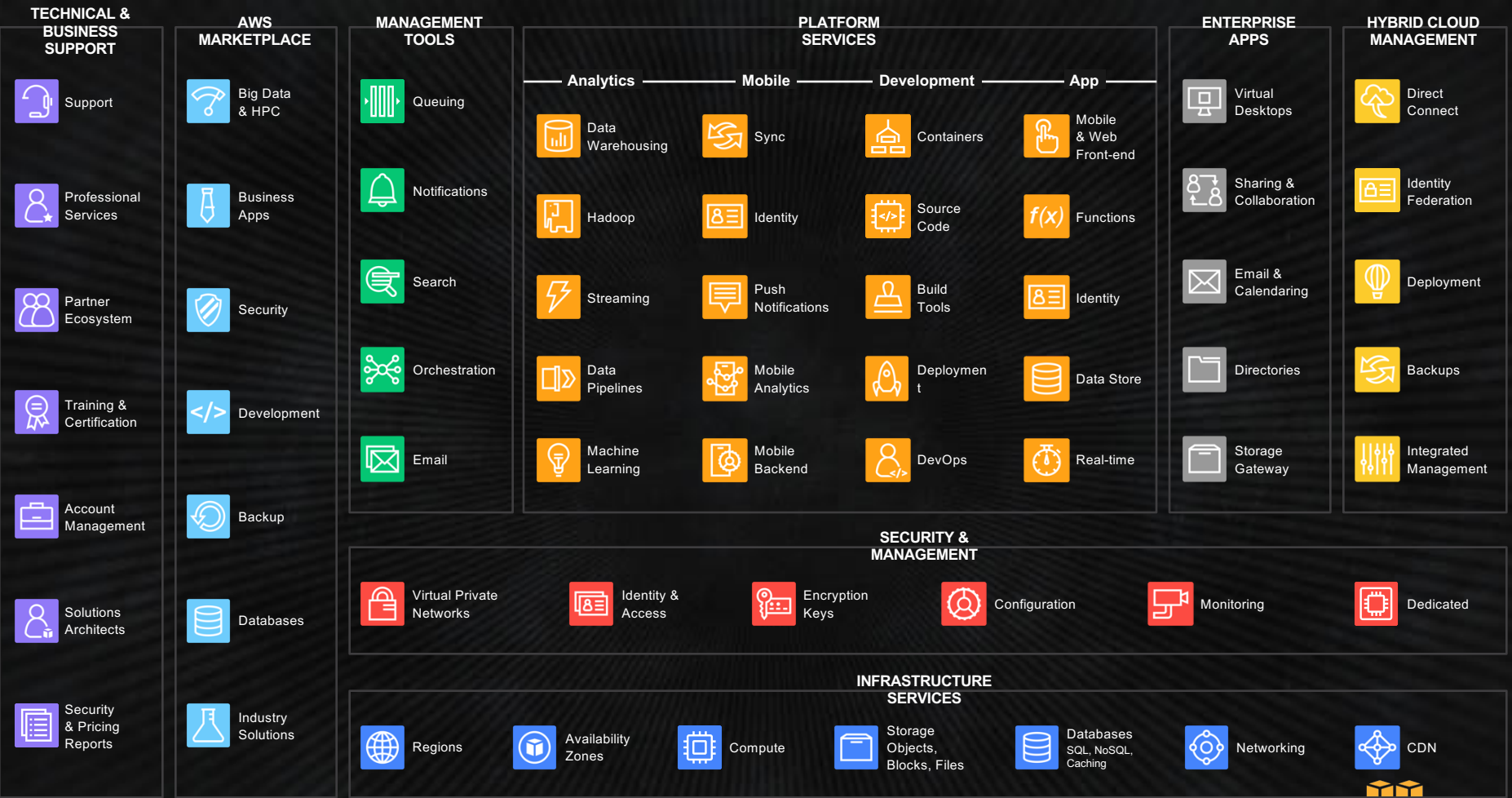
## Go faster

- Use 3x the cores to run your jobs at 3x the speed.

3

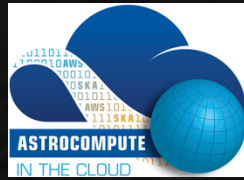
## Pocket the savings

- Buy chocolate
- Buy a spectrometer
- Hire a scientist.



