

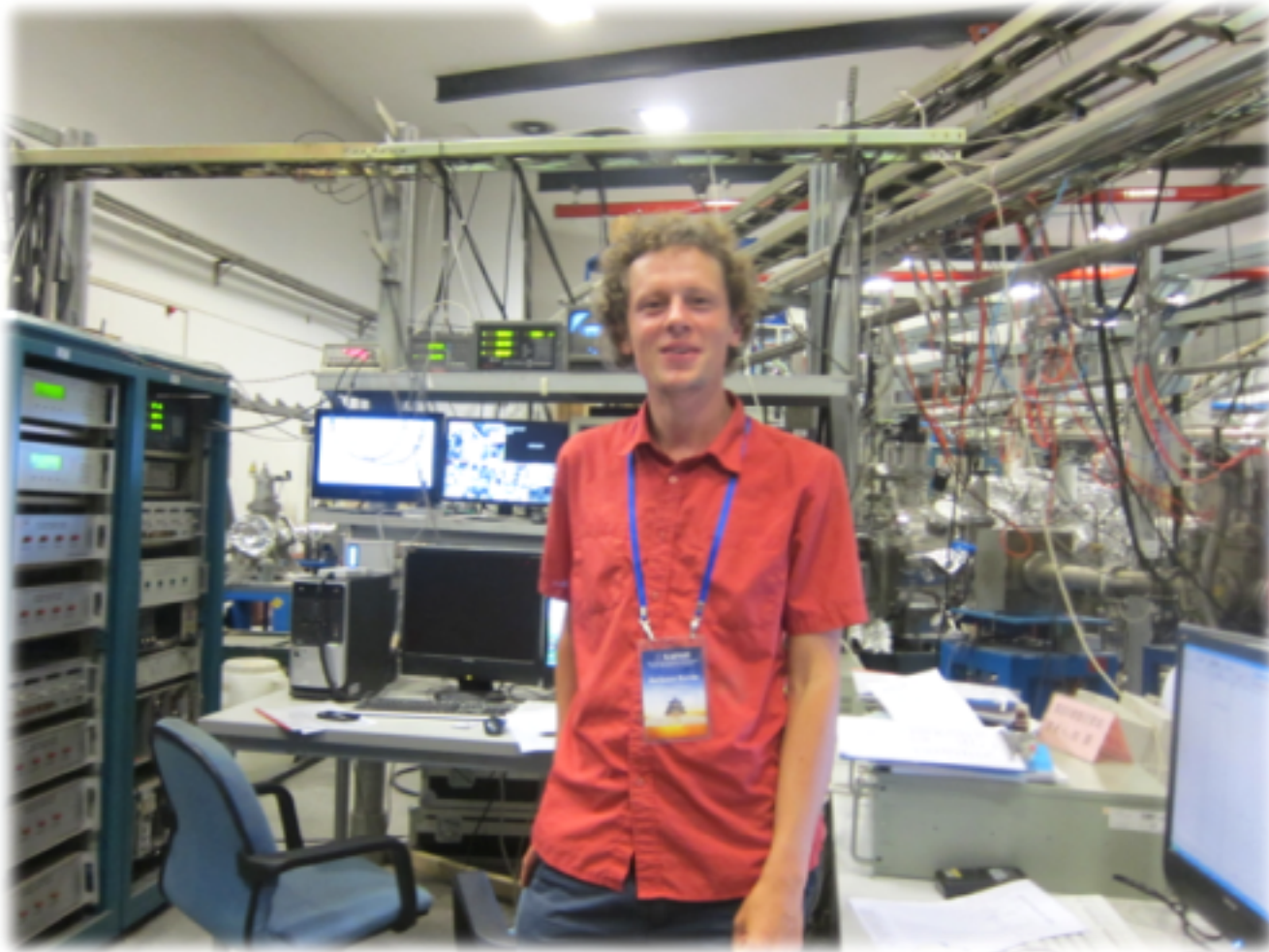
AWS for Scientific Workflows

Unidata workshop, June 2018

Kevin Jorissen

jorissen@amazon.com





Kevin Jorissen

Seattle

Kevin has 10 years of experience in computational science, and holds a Ph.D. in **Physics**. He developed codes solving the quantum physics equations for light absorption by materials, taught workshops to scientists worldwide, and wrote about high performance computing in the cloud before it was fashionable. He worked as a **postdoctoral researcher** in Antwerp, Lausanne, Seattle, and Zurich. He contributed to the WIEN2k code (Density Functional Theory calculations of material properties, www.wien2k.at) and the FEFF code (X-ray and Electron absorption spectra, www.fefferproject.org).

Kevin joined **Amazon** in 2015 to help accelerate the adoption of cloud computing in the scientific community globally.

Agenda

Unidata ask: new technologies & skills transfer



Warmup

1. Running models, HPC, Clusters
2. Data Lakes
3. Containers, AWS Batch, Microservices
4. Serverless Computing
5. Machine Learning, Amazon SageMaker, Notebooks



Thank You & Homework

jorissen@amazon.com

Sign up for the Researchers Handbook for AWS at aws.amazon.com/rcp . Browse data at <https://registry.opendata.aws>

1. Alces Flight **compute cluster** - NAMD tutorial: Launch “Performance Compute (SGE)” cluster at <https://launch.alces-flight.com/default/launch> , wait for e-mail confirmation, then tutorial from <http://docs.alces-flight.com/en/stable/getting-started/environment-usage/using-openfoam-with-alces-flight-compute.html>
2. **Containers + AWS Batch** for DNA sequencing: <https://github.com/awslabs/aws-batch-genomics>
3. **Containers** – WRF Big Weather Web: www.bigweatherweb.org
4. **Serverless Computing** – PyWren: <http://pywren.io/pages/gettingstarted.html>
then <https://github.com/pywren/examples/>
5. **SageMaker Machine Learning** labs: files from <https://bit.ly/2HhD2SG> ; instructions at <https://github.com/wleepang/sagemaker4research-workshop> ; further labs at <https://developmentseed.org/blog/2018/01/19/sagemaker-label-maker-case/> and <https://aws.amazon.com/blogs/machine-learning/simulate-quantum-systems-on-amazon-sagemaker/>

1

AWS and Scientific Workflows

AWS and Scientific Workflows

- **Agility** == “time to discovery”
 - Availability of resources, scalability, right-sizing
 - Experiment, fail fast, avoid undifferentiated work

AWS and Scientific Workflows

- **Agility** == “time to discovery”
 - Availability of resources, scalability, right-sizing
 - Experiment, fail fast, avoid undifferentiated work

- **Collaboration**
 - Data lake model
 - Security
 - Sharing
 - Infrastructure
 - Analytics

Hot off the presses: WWPS AWS Summit

Real-Time Machine Learning on Satellite Imagery: How DigitalGlobe Uses Amazon SageMaker to Massively Scale-up Information Extraction from Satellite Imagery

Using AWS and Open Data to Meet the Demands of Disaster Response Situations

Transitioning Geoscience Research to the Cloud: Opportunities and Challenges

AWS Public Datasets: Learnings from Staging Petabytes of Data for Analysis in AWS

Enabling Sustainable Research Platforms in the Cloud

Enabling Research using Hybrid HPC Cloud Computing

Precision Medicine on the Cloud

Transforming Research in Collaboration with Funding Agencies

Enabling Research using Hybrid HPC Cloud Computing

Innovation on the Edge: How Rapid Experimentation with Technology is Achieving Results in the Enterprise With NASA JPL

Accelerating Analytics for the Future of Genomics

Analyzing Data Streams in Real Time with Amazon Kinesis: PNNL's Serverless Data Lake Ingestion

Empowering Every Brain! How Brain Power is using AWS-Powered AI in their Mission to Help People with Autism and Other Brain-Related Challenges

... Soon available at <https://www.youtube.com/user/AmazonWebServices/videos>

Hot off the presses: WWPS AWS Summit

“Earth and Space on AWS” Day

- How Element 84 Raises the Bar on Streaming Satellite Data (Element 84)
- Machine Learning with Earth Observation Imagery (EOS Data Analytics, DevelopmentSeed)
- Making Sense of Remote Sensing (Sinergise, SkyWatch)
- Black Sky: Advancing the Geospatial Revolution with Cloud-First Approach (SpaceFlight Industries)
- How Can We Answer the Really BIG Questions? (NASA JPL)
- Lessons Learned Migrating Space Operational Systems to the Cloud on AWS (Lockheed Martin)
- Earth is Just Our Starting Place: Blue Origin and the Future of Space Technology (Blue Origin)



Earth and Space on AWS

Processing and Streaming GOES-16 Data with AWS Managed Services

Element 84

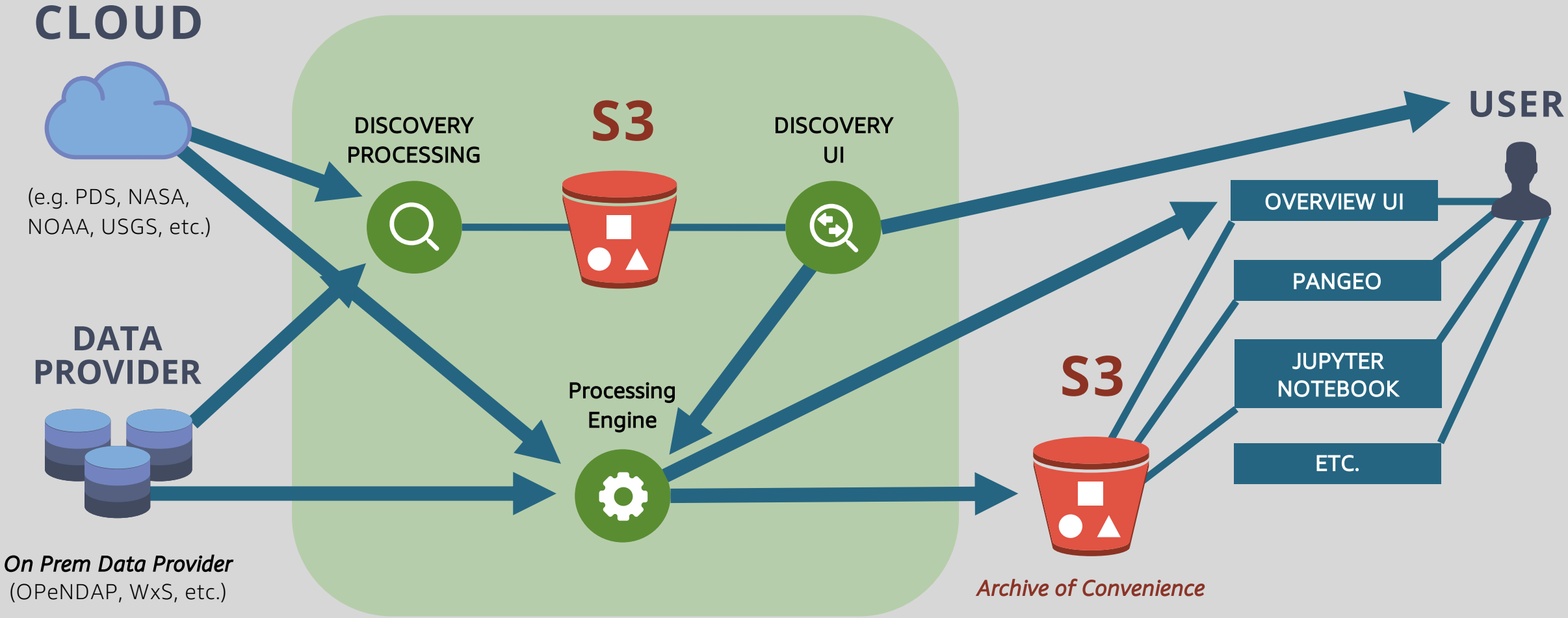
Dan Pilone

CTO - Element 84, Inc.

- We leveraged AWS EC2/Spot/ECS and ETS to make ~20 TBs of AWS Public Dataset GOES-16 imagery visually navigable at varying levels of bandwidth.
- We can apply this approach to lots and lots of data products
- We've leveraged AWS Batch (ECS & Spot) to parallelize creation of data bundles into ephemeral Archives of Convenience
- Users get convenient, highly elastic access to data that suits their needs, in their preferred format.
- All of this costs \$0 when not in active use but scales horizontally as big as budget allows.

Demo @ <https://labs.element84.com/index.html>

Overall Data Flow



Working with cloud-based Zarr files

Connecting to S3

Because Zarr loads datasets by chunks, we can keep most of the data on S3, and only pull down the pieces we want:

```
In [1]: import s3fs
import zarr

s3 = s3fs.S3FileSystem(profile_name='rd', client_kwargs=dict(region_name='us-east-1'))
store = s3fs.S3Map(root='e84-goes/dan/test.zarr', s3=s3, check=False)
big_zarr = zarr.group(store=store)
```

```
In [2]: print big_zarr.keys()

['2017-12-31T06:11:24.6Z', '2017-12-31T06:26:24.6Z', '2017-12-31T06:41:24.2Z', '2017-12-31T06:56:24.4Z']
```

This dataset is only four frames, but already is 1.5GB. For a video of non-trivial size, zarr allows us to only pull the pieces of the datasets we need, without downloading the entire group.

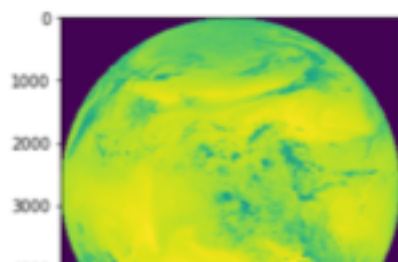
```
In [3]: %matplotlib inline
```

```
In [5]: import matplotlib.pyplot as plt

import time
start_time = time.time()

# Plot only the "lower-level water vapor" infrared band for a specific frame in a larger dataset
band10 = big_zarr['2017-12-31T06:26:24.6Z']['CMI_C10'][:, :]
plt.imshow(band10)
```

```
Out[5]: <matplotlib.image.AxesImage at 0x121d7ce10>
```





Machine Learning with Earth Observation Imagery

NaNa Yi

Engineer, Development Seed

Marc M Fagan

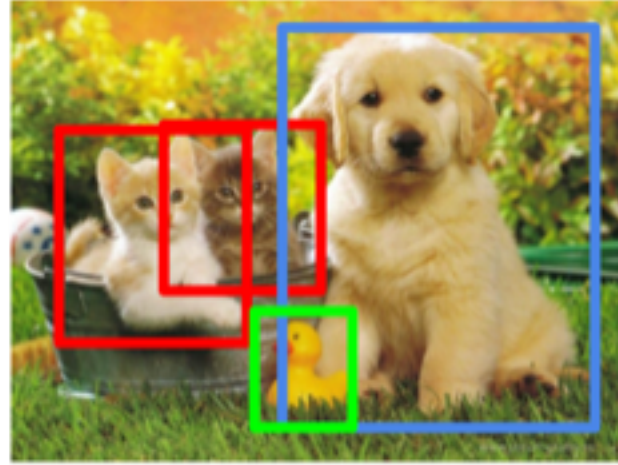
CEO, EOS Data Analytics

Classification

Classification + Localization

Object Detection

Instance Segmentation



CAT

CAT

CAT, DOG, DUCK

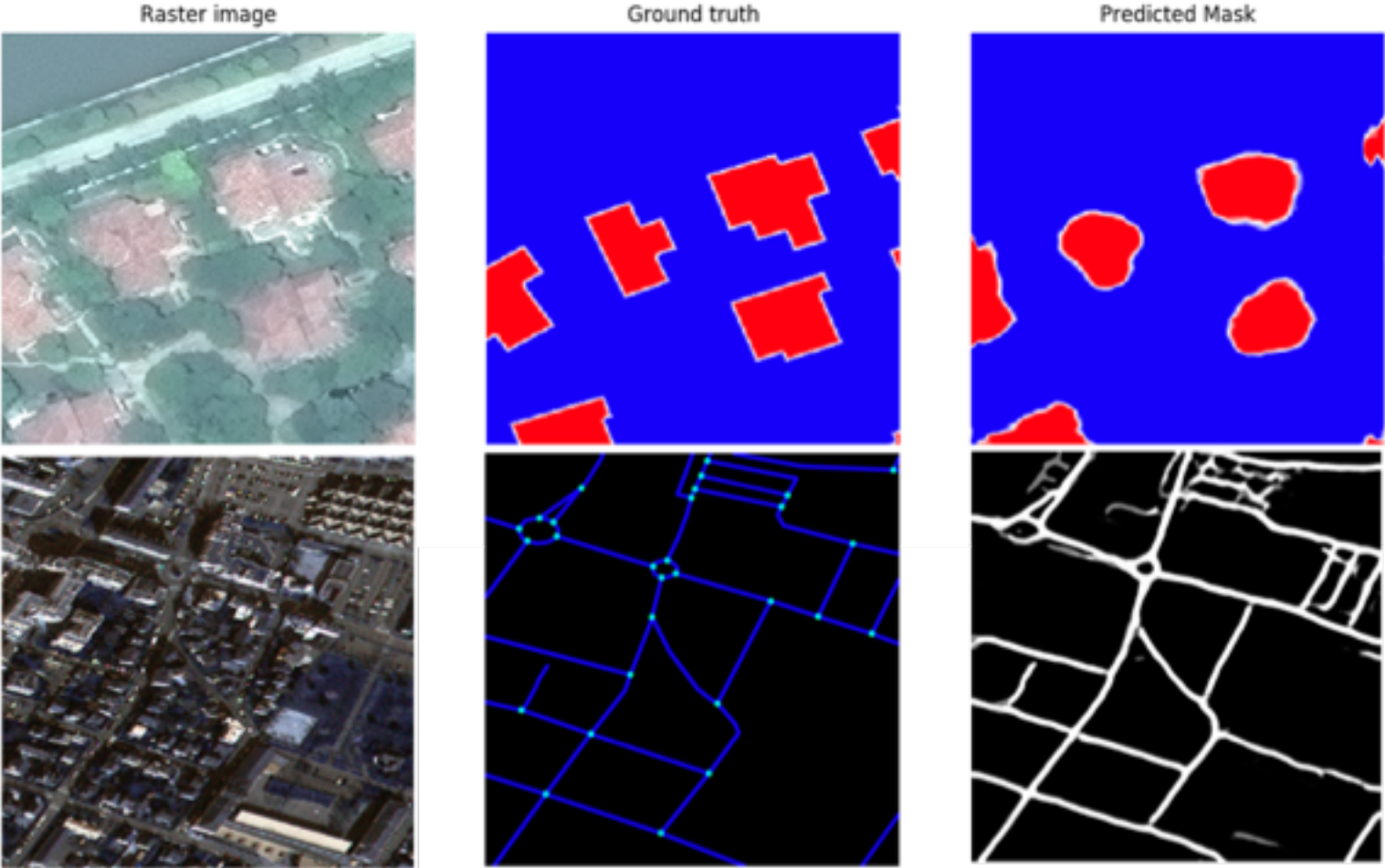
CAT, DOG, DUCK

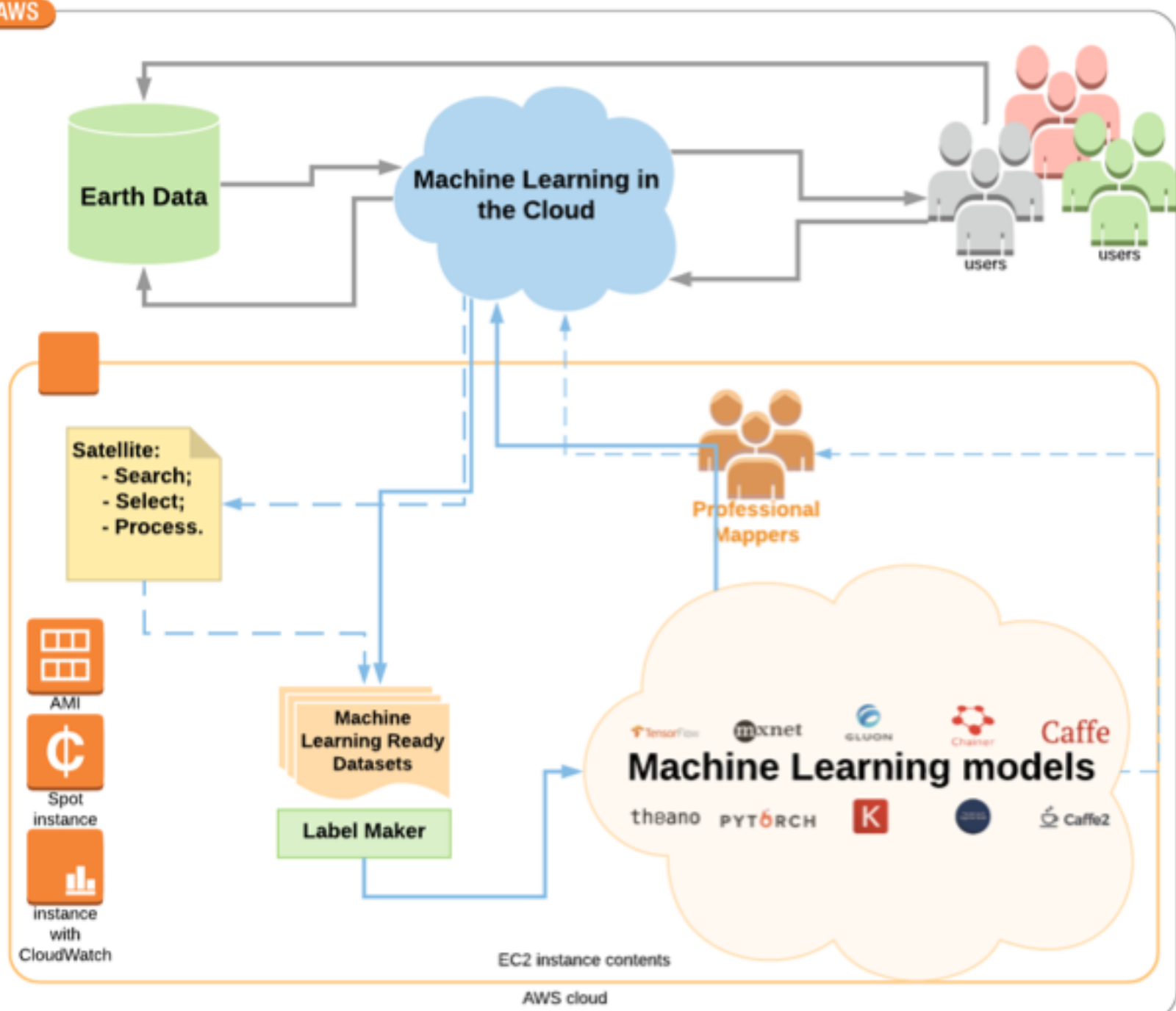
Single object

Multiple objects

Segmentation for detecting building footprint and road network

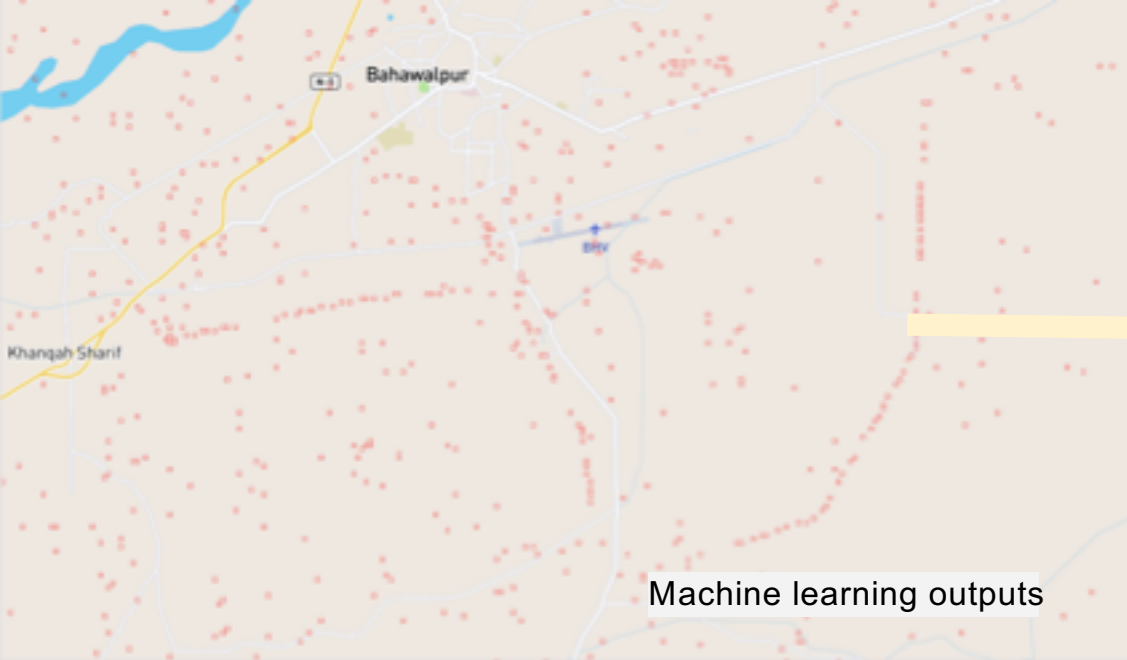
Object detection for building counts



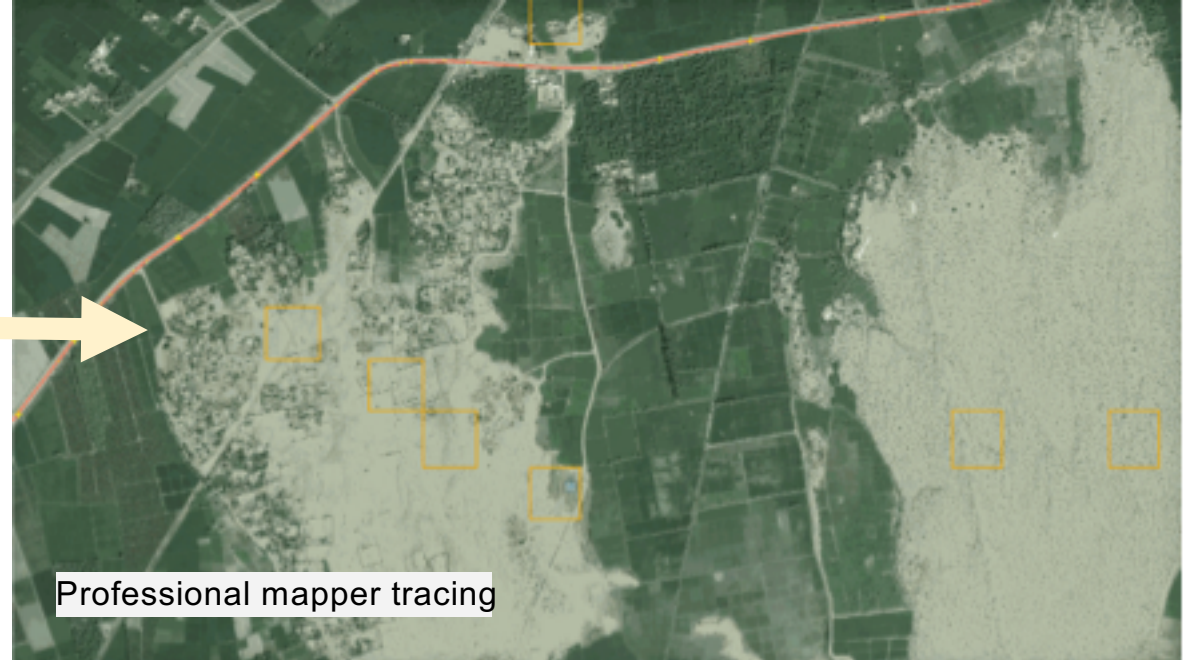


- **‘Fully automated’** machine learning pipeline;
- **Semi-automated** machine learning that will require our professional mappers’ QA and mapping objects.

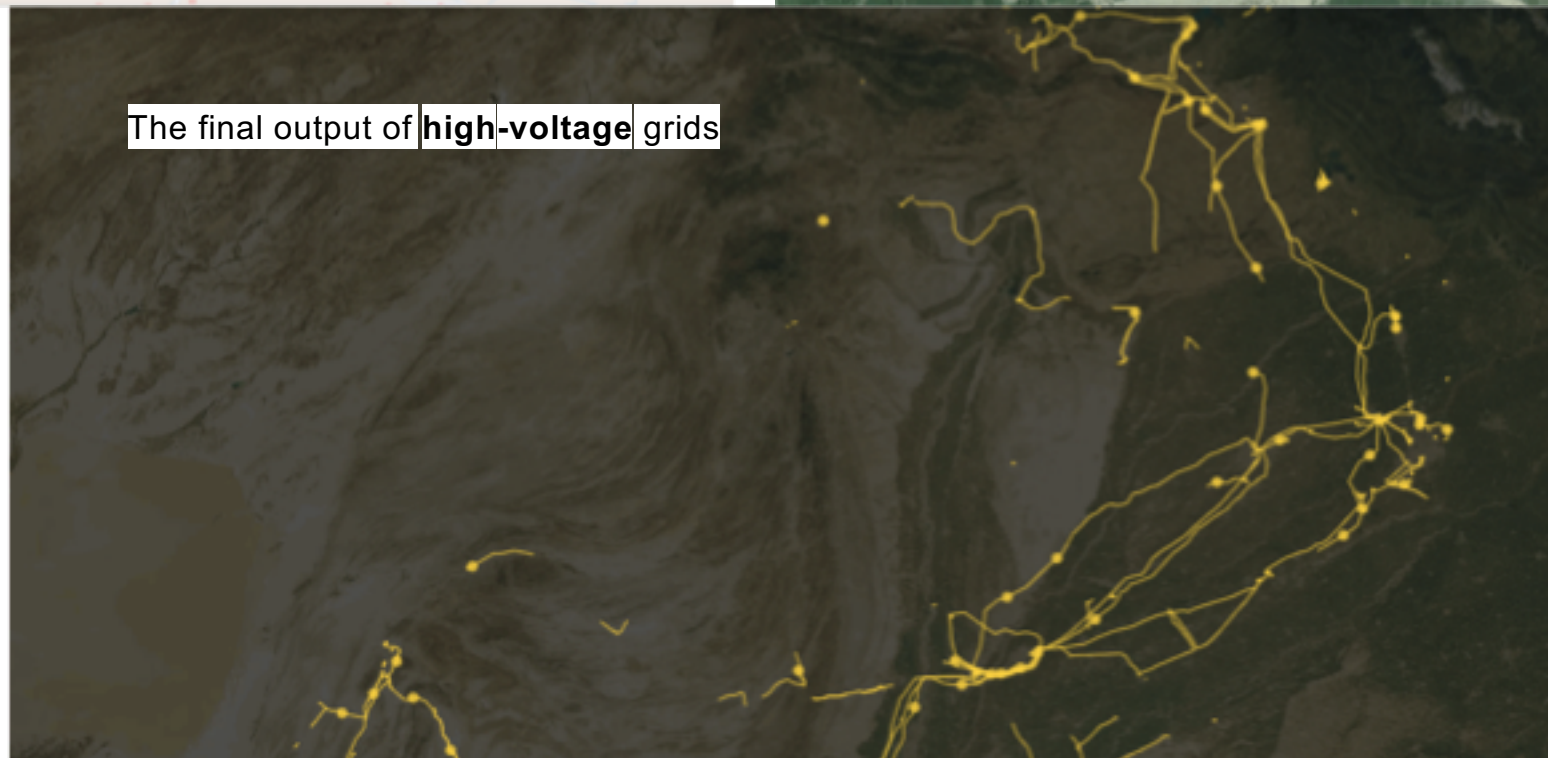




Machine learning outputs



Professional mapper tracing



The final output of **high-voltage** grids





“100% Convolutional Neural Network”

Disclaimer – lots of human “training or Indexing”

“100% Amazon”