# Finding what you're **not** looking for

(Widefield outTlier Finder)



## 'WTF is that?' How we're trawling the Universe for the unknown

By Eamonn Bermingham 9TH OCTOBER 2015



The Australian Square Kilometre Array Pathfinder. Credit: Alex Cherney

<http://blog.csiro.au/wtf-is-that-how-were-trawling-the-universe-for-the-unknown/>

WTF's cloud-based backend is hosted on Amazon Web Services servers, where the researchers are able to access software for data reduction, calibration and viewing right from their desktop. The team is currently issuing a challenge using data peppered with "EMU (Easter) Eggs" – objects that might pose a challenge to data mining algorithms.

This way they hope to train the system to recognise things that systematically depart from known categories of astronomical objects, to help better prepare for unanticipated discoveries that would otherwise remain hidden.





# EC2

There's a couple dozen EC2 compute instance types alone, each of which is optimized for different things.

One size does not fit all.

## Memory Optimized

### R3

R3 instances are optimized for memory-intensive applications and have the lowest cost per GiB of RAM among Amazon EC2 instance types.

#### Features:

- High Frequency Intel Xeon E5-2670 v2 (Ivy Bridge) Processors
- Lowest price point per GiB of RAM
- SSD Storage
- Support for [Enhanced Networking](#)

Model	vCPU	Mem (GiB)	SSD Storage (GB)
r3.large	2	15.25	1 x 32
r3.xlarge	4	30.5	1 x 80
r3.2xlarge	8	61	1 x 160
r3.4xlarge	16	122	1 x 320
r3.8xlarge	32	244	2 x 320

#### Use Cases

We recommend memory-optimized instances for high performance databases, distributed memory caches, in-memory analytics, genome assembly and analysis, larger deployments of SAP, Microsoft SharePoint, and other enterprise applications.

### GPU

#### G2

This family includes G2 instances intended for graphics and general purpose GPU compute applications.

#### Features:

- High Frequency Intel Xeon E5-2670 (Sandy Bridge) Processors
- High-performance NVIDIA GPU with 1,536 CUDA cores and 4GB of video memory
- On-board hardware video encoder designed to support up to eight real-time HD video streams (720p at 30fps) or up to four real-time FHD video streams (1080p at 30 fps).
- Support for low-latency frame capture and encoding for either the full operating system or select render targets, enabling high-quality interactive streaming experiences.

Model	vCPU	Mem (GiB)	SSD Storage (GB)
g2.2xlarge	8	15	1 x 60

#### Use Cases

Game streaming, video encoding, 3D application streaming, and other server-side graphics workloads.

<http://aws.amazon.com/ec2/instance-types/>

### C4

C4 instances are the latest generation of Compute-optimized instances, featuring the highest performing processors and the lowest price/compute performance in EC2.

#### Features:

- High frequency Intel Xeon E5-2666 v3 (Haswell) processors optimized specifically for EC2
- EBS-optimized by default and at no additional cost
- Ability to control processor C-state and P-state configuration on the c4.xlarge instance type
- Support for [Enhanced Networking](#) and Clustering

Model	vCPU	Mem (GiB)	Storage	Dedicated EBS Throughput (Mbps)
c4.large	2	3.75	EBS-Only	500
c4.xlarge	4	7.5	EBS-Only	750
c4.2xlarge	8	15	EBS-Only	1,000
c4.4xlarge	16	30	EBS-Only	2,000
c4.8xlarge	36	60	EBS-Only	4,000

### C3

#### Features:

- High Frequency Intel Xeon E5-2680 v2 (Ivy Bridge) Processors
- Support for [Enhanced Networking](#)
- Support for clustering
- SSD-backed instance storage

Model	vCPU	Mem (GiB)	SSD Storage (GB)
c3.large	2	3.75	2 x 16
c3.xlarge	4	7.5	2 x 40
c3.2xlarge	8	15	2 x 80
c3.4xlarge	16	30	2 x 160
c3.8xlarge	32	60	2 x 320

### M3

This family includes the M3 instance types and provides a balance of compute, memory, and network resources, and it is a good choice for many applications.