Disclaimer – sometimes “fool around” with on-premise Gaming GPUs

All production - Storage, CPU, GPU, Products up on
AWS

2

Running models and **HPC**

HPC Workloads in the Cloud

Life Sciences



Financial Services



Energy & Geo Sciences



Design & Engineering



Media & Entertainment

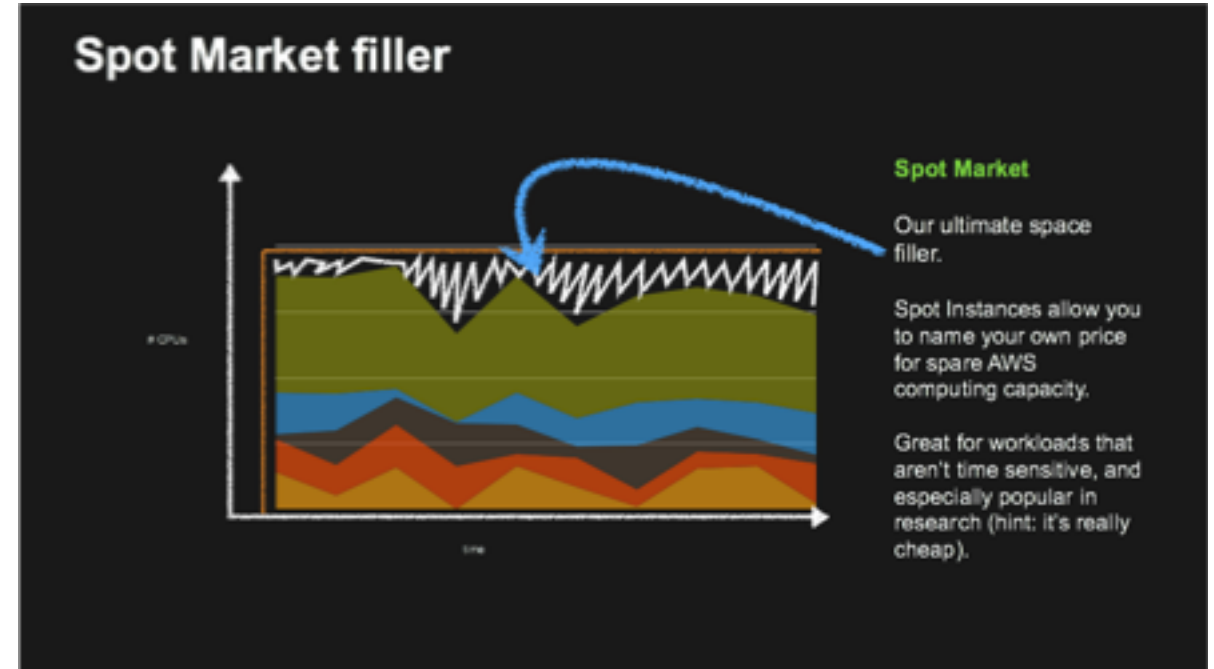
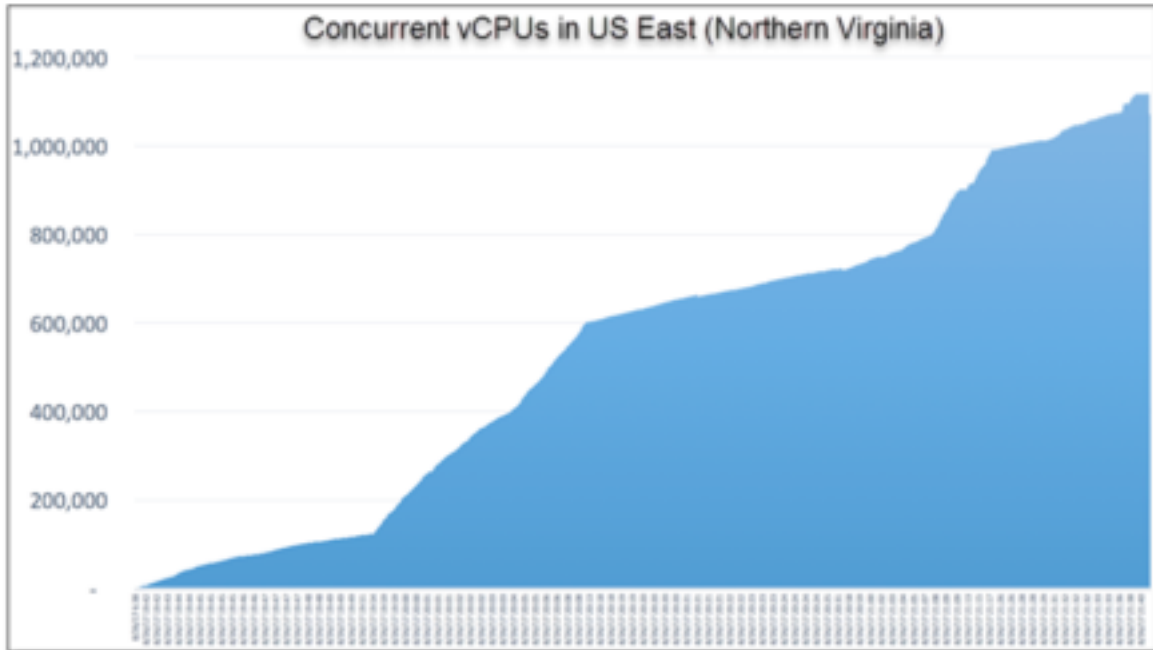


Autonomous Vehicles



Elasticity: Natural Language Processing at **Clemson University**

550,000 cores & EC2 Spot Instances



"I am absolutely thrilled with the outcome of this experiment. The graduate students on the project [...] used resources from AWS and Omnibond and developed a new software infrastructure to perform research at a scale and time-to-completion not possible with only campus resources." – Prof. [Amy Apon](#), Co-Director of the Complex Systems, Analytics and Visualization Institute

<https://aws.amazon.com/blogs/aws/natural-language-processing-at-clemson-university-1-1-million-vcpus-ec2-spot-instances/>

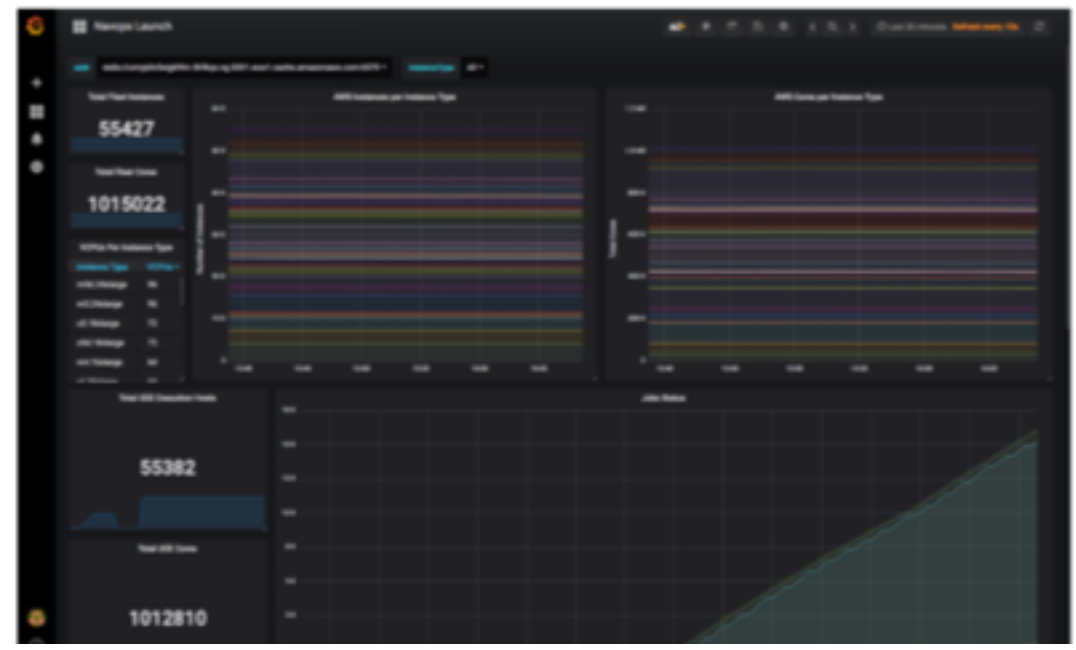
YESTERDAY – 1M cores

55

Univa Demonstrates Extreme Scale Automation by Deploying More Than One Million Cores in a Single Univa Grid Engine Cluster using AWS

June 24, 2018 Cameron Brunner, Director of Engineering, Univa

To demonstrate the unique ability to run very large enterprise HPC clusters and workloads, Univa leveraged AWS to deploy 1,015,022 cores in a single Univa Grid Engine cluster to showcase the advantages of running large-scale electronic design automation (EDA) workloads in the cloud. The cluster was built in approximately 2.5 hours using Navops Launch automation and comprised more than 55,000 AWS instances in 3 availability zones, 16 different instance types and leveraged AWS Spot Fleet technology to



<https://blogs.univa.com/2018/06/univa-demonstrates-extreme-scale-automation-by-deploying-more-than-one-million-cores-in-a-single-univa-grid-engine-cluster-using-aws/>

ou
B
at
nd
y

ire
e to
ond
ns,

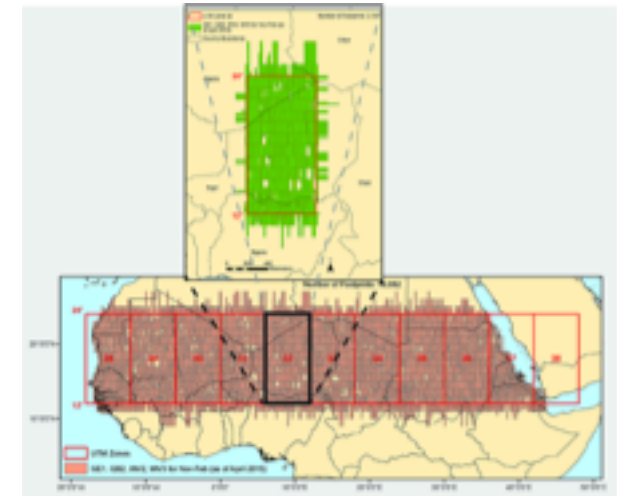
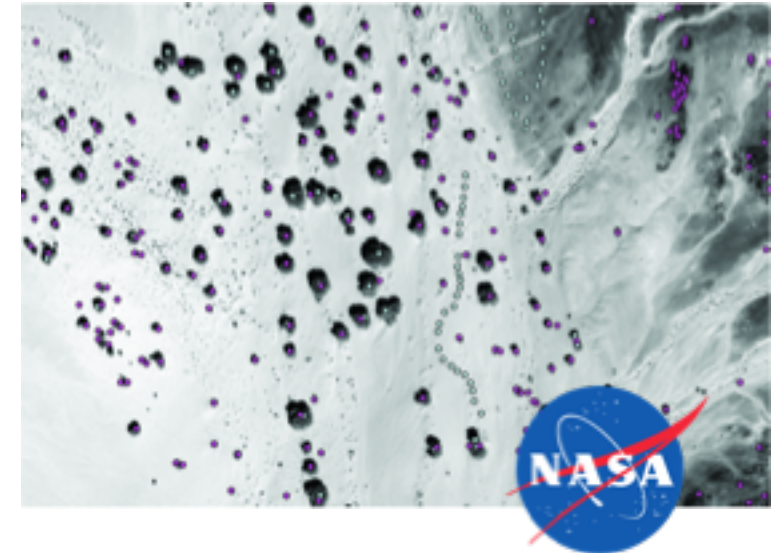


NASA – Climate Research

- Mosaicking 2,500+ QuickBird satellite images into 100-kilometer (km) x 100-km tiles, which are then broken into 25-km x 25-km sub-tiles for processing.
- Orthorectifying and mosaicking all satellite data in ADAPT
- Identifying trees and shrubs using adaptive vegetation classifier algorithms. Estimating biomass. Incorporating algorithms to calculate tree and shrub height for biomass estimates.

*The combined resources of ADAPT and AWS potentially **reduce total processing time from 10 months to less than 1 month***

Source: <https://www.nas.nasa.gov/SC15/demos/demo31.html>



Accelerators (GPU/FPGA) for HCLS: **Children's Hospital of Philadelphia**

The fastest ever analysis of 1000 genomes

- 1,000 pediatric whole genomes
- Average 40X coverage
- Max 60X coverage
- Total runtime 2h 25min
- 1000 FPGA instances

edico genome



... Available in "AWS App Store" for ~\$24 / genome

WRF in the Cloud Using Amazon Web Services



Welcome to the WRF in the Cloud Mini-tutorial for the 2018 Joint WRF and MPAS Workshop. This tutorial will introduce the steps for running WRF in the cloud, using an Amazon Web Services (AWS) platform.

Click on a tab below for quick navigation.

Using AMI to Create Instance

Running WRF on an Instance

Save Instance as Virtual Image

http://www2.mmm.ucar.edu/wrf/OnLineTutorial/wrf_in_c



Why is Cloud Computing Important for our Community?

- ▶ WRF Taking Advantage of New Technology
- ▶ Elastic Resource Availability
- ▶ Collaboration/Sharing
- ▶ Educational Outreach
- ▶ Officially-supported WRF Version

WRF in the Cloud Using Amazon Web Services



Welcome to the WRF in the Cloud Mini-tutorial for the 2018 Joint WRF and MPAS Workshop. This tutorial will introduce the steps for running WRF in the cloud, using an Amazon Web Services (AWS) platform.

Click on a tab below for quick navigation.

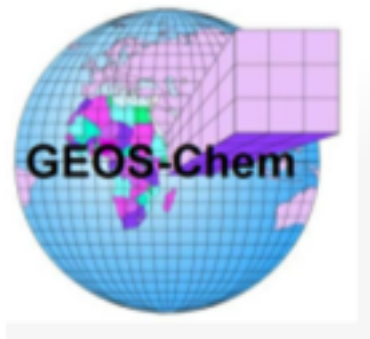
Using AMI to Create Instance

Running WRF on an Instance

Save Instance as Virtual Image

Using AMIs

http://www2.mmm.ucar.edu/wrf/OnLineTutorial/wrf_in_cloud_aws_tutorial.php



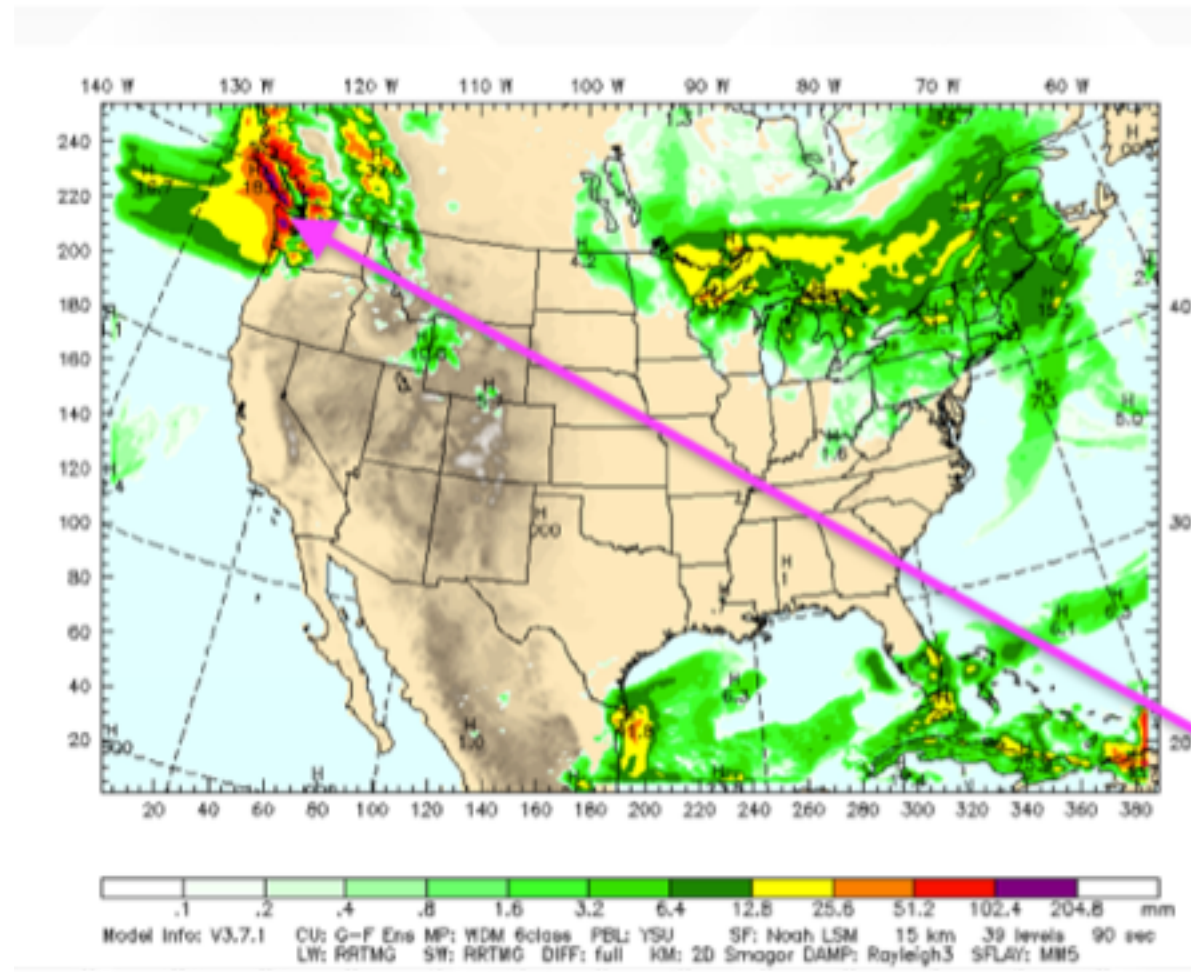
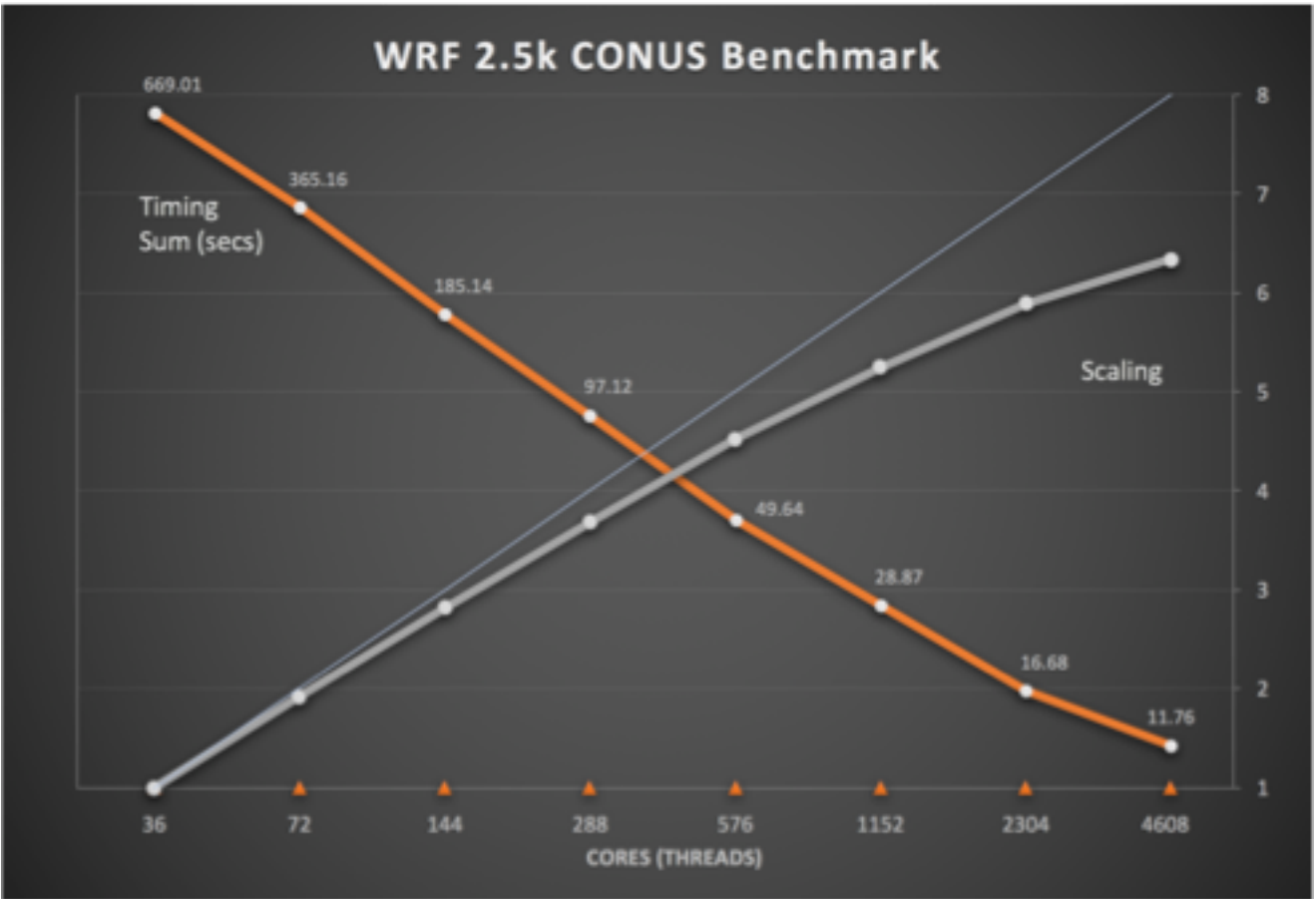
Step 2: Launch a server with GEOS-Chem pre-installed