#### Features:

- High Frequency Intel Xeon E5-2670 v2 (Ivy Bridge) Processors
- SSD-based instance storage for fast I/O performance
- Balance of compute, memory, and network resources

Model	vCPU	Mem (GiB)	SSD Storage (GB)
m3.medium	1	3.75	1 x 4
m3.large	2	7.5	1 x 32
m3.xlarge	4	15	2 x 40
m3.2xlarge	8	30	2 x 80

#### Use Cases

Small and mid-size databases, data processing tasks that require additional memory, caching fleets, and for running backend servers for SAP, Microsoft SharePoint, and other enterprise applications.



# C4

Intel Xeon E5-2666 v3, custom built for AWS.

Intel Haswell, 16 FLOPS/tick

2.9 GHz, turbo to 3.5 GHz

Feature	Specification
Processor Number	E5-2666 v3
Intel® Smart Cache	25 MiB
Instruction Set	64-bit
Instruction Set Extensions	AVX 2.0
Lithography	22 nm
Processor Base Frequency	2.9 GHz
Max All Core Turbo Frequency	3.2 GHz
Max Turbo Frequency	3.5 GHz (available on c4.2xLarge)
Intel® Turbo Boost Technology	2.0
Intel® vPro Technology	Yes
Intel® Hyper-Threading Technology	Yes
Intel® Virtualization Technology (VT-x)	Yes
Intel® Virtualization Technology for Directed I/O (VT-d)	Yes
Intel® VT-x with Extended Page Tables (EPT)	Yes
Intel® 64	Yes

AWS Official Blog

## New Compute-Optimized EC2 Instances

by Jeff Barr | on 13 NOV 2014 | in [Amazon EC2](#) | [Permalink](#)

Our customers continue to increase the sophistication and intensity of the compute-bound workloads that they run on the Cloud. Applications such as top-end website hosting, online gaming, simulation, risk analysis, and rendering are voracious consumers of CPU cycles and can almost always benefit from the parallelism offered by today's multicore processors.

### The New C4 Instance Type

Today we are pre-announcing the latest generation of compute-optimized [Amazon Elastic Compute Cloud \(EC2\)](#) instances. The new C4 instances are based on the Intel Xeon E5-2666 v3 (code name Haswell) processor. This custom processor, designed specifically for EC2, runs at a base speed of 2.9 GHz, and can achieve clock speeds as high as 3.5 GHz with Turbo boost. These instances are designed to deliver the highest level of processor performance on EC2. If you've got the workload, we've got the instance!

Here's the lineup (these specs are preliminary and could change a bit before launch time):

Instance Name	vCPU Count	RAM	Network Performance
c4.large	2	3.75 GiB	Moderate
c4.xlarge	4	7.5 GiB	Moderate
c4.2xlarge	8	15 GiB	High
c4.4xlarge	16	30 GiB	High
c4.8xlarge	36	60 GiB	10 Gbps

These instances are a great match for the [SSD-Backed Elastic Block Storage](#) that we introduced earlier this year. [EBS Optimization](#) is enabled by default for all C4 instance sizes, and is available to you at no extra charge. C4 instances also allow you to achieve significantly higher packet per second (PPS) performance, lower network jitter, and lower network latency using [Enhanced Networking](#).

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/c4-instances.html>



# The AWS Console

**Amazon Web Services**

**Compute**

- EC2**  
Virtual Servers in the Cloud
- EC2 Container Service**  
Run and Manage Docker Containers
- Elastic Beanstalk**  
Run and Manage Web Apps
- Lambda**  
Run Code in Response to Events

**Storage & Content Delivery**

- S3**  
Scalable Storage in the Cloud
- CloudFront**  
Global Content Delivery Network
- Elastic File System**  
Fully Managed File System for EC2
- Glacier**  
Archive Storage in the Cloud
- Snowball**  
Large Scale Data Transport
- Storage Gateway**  
Hybrid Storage Integration

**Database**

- RDS**  
Managed Relational Database Service
- DynamoDB**  
Managed NoSQL Database
- ElastiCache**  
In-Memory Cache
- Redshift**  
Fast, Simple, Cost-Effective Data Warehousing
- DMS**  
Managed Database Migration Service

**Networking**

- VPC**  
Isolated Cloud Resources
- Direct Connect**  
Dedicated Network Connection to AWS
- Route 53**  
Scalable DNS and Domain Name Registration