Log in to AWS console, and click on EC2 (Elastic Compute Cloud), which is the most basic cloud computing service.

AWS services

http://cloud-gc.readthedocs.io/en/latest/chapter02_beginner-tutorial/quick-start.html#quick-start-label

WRF Weather Prediction



WRF Scaling and Performance on AWS

Weather and climate models are popular workloads on AWS:
Researchers, businesses (The Weather Channel), financial sector, ...

The AWS Console (in your web browser)

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: Version Control for Source Code
- 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 Core: Connect and Manage IoT Resources

Gaming

- AWS GameLift: Managed Game Server Hosting

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 Calendar Service

Common Services for HPC Applications

Resource Groups

Resource Group is a collection of resources that share tags. Create a group for each project, environment in your account.

Group Tag Editor

Additional Resources

- Getting Started: Read our documentation or view our training to learn more about AWS.
- AWS Console Mobile App: View your resources on the go with our AWS Console mobile app, available from Amazon Appstore, Google Play, or iTunes.
- AWS Marketplace: Find and buy software, launch with 1-Click and pay by the hour.
- AWS re:Invent Announcements: Explore the next generation of AWS cloud capabilities. See what's new.

Service Health

All services operating normally.
Updated: Jul 15 2016 14:01:01 GMT-0700
Service Health Dashboard

Basics

Compute



Spectrum of Compute Instance Types

General purpose



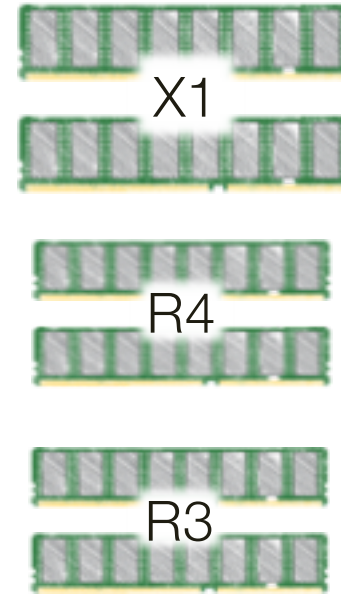
Compute optimized



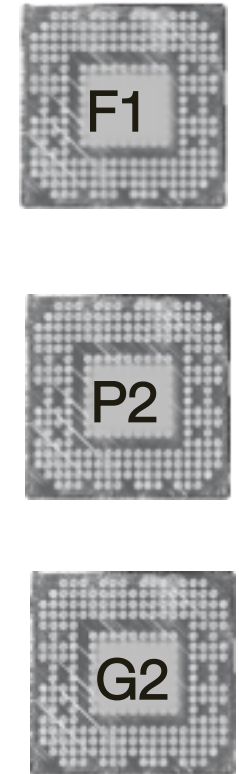
Storage and IO optimized



Memory optimized



GPU or FPGA enabled



Selecting an instance type

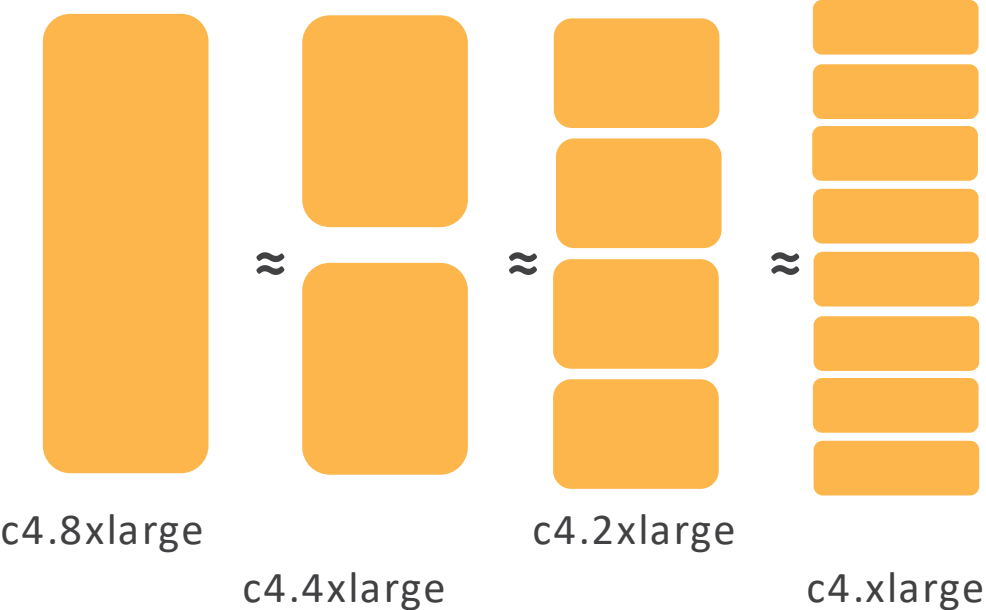
| Instance Type | vCPU | Memory (GiB) | Storage (GB) | Networking Performance | Physical Processor | Clock Speed (GHz) | EBS Opt |
|---------------|------|--------------|---------------|------------------------|-----------------------|-------------------|---------|
| c4.8xlarge | 36 | 60 | EBS Only | 10 Gigabit | Intel Xeon E5-2666 v3 | 2.9 | Yes |
| c3.8xlarge | 32 | 60 | 2 x 320 SSD | 10 Gigabit | Intel Xeon E5-2680 v2 | 2.8 | No |
| m4.10xlarge | 40 | 160 | EBS Only | 10 Gigabit | Intel Xeon E5-2676 v3 | 2.4 | Yes |
| m4.16xlarge | 64 | 256 | EBS Only | 20 Gigabit | Intel Xeon E5-2686 v4 | 2.3 | Yes |
| p2.16xlarge | 64 | 732 | EBS Only | 20 Gigabit | Intel Xeon E5-2686 v4 | 2.3 | Yes |
| x1.32xlarge | 128 | 1,952 | 2 x 1,920 SSD | 20 Gigabit | Intel Xeon E7-8880 v3 | 2.3 | Yes |
| r3.8xlarge | 32 | 244 | 2 x 320 SSD | 10 Gigabit | Intel Xeon E5-2670 v2 | 2.5 | No |

CAREFUL: a “vCPU” is a hyperthread, i.e. ½ of a physical core.

C4.8xlarge has 36 vCPU but 18 physical cores, the way HPC practitioners usually count them.

<https://aws.amazon.com/ec2/instance-types/#instance-type-matrix>

Amazon EC2 Instances



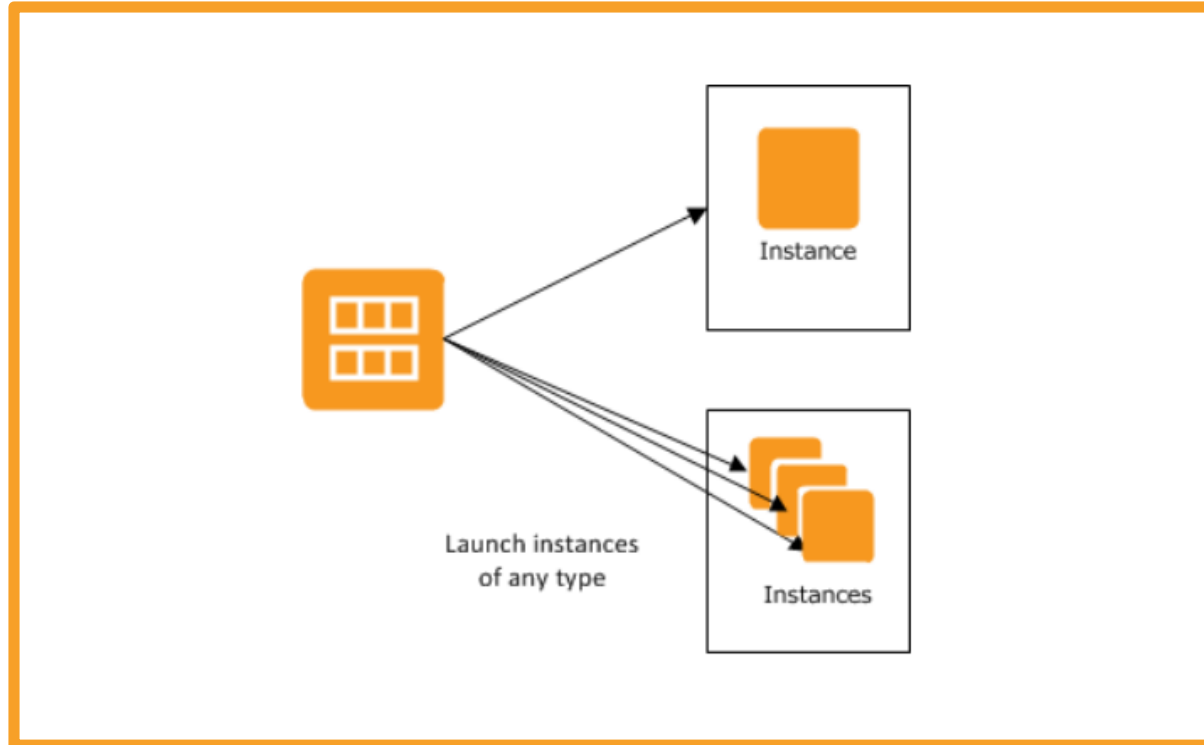
Instance
Generation

c4.large

Instance
Family

Instance
Size

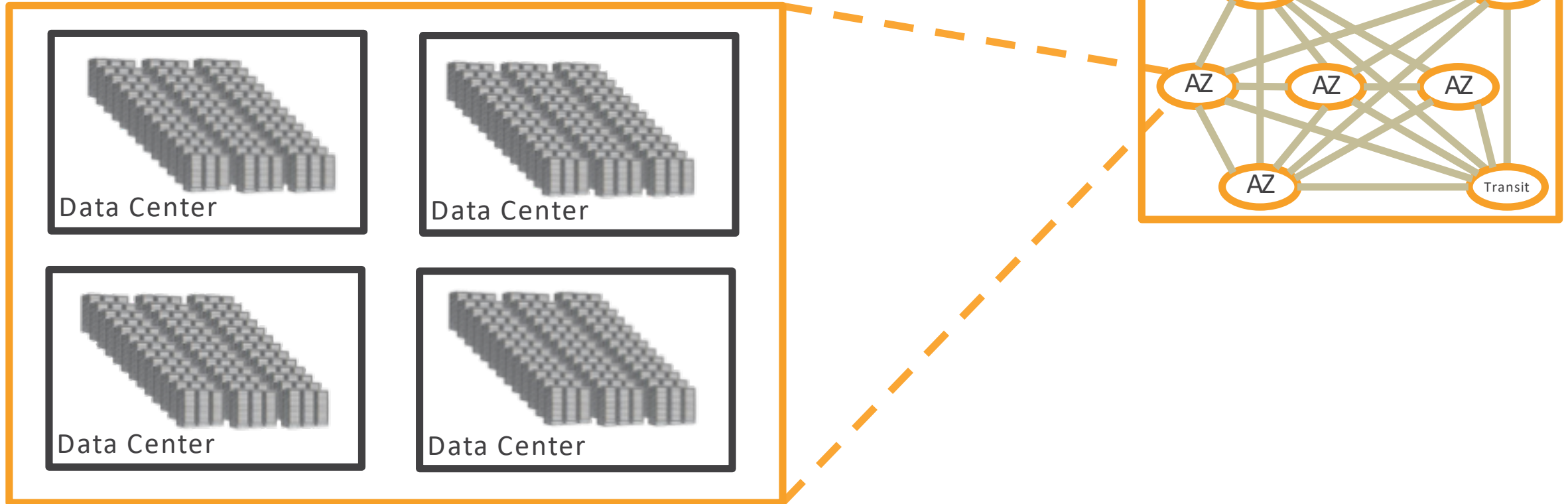
Launching an Instance from an AWS Machine Images (AMI)



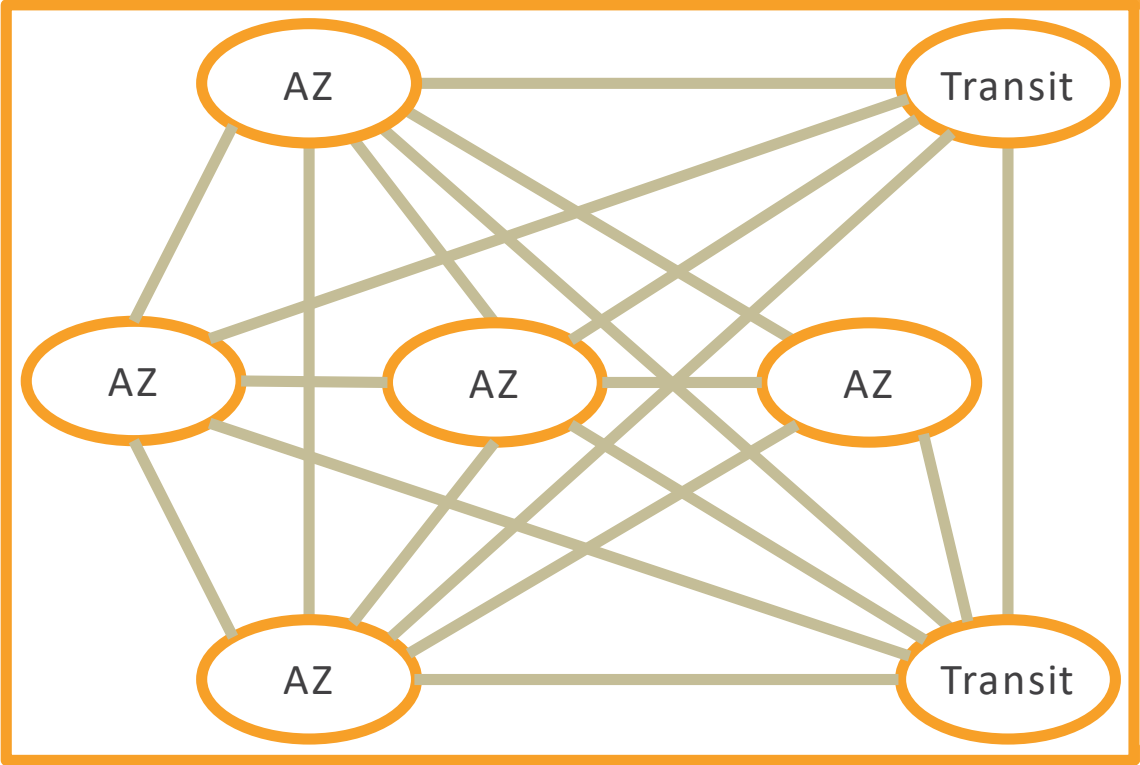
Instance == virtual server

AMI == virtual machine image

An AWS Availability Zone



An AWS Region



The AWS Global Infrastructure



Basics

Storage



AWS Storage Options for HPC Workloads

EFS

Highly available, multi-AZ, fully managed network-attached elastic file system.

For near-line, highly-available storage of files in a traditional NFS format (NFSv4).

Use for read-often, temporary working storage

EC2+EBS

Block storage device (SSD or HDD) for file system attached to EC2 instance. Can build parallel file system (e.g., using Intel Lustre).

For near-line storage of files optimized for high I/O performance.

Use for high-IOPs, temporary working storage

Amazon S3

Secure, durable, highly-scalable object storage. Fast access, low cost.

For long-term durable storage of data, in a readily accessible get/put access format.

Primary durable and scalable storage for HPC data

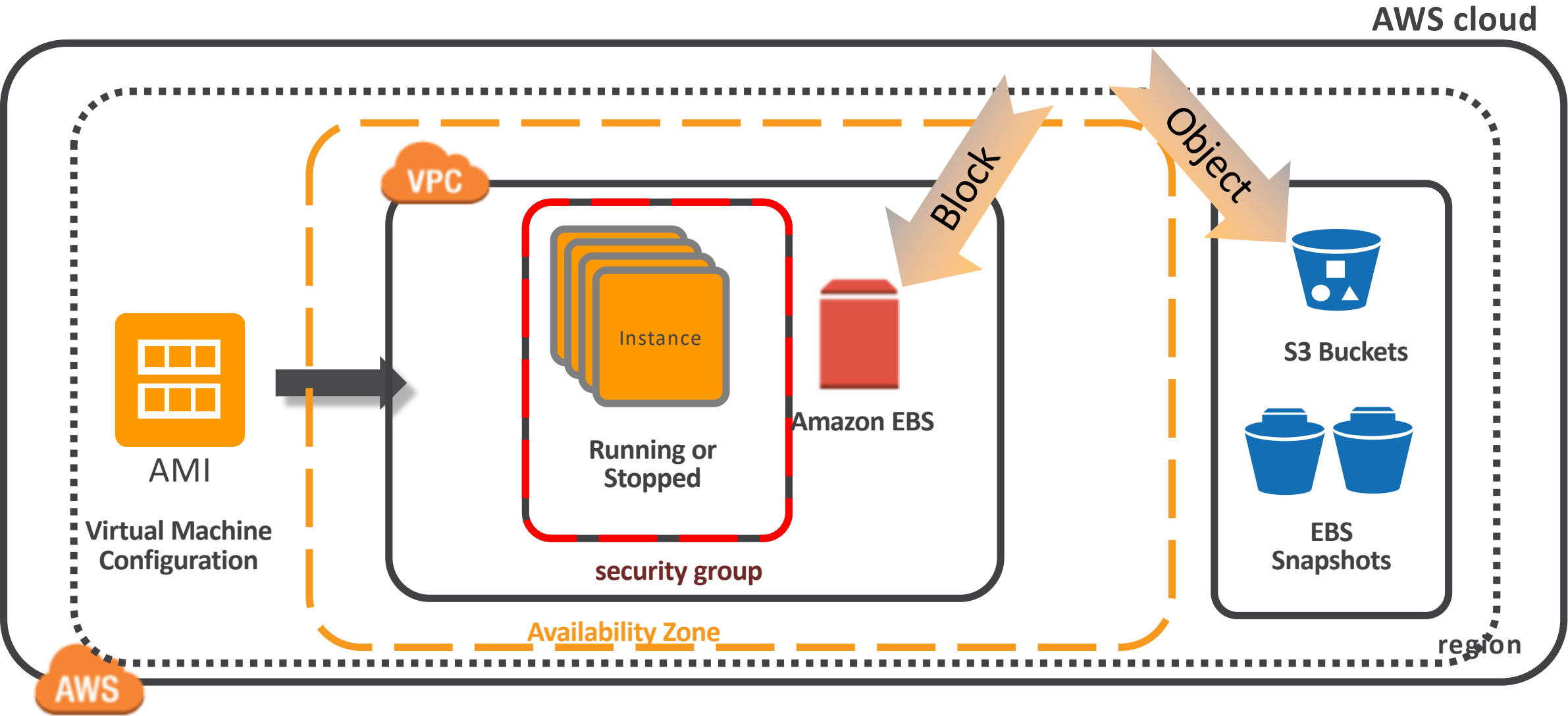
Amazon Glacier

Secure, durable, long term, highly cost-effective object storage.

For long-term storage and archival of data that is infrequently accessed.

Use for long-term, lower-cost archival of HPC data

Combining Compute and Storage

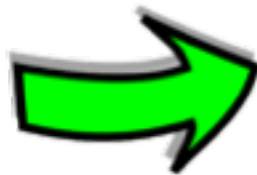


Basics

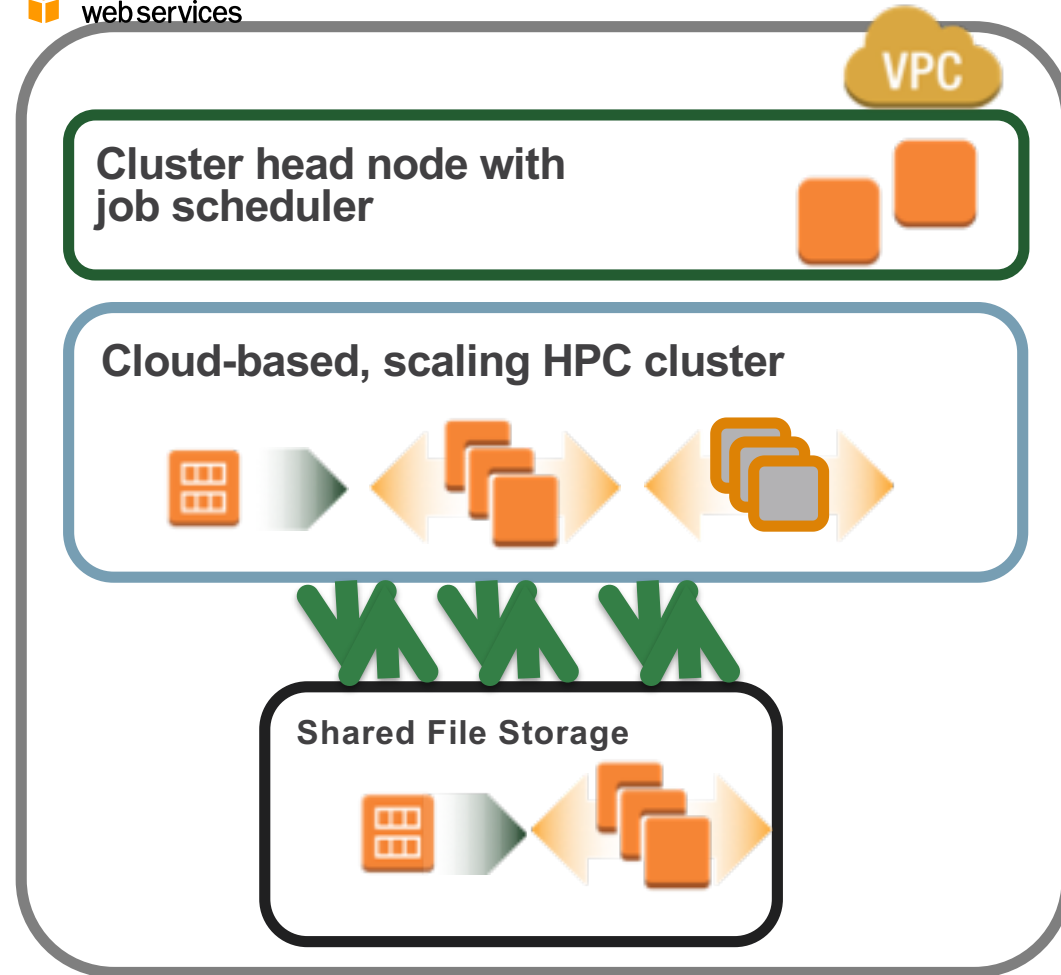
Network



We can build a cluster:



Using a compute cluster in the cloud



Amazon S3
and
Amazon
Glacier



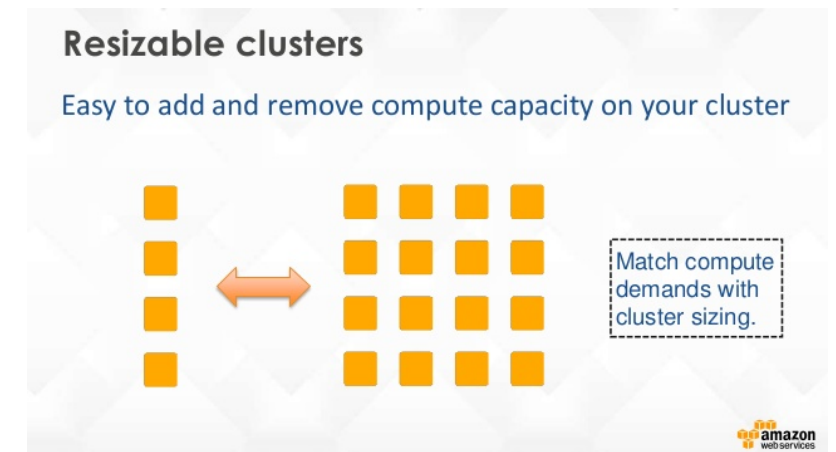
Thin or Zero Client
- No local data -

Using a compute cluster in the cloud

Self-scaling HPC clusters instantly ready to compute, billed by the hour and use the AWS Spot market by default, so they're very low cost

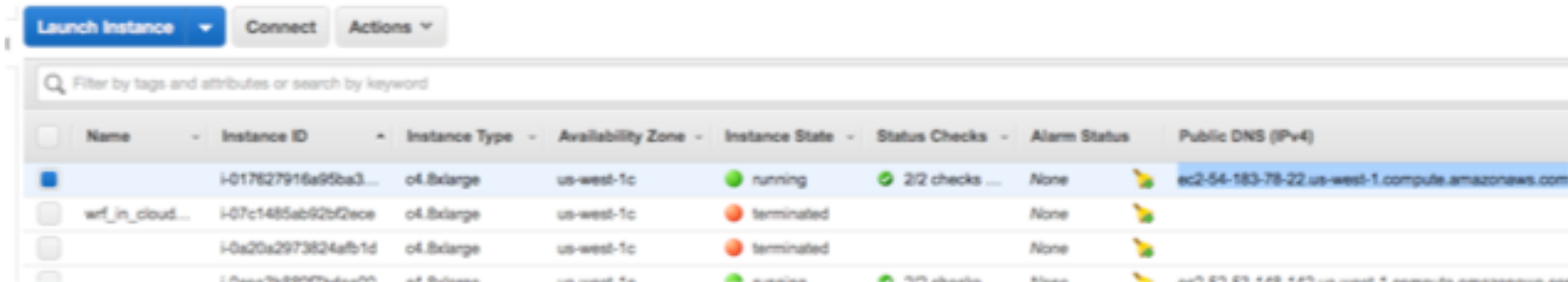


- Popular scientific applications prepackaged
- Deploys in ~5 minutes.
- Familiar job schedulers, scientific applications, and shared file system.
- Install any software you need.
- No job queues – it's your personal cluster.
- Access to the graphical console.
- Deploys in minutes.
- Scales as large as needed when you add jobs to the queue, and scales back down when the jobs are done.



Controlling your AWS resources

- 1. Web browser (point-and-click)



The screenshot shows the AWS Management Console interface for EC2 instances. At the top, there are buttons for 'Launch Instance', 'Connect', and 'Actions'. Below that is a search bar with the text 'Filter by tags and attributes or search by keyword'. The main area displays a table of instances with the following columns: Name, Instance ID, Instance Type, Availability Zone, Instance State, Status Checks, Alarm Status, and Public DNS (IPv4). The first instance is highlighted in blue and is in a 'running' state. The other two visible instances are in a 'terminated' state.

| Name | Instance ID | Instance Type | Availability Zone | Instance State | Status Checks | Alarm Status | Public DNS (IPv4) |
|----------------|----------------------|---------------|-------------------|----------------|----------------|--------------|--|
| | i-017627916a95ba3... | c4.xlarge | us-west-1c | running | 2/2 checks ... | None | ec2-54-183-78-22.us-west-1.compute.amazonaws.com |
| wf_in_cloud... | i-07c1485ab92b2e3e | c4.xlarge | us-west-1c | terminated | | None | |
| | i-0a20a2973824afb1d | c4.xlarge | us-west-1c | terminated | | None | |

- 2. Command-line interface (script, automate)

```
~$ aws s3 cp myvideo.mp4 s3://mybucket/
```

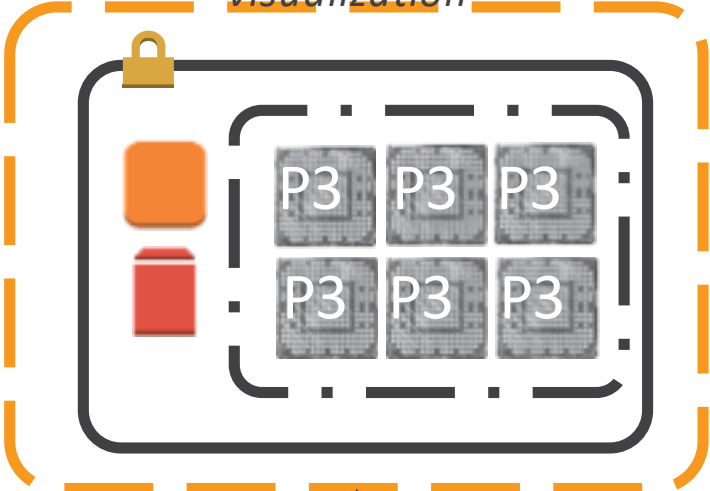
<https://aws.amazon.com/cli/>

- 3. SDKs (GUIs, platforms, science gateways)

Compute clusters in the cloud are fit for purpose

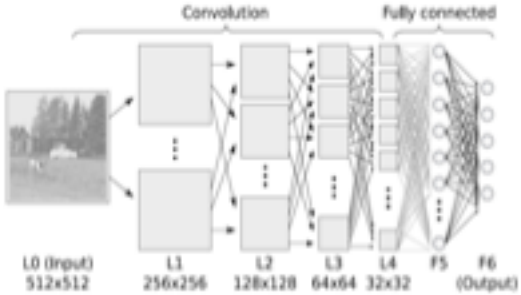
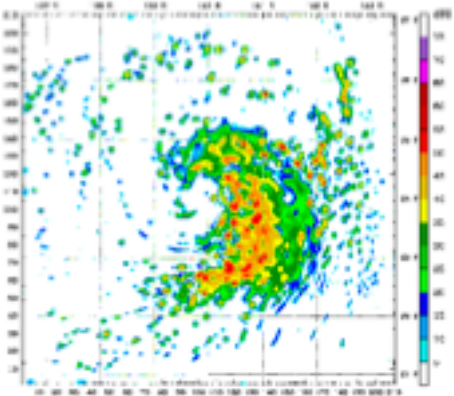
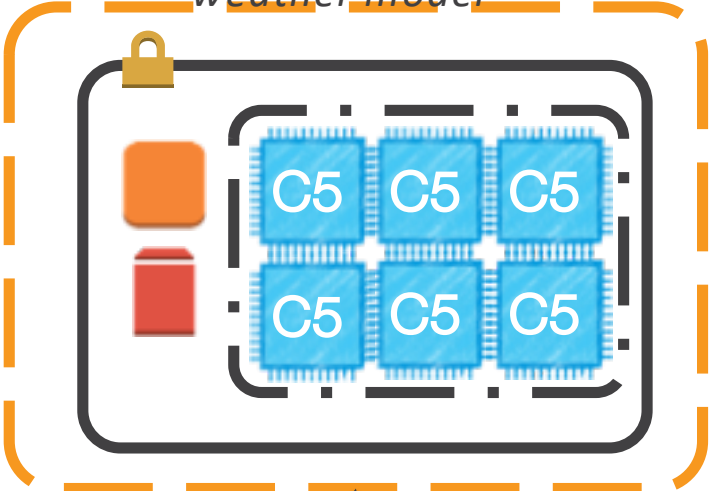
GPU cluster