**Developer Tools**

- CodeCommit**  
Store Code in Private Git Repositories
- CodeDeploy**  
Automate Code Deployments
- CodePipeline**  
Release Software using Continuous Delivery

**Management Tools**

- CloudWatch**  
Monitor Resources and Applications
- CloudFormation**  
Create and Manage Resources with Templates
- CloudTrail**  
Track User Activity and API Usage
- Config**  
Track Resource Inventory and Changes
- OpsWorks**  
Automate Operations with Chef
- Service Catalog**  
Create and Use Standardized Products
- Trusted Advisor**  
Optimize Performance and Security

**Security & Identity**

- Identity & Access Management**  
Manage User Access and Encryption Keys
- Directory Service**  
Host and Manage Active Directory
- Inspector**  
Analyze Application Security
- WAF**  
Filter Malicious Web Traffic
- Certificate Manager**  
Provision, Manage, and Deploy SSL/TLS Certificates

**Analytics**

- EMR**  
Managed Hadoop Framework
- Data Pipeline**  
Orchestration for Data-Driven Workflows
- Elasticsearch Service**  
Run and Scale Elasticsearch Clusters
- Kinesis**  
Work with Real-Time Streaming Data
- Machine Learning**  
Build Smart Applications Quickly and Easily

**Internet of Things**

- AWS IoT**  
Connect Devices to the Cloud

**Game Development**

- GameLift**  
Deploy and Scale Session-based Multiplayer Games

**Mobile Services**

- Mobile Hub**  
Build, Test, and Monitor Mobile Apps
- Cognito**  
User Identity and App Data Synchronization
- Device Farm**  
Test Android, iOS, and Web Apps on Real Devices in the Cloud
- Mobile Analytics**  
Collect, View and Export App Analytics
- SNS**  
Push Notification Service

**Application Services**

- API Gateway**  
Build, Deploy and Manage APIs
- AppStream**  
Low Latency Application Streaming
- CloudSearch**  
Managed Search Service
- Elastic Transcoder**  
Easy-to-Use Scalable Media Transcoding
- SES**  
Email Sending and Receiving Service
- SQS**  
Message Queue Service
- SWF**  
Workflow Service for Coordinating Application Components

**Enterprise Applications**

- WorkSpaces**  
Desktops in the Cloud
- WorkDocs**  
Secure Enterprise Storage and Sharing Service
- WorkMail**  
Secure Email and Calendaring Service

# The AWS API

Almost anything you can do in the GUI, you can do on the command line or via our API

```
# use power user policy
#
resp = iam.get_policy(policy_arn: "arn:aws:iam:aws:policy/PowerUserAccess")
policy_arn = resp.policy_arn

# create an output CSV document
output = File.open("users.csv", 'a+')
output.puts(%w{ UserName Password AwsAccessKeyId, AwsSecretAccessKey}.to_csv)

users.each do |user|
  # create user
  iam.create_user({user_name: user[:name]})
  # attach policy
  iam.attach_user_policy({
    user_name: user[:name],
    policy_arn: policy_arn
  })
  # create login
  iam.create_login_profile({
    user_name: user[:name].to_s,
    password: user[:pw].to_s,
    password_reset_required: false,
  });
  # create access keys
  resp = iam.create_access_key({user_name: user[:name]})
  user[:access_key_id] = resp.access_key.access_key_id
  user[:secret_access_key] = resp.access_key.secret_access_key
  output.puts(
    [
      user[:name],
      user[:pw],
      user[:access_key_id],
      user[:secret_access_key]
    ].to_csv
  )
  puts "--\n"
  puts "Portal:          #{MY_URL}\n"
  puts "Username:         #{user[:name]}\n"
  puts "Password:         #{user[:pw]}\n"
  puts "Access Key ID:    #{user[:access_key_id]}\n"
  puts "Secret Access Key: #{user[:secret_access_key]}\n"
end
```

<http://boto.readthedocs.org/en/latest/>

Java  
Python  
Ruby  
PHP  
Shell

...

... and in most popular languages.