visualization



CPU cluster

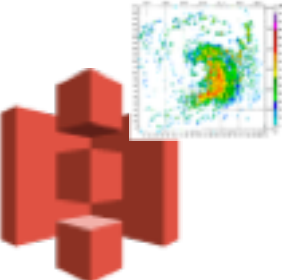
weather model



Amazon S3

storage of input/output

Compute clusters in the cloud are ephemeral

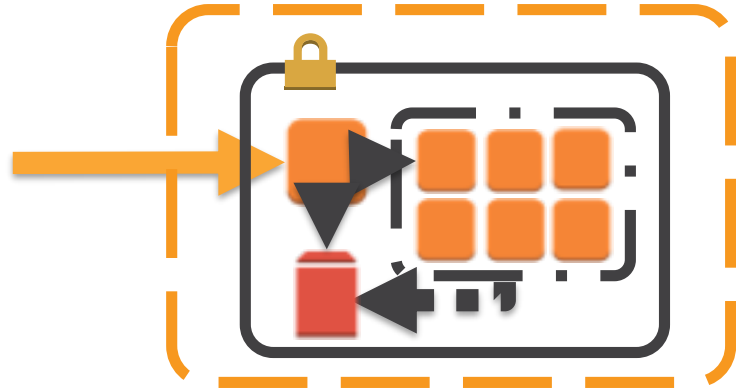


Amazon S3

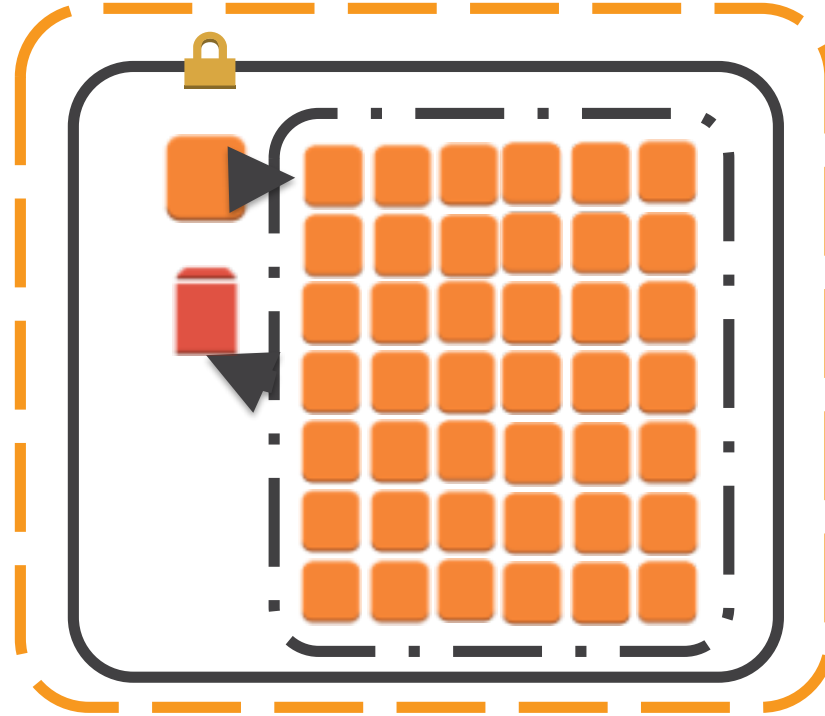
storage of input/output

Compute clusters in the cloud are elastic

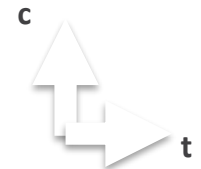
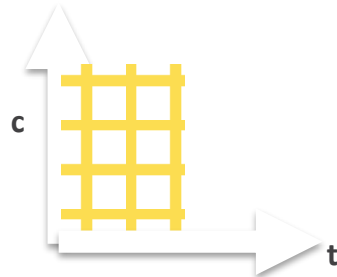
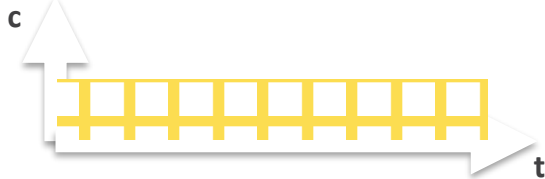
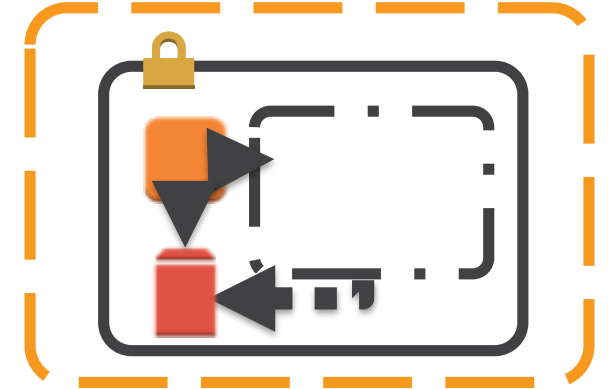
morning



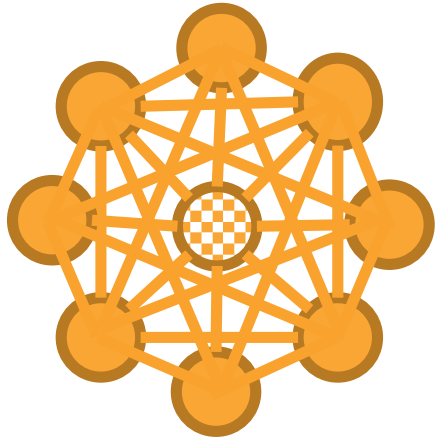
afternoon



evening



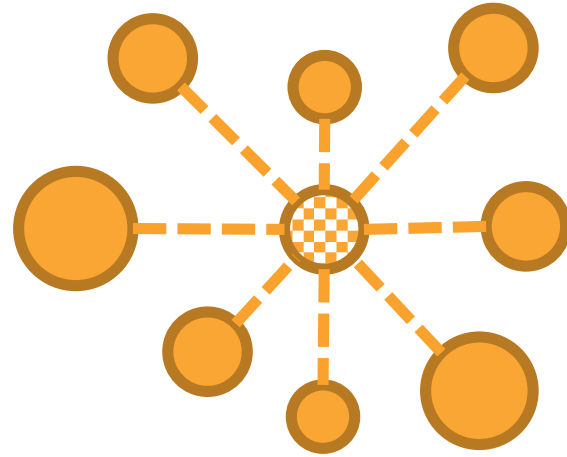
Tightly and loosely coupled workloads



Cluster HPC

Tightly coupled,
latency sensitive
applications

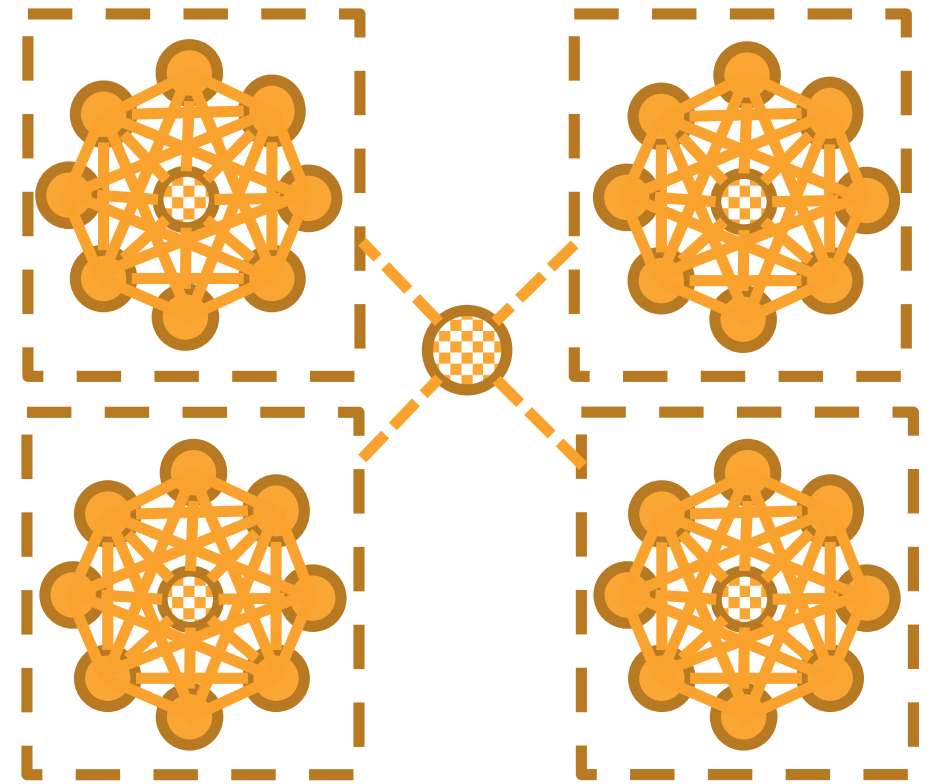
Use larger EC2
compute instances,
placement groups,
enhanced networking



Grid HPC

Loosely coupled, HTC,
pleasingly parallel

Use a variety of EC2
instances, multiple AZs,
Spot, Auto Scaling,
Amazon SQS



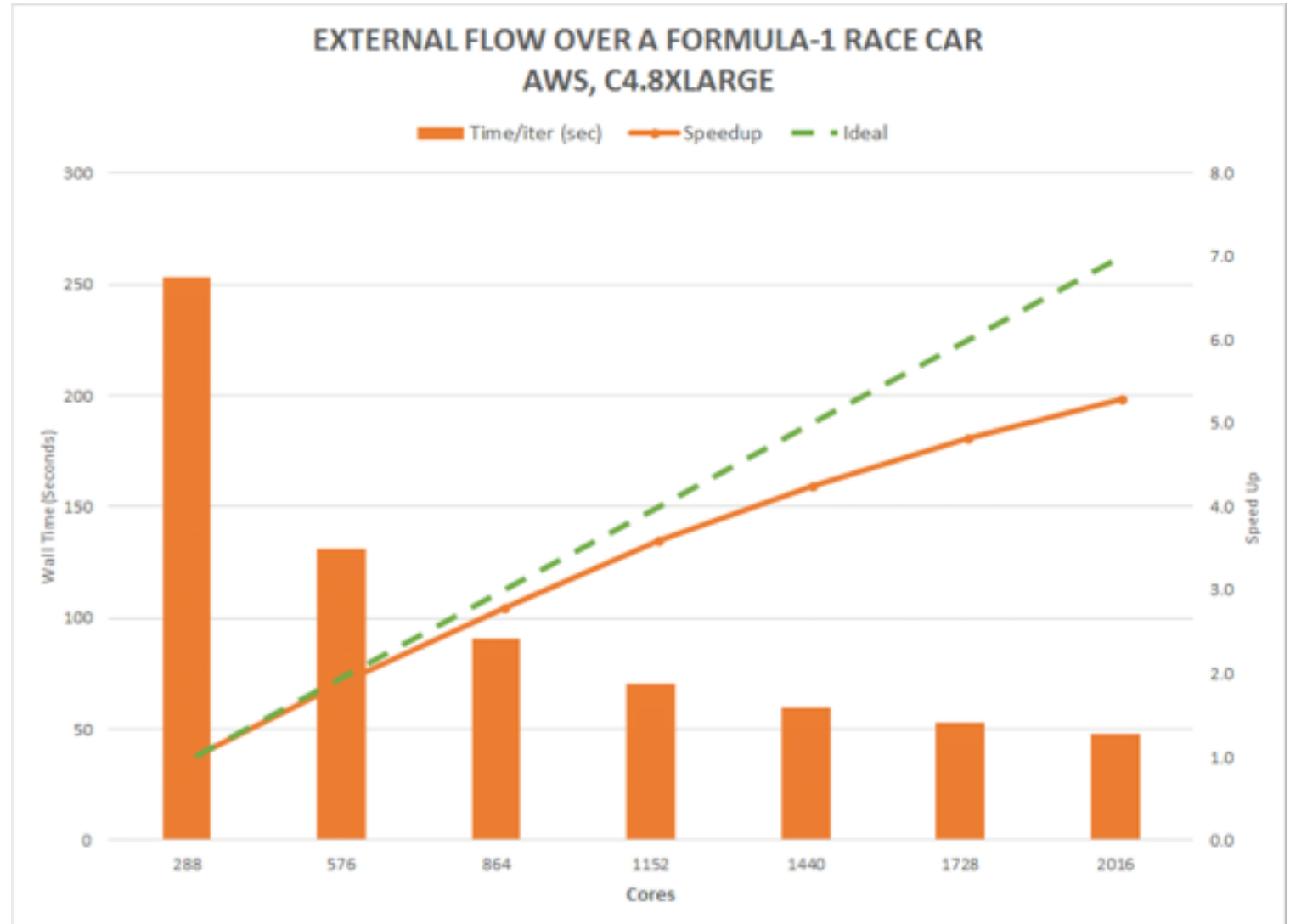
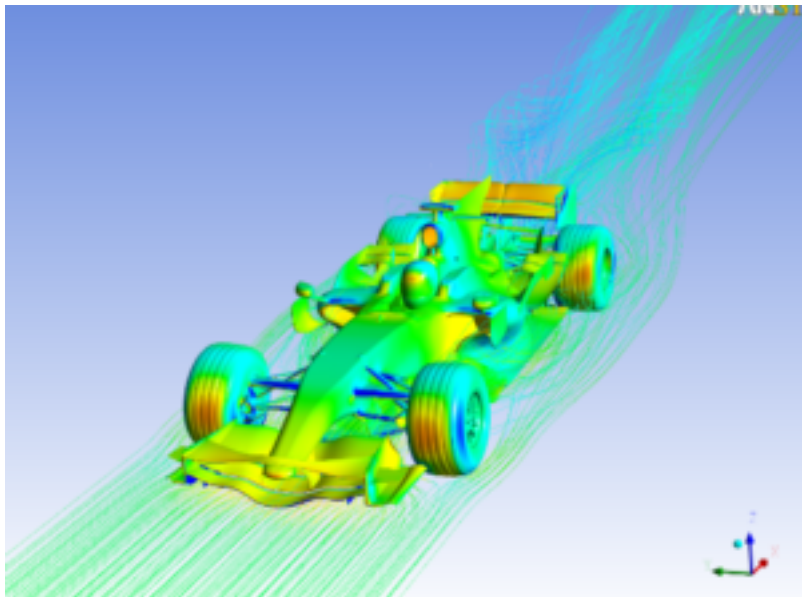
Ensemble?

Run all members at once!

Performance for Fluid Dynamics on AWS

ANSYS Fluent

- AWS c4.8xlarge
- 140M cells
- F1 car CFD benchmark



<http://www.ansys-blog.com/simulation-on-the-cloud/>

3



Data Lakes and Collaboration

Collaborating on scientific data in the cloud

It's typically time-consuming and expensive to acquire, store, and analyze large data sets.

Sharing data on AWS makes it accessible to a large and growing community of researchers who use the AWS cloud.

- AWS is built from the ground up with sophisticated, real-world security: share without giving up security.
- Use AWS worldwide network and data centers to reach your collaborators.
- Collaborators can analyze your shared data in their own account, and run your shared applications in their own account, at their own expense.
- Not necessary for everyone to download a copy of the dataset: everyone can bring analytics to central copy.
- You retain full ownership. Data never leaves a country (“data residency”) unless you explicitly move it.

Global Platform for Global Collaboration



18
Regions
52
Availability
Zones

As of Jan 25, 2018

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

Bring the users and compute to the data; don't send the data to the users.

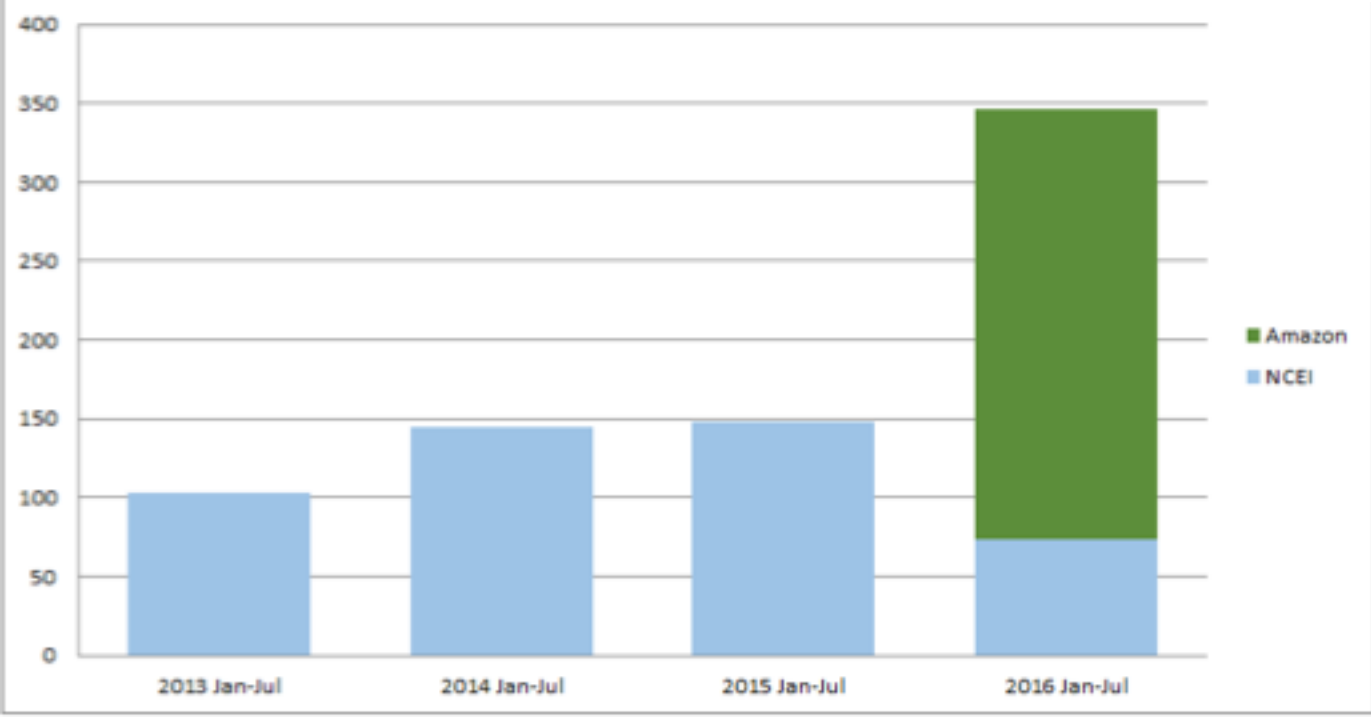
Collaborating on scientific data in the cloud



Collaborating on scientific data in the cloud



NOAA- NEXRAD on AWS S3, usage increased 2.3



Public Datasets on AWS

To stimulate innovation, AWS hosts a selection of datasets that anyone can access for free. Data in our public datasets is available for rapid access to our flexible and low-cost computing resources.



Life Science

- **TCGA & ICGC** (used at OICR)
- 1000 Genomes
- Genome in a Bottle
- Human Microbiome Project
- 3000 Rice Genome



Earth Science

- Landsat
- NEXRAD
- NASA NEX



Internet Science

- Common Crawl Corpus
- Google Books Ngrams
- Multimedia Commons

Collaborating on scientific data in the cloud

AWS hosts a selection of public datasets that anyone can access for free.

Earth on AWS

Build planetary-scale applications in the cloud with open geospatial data.

aws.amazon.com/earth

<https://registry.opendata.aws>

Climate
Models

Aerial
Imagery

Elevation
Models

Satellite
Imagery

High-resolution
Radar

Registry of Open Data on AWS (RODA)

Registry of Open Data on AWS



About

This registry exists to help people discover and share datasets that are available via AWS resources. [Learn more about sharing data on AWS.](#)

See [all usage examples for datasets listed in this registry.](#)

Search datasets (currently 59 matching datasets)

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the [Registry of Open Data on AWS GitHub repository.](#)

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

Sentinel-2

[earth observation](#) [satellite imagery](#) [gis](#) [natural resource](#) [sustainability](#)

The [Sentinel-2 mission](#) is a land monitoring constellation of two satellites that provide high resolution optical imagery and provide continuity for the current SPOT and Landsat missions. The mission provides a global coverage of the Earth's land surface every 5 days, making the data of great use in on-going studies. L1C data are available from June 2015 globally. L2A data are available from April 2017 over wider Europe region, planned to be expanded globally in July 2018.

[Details →](#)

Usage examples

- [Sterling Geo Using Sentinel-2 on Amazon Web Services to Create NDVI by Sterling Geo](#)
- [Satellite Search by Remote Pixel by Remote Pixel](#)
- [Sentinel Hub WMS/WMTS/WCS Service by Sinergise](#)
- [Spectator - tracking Sentinel 2, accessing the data and quick preview by Spectator](#)
- [Sentinel Playground by Sinergise](#)

[See 15 usage examples →](#)

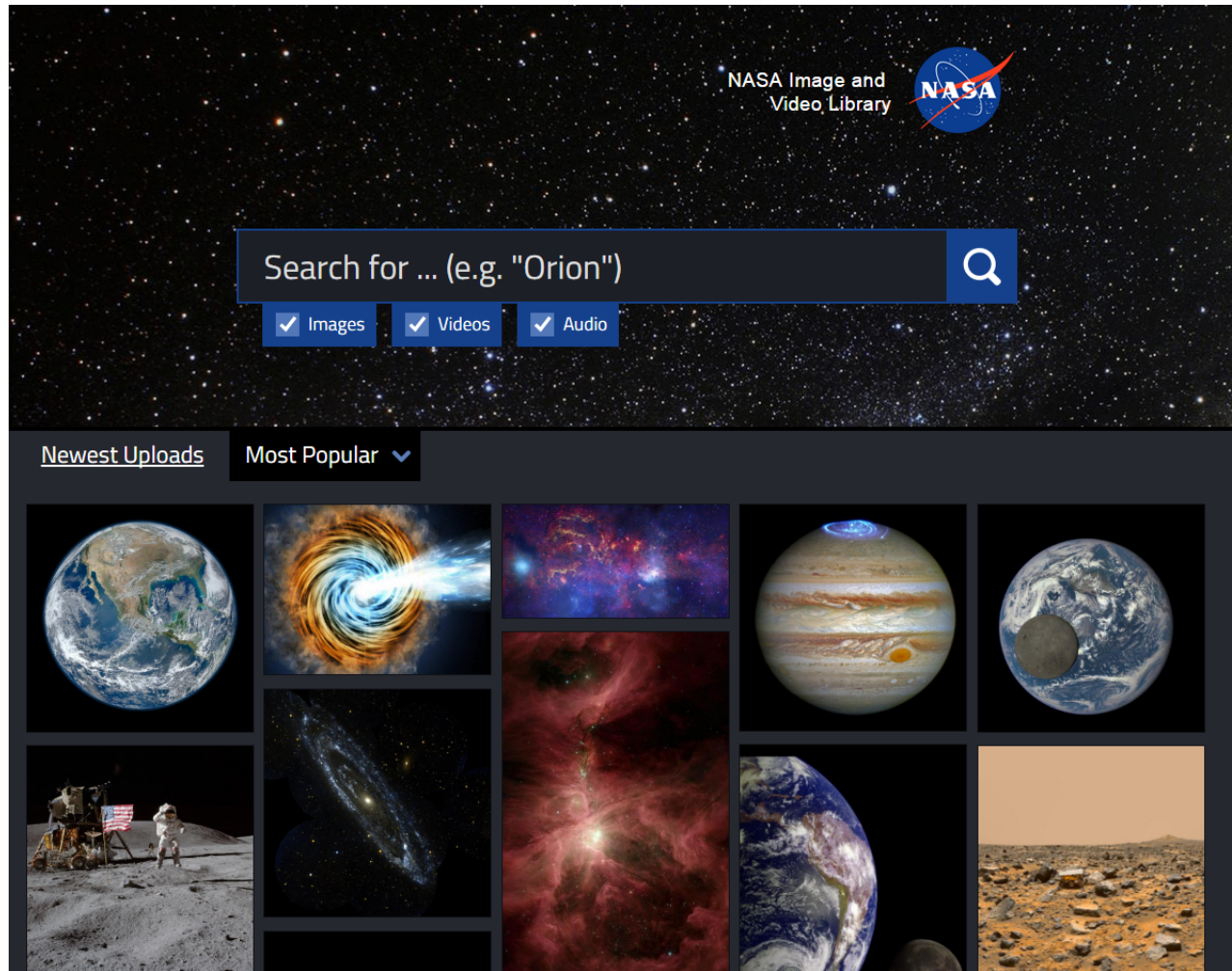
Landsat 8

[earth observation](#) [satellite imagery](#) [gis](#) [natural resource](#) [sustainability](#)

An ongoing collection of satellite imagery of all land on Earth produced by the Landsat 8 satellite.

[Details →](#)

NASA Image and Video Library (2017)



- Easy Access to the Wonders of Space. Fully compliant with Section 508 of the Rehabilitation Act.
- Built-in **Scalability**. “On-demand scalability will be invaluable for events such as the solar eclipse that’s happening later this summer—both as we upload new media and as the public comes to view that content,” says Bryan Walls, Imagery Experts Deputy Program Manager at NASA.
- Good Use of Taxpayer Dollars. By building its Image and Video Library in the cloud, NASA **avoided the costs** associated with deploying and maintaining server and storage hardware in-house. Instead, the agency can simply pay for the AWS resources it uses at any given time.

<https://aws.amazon.com/partners/success/nasa-image-library/>

U.K. Met Office Uses AWS to Deliver Tailored Meteorological Data

“

“We are using the AWS Cloud to drive the mass-market availability of customizable weather information.

James Tomkins

Head of Enterprise IT Architecture
Met Office



”

The Met Office has been a widely respected national weather service in the United Kingdom for 160 years.

- Needed the means to send weather data to device users and third-party customers.
- Deployed Amazon ElastiCache to respond to peak demands.
- Attracted more than half a million users with its **WeatherCloud** app.
- Scaled data storage tenfold and reduced solution costs by 50 percent.
- Enabled innovation of big data services in a competitive landscape.

<https://aws.amazon.com/solutions/case-studies/the-met-office/>


<https://aws.amazon.com/about-aws/whats-new/2017/08/uk-met-office-high-resolution-weather-forecast-data-is-now-on-aws/>

NIH initiatives: National Cancer Institute – Cloud Resources

Funded projects to create collaborative environments on cloud


- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

Data



- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Compute

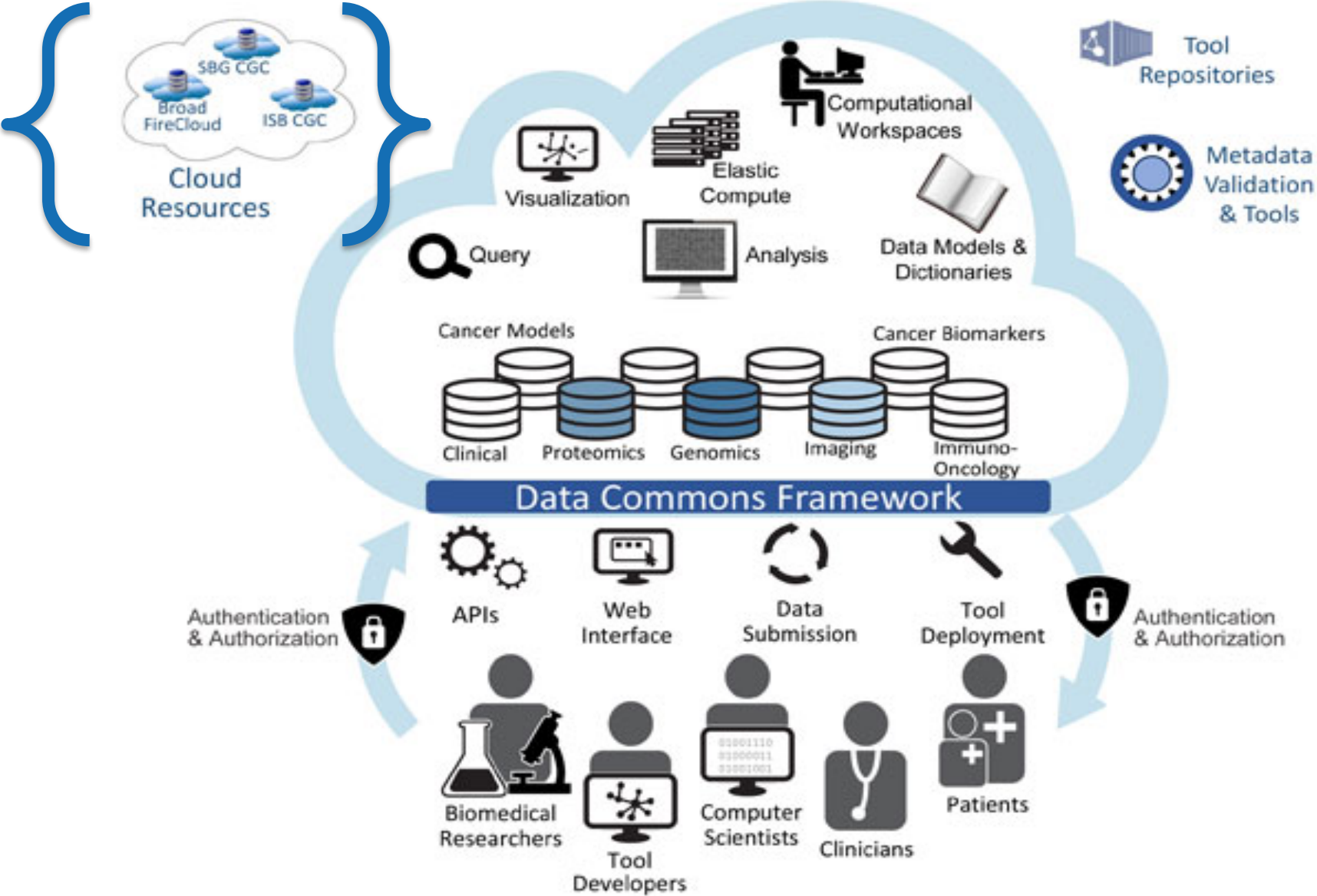


- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

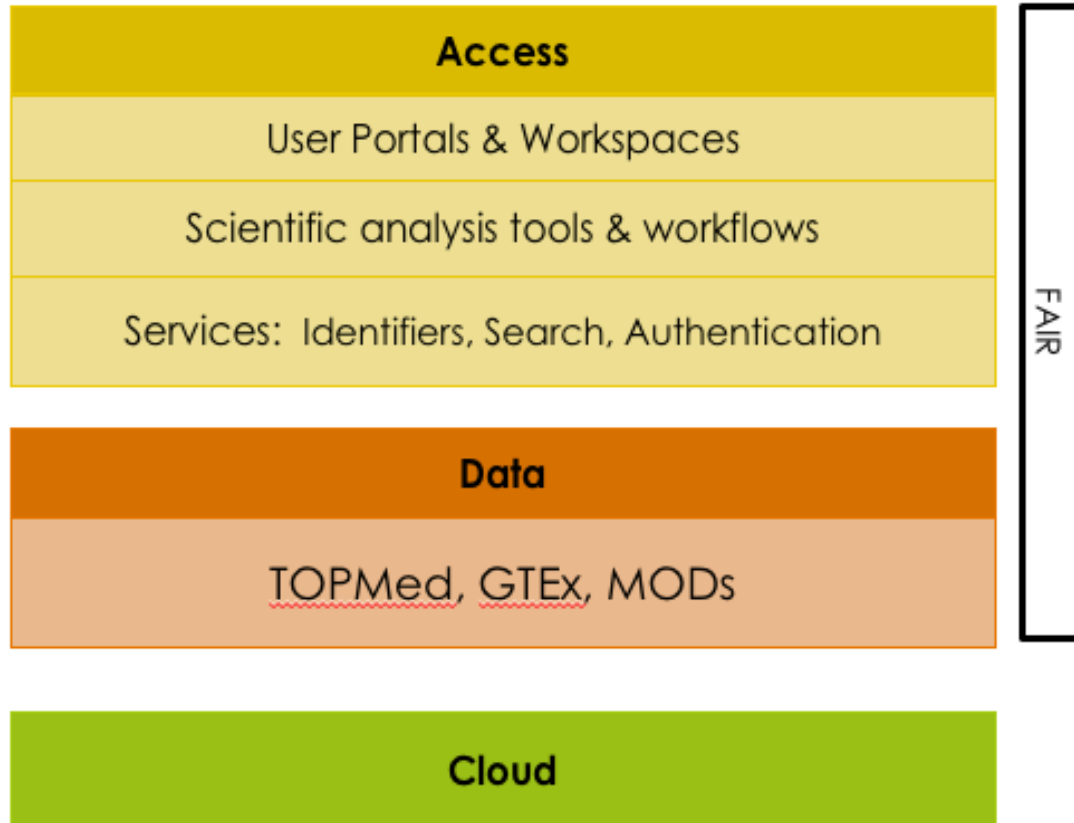
Security



NCI Cancer Research Data Commons



NIH Data Commons Pilot



- Create a research ecosystem
- Components include:
 - Computing environments (HPC, cloud)
 - Data with Common Digital Object ID's
 - Software for resource provisioning, data discovery, scientific applications and workflows

4

Containers, AWS Batch, **Microservices**

Using Containers

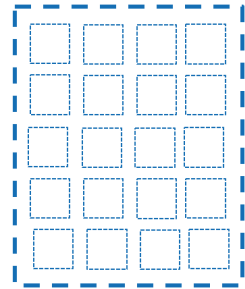
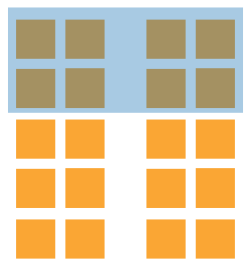