Cray supercomputer  
28 Sept 1993

Cray Supercomputer





Beowulf Cluster



# A top500 supercomputer ... 2013-style

76	Amazon Web Services United States	Amazon EC2 C3 Instance cluster - Amazon EC2 Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, 10G Ethernet Self-made	26496	484.2	593.5
----	--------------------------------------	---	-------	-------	-------

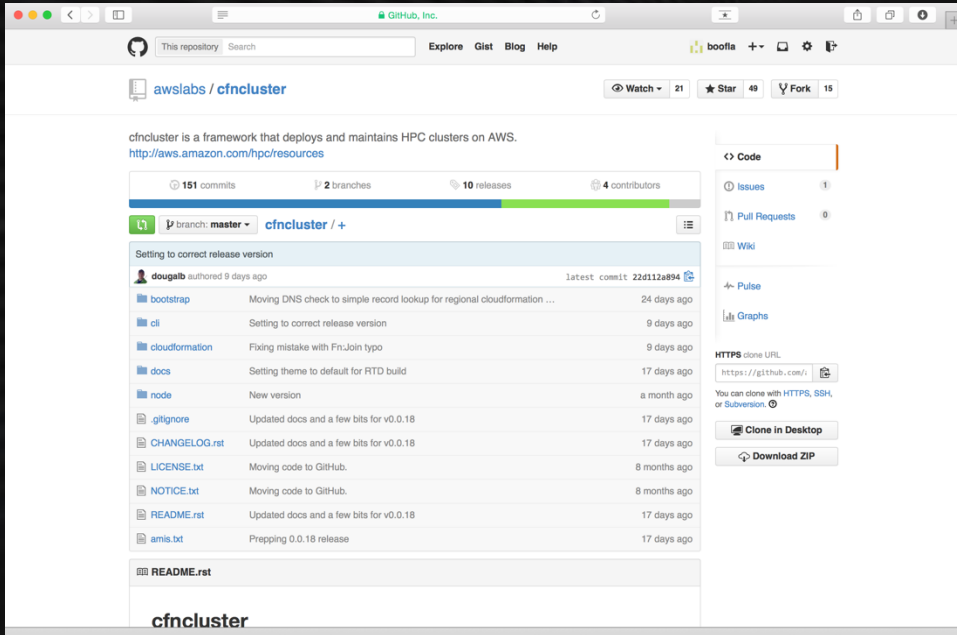


Ready in ~100 seconds

For ~ \$100/hr

# cfncCluster - provision an HPC cluster in minutes

cfnccluster is a **sample code framework** that **deploys and maintains clusters on AWS**. It is reasonably agnostic to what the cluster is for and can easily be extended to support different frameworks. The CLI is stateless, everything is done using **CloudFormation** or resources within AWS.



```
artthur - [44] $ grep -iv "# .cfnccluster/config"
[global]
cluster_template = default
update_check = true
sanity_check = true

[aws]
aws_region_name = ap-southeast-2

[cluster default]
key_location = /Users/bouffler/.ssh
key_name = boof-cluster
scheduler = sge
vpc_settings = public

[vpc public]
vpc_id = vpc-c48a4fa1
master_subnet_id = subnet-3108f146

artthur - [45] $
```



#cfnccluster

<https://github.com/awslabs/cfnccluster>



10 minutes



<http://boofla.io/u/cfnCluster> – (Boof's HOWTO slides)



# Self-service Supercomputing ... 2016

Introducing **Alces Flight** - self-scaling HPC clusters instantly ready to compute, billed by the hour and using the AWS Spot market by default to achieve supercomputing for ~1c per core per hour.



- 750+ popular scientific applications
  - Pre-installed & ready to run.
  - Multiple versions, complete with libraries and various compiler optimizations, ready to run
- Available via the **AWS Marketplace** (the cloud's "App Store") and launched within minutes.
- Deployable anywhere on Earth ... **immediately**.

<http://boofla.io/u/alcesFlight>





# Filesystems in the marketplace, too

There are cluster filesystem options, too— for when you need extreme I/O scaling.

- **BeeGFS** is a scalable parallel cluster filesystem developed with a strong focus on performance and designed easy installation and management developed by the Fraunhofer Institute.
- **Intel Lustre® Cloud Edition** is a scalable, parallel file system purpose-built for HPC and with a long history in the field supporting a range of workloads.
- **There's more to come** - the AWS Marketplace is growing all the time and new offerings are added frequently. Watch this space.





Thank You