Physical

Virtualisation

Containerization

Serverless



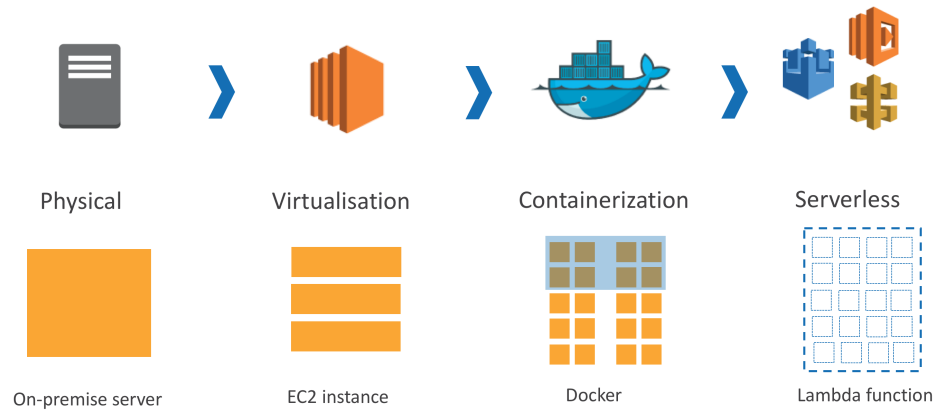
On-premise server

EC2 instance

Docker

Lambda function

Using Containers



http://bigweatherweb.org/Big_Weather_Web/Home/Home.html

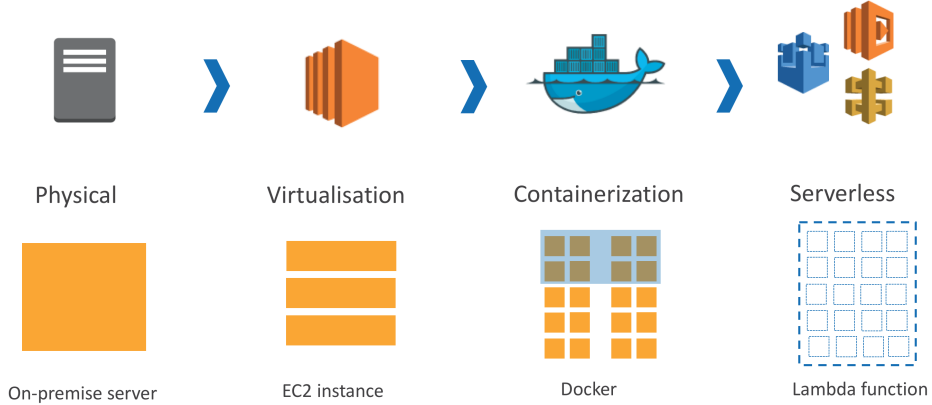
BAMS Article: A Containerized Mesoscale Model and Analysis Toolkit to Accelerate Classroom Learning, Collaborative Research, and Uncertainty Quantification

Containerized WRF available!

<https://github.com/NCAR/container-wrf>

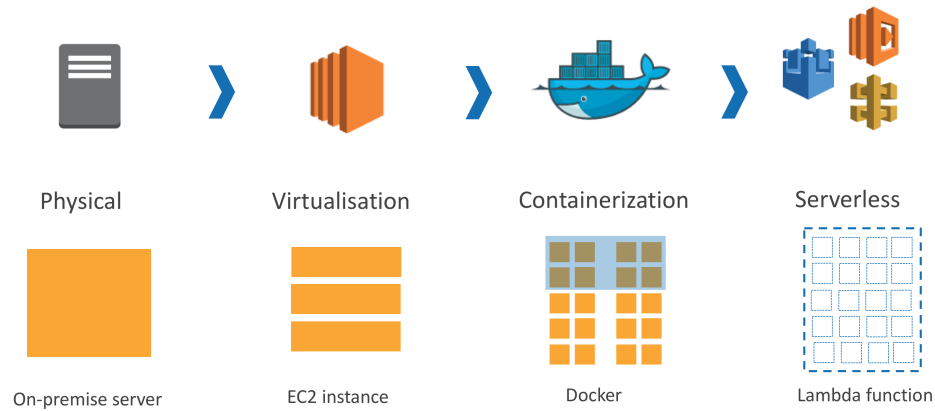
<https://hub.docker.com/r/bigwxwrf/ncar-wrf/>

Using Containers



Also see: Jupiter talk by Luke on Tuesday

Using Containers



AWS Batch – a managed service for container based jobs



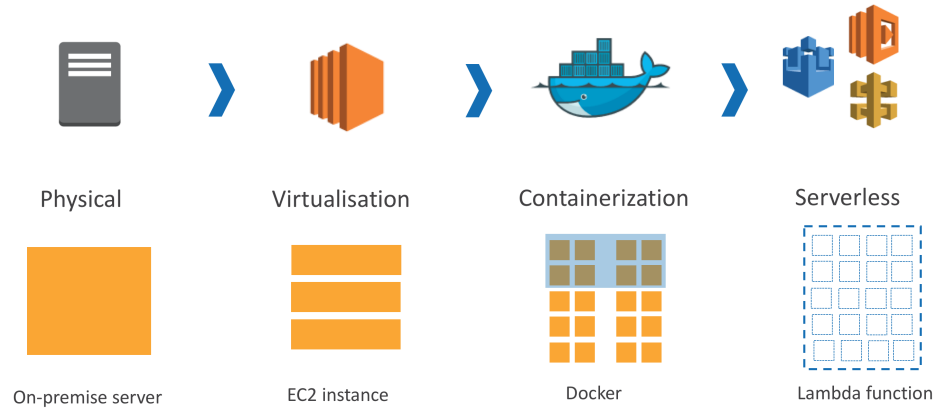
Container Based: Each job is a Docker container with runtime parameters. Submit tens to millions of jobs to a queue, with priority and job dependency options.

Fully Managed: No software to install or servers to manage. AWS Batch provisions, manages, and scales the infrastructure needed to run the jobs.

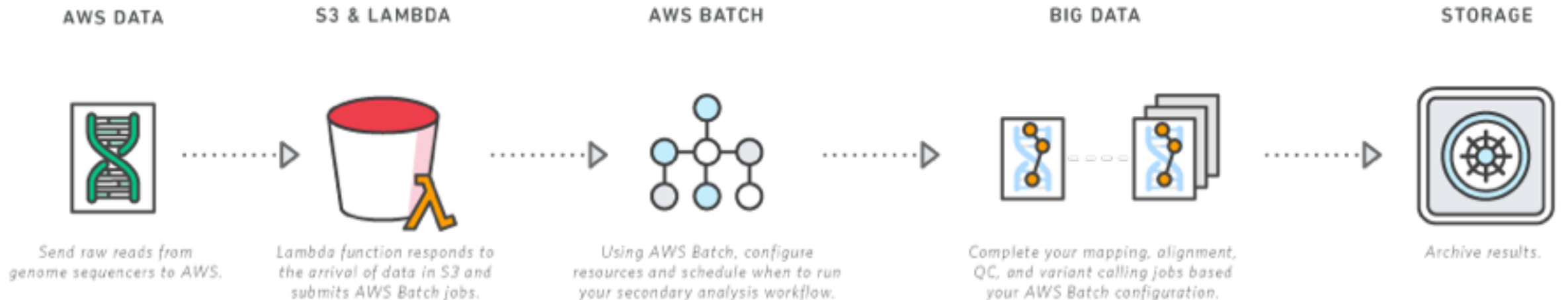
Cost optimization: use spot instances or reserved instances to get the most research possible out of your research budget.

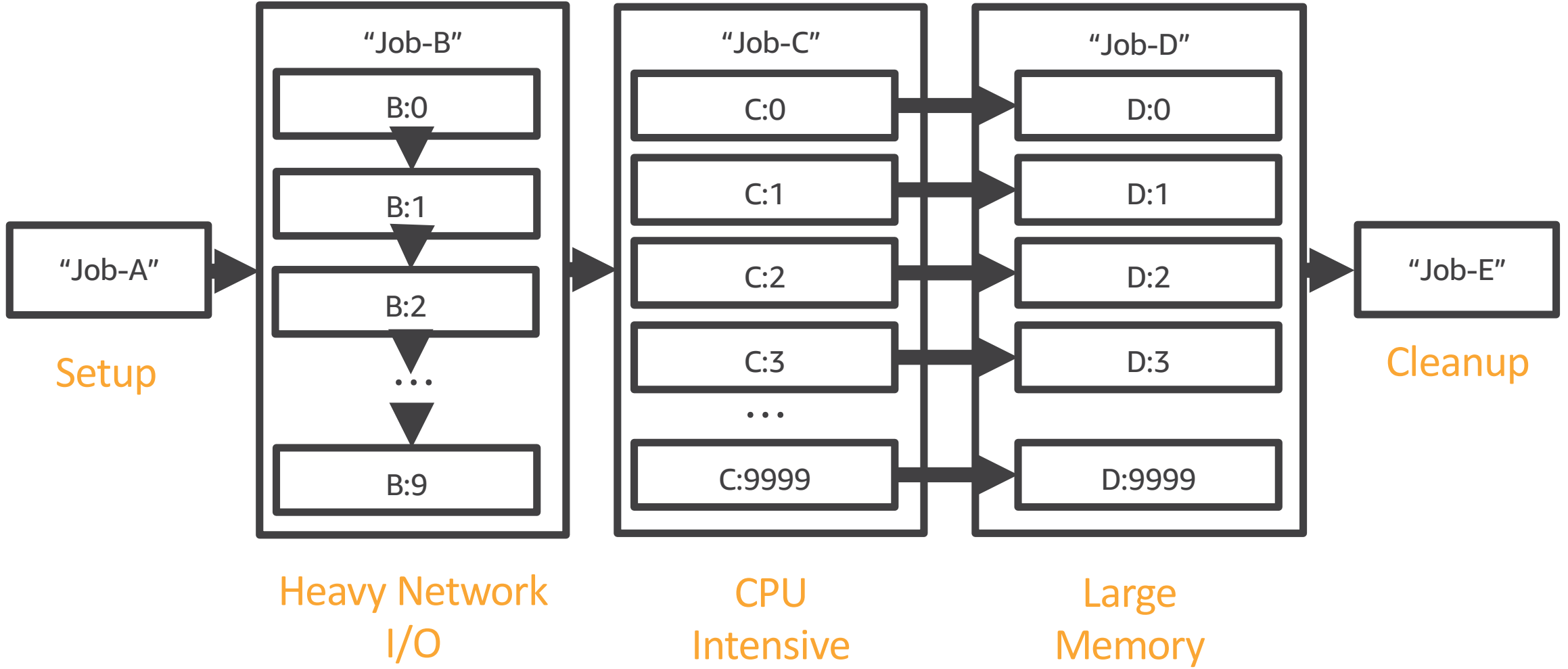
Tutorial using AWS Batch for DNA sequencing: <https://github.com/awslabs/aws-batch-genomics>

Using Containers

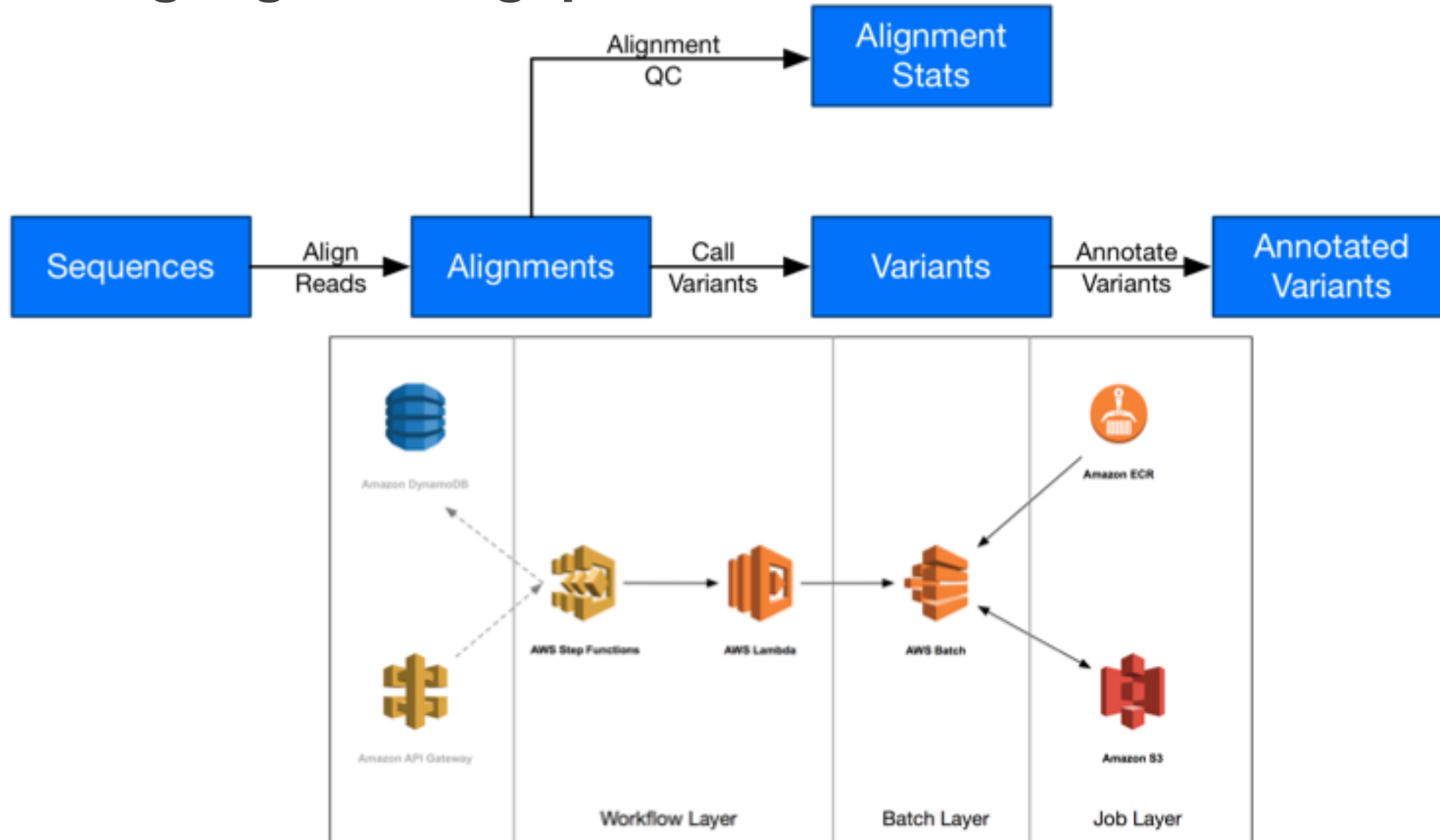


DNA Sequencing





Building High-Throughput Genomics Batch Workflows on AWS



5

Serverless Computing

Serverless Computing: AWS Lambda

AWS Lambda is a service which allows for **software functions** in a variety of languages to be deployed into the cloud natively, and to be **triggered directly or driven by events** in the cloud. The infrastructure (hardware, operating system and software environment) for Lambda is **managed** by AWS and **scales rapidly**.



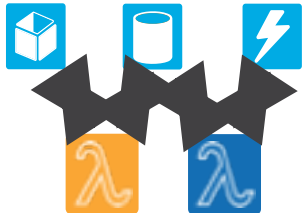
Bring your own code

- Node.JS, Java, Python
- Java = Any JVM based language such as Scala, Clojure, etc.
- Bring your own libraries



Simple resource model

- Select memory from 128MB to 1.5GB in 64MB steps
- CPU & Network allocated proportionately to RAM
- Reports actual usage



Flexible invocation paths



Fine grained permissions

Two examples of HPC on Lambda

CSIRO have built quickly scaling genomics analysis on AWS Lambda

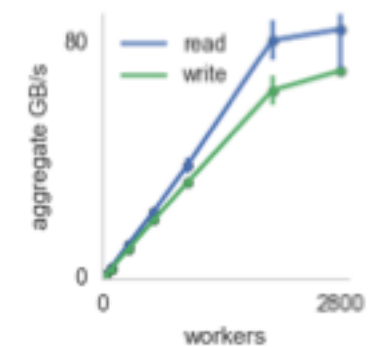
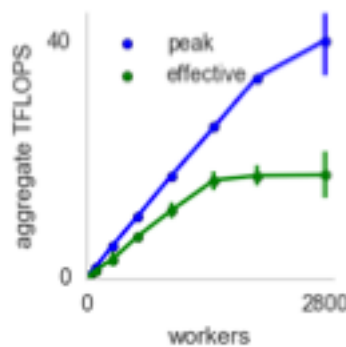


PyWren.io



```
def my_function(b):  
    x = np.random.normal(0, b, 1024)  
    A = np.random.normal(0, b, (1024, 1024))  
    return np.dot(A, x)  
  
pwex = pywren.default_executor()  
res = pwex.map(my_function, np.linspace(0.1, 100, 1000))
```

PyWren lets you run your existing python code at massive scale via AWS Lambda



Pywren: Lambda in the context of Grid Computing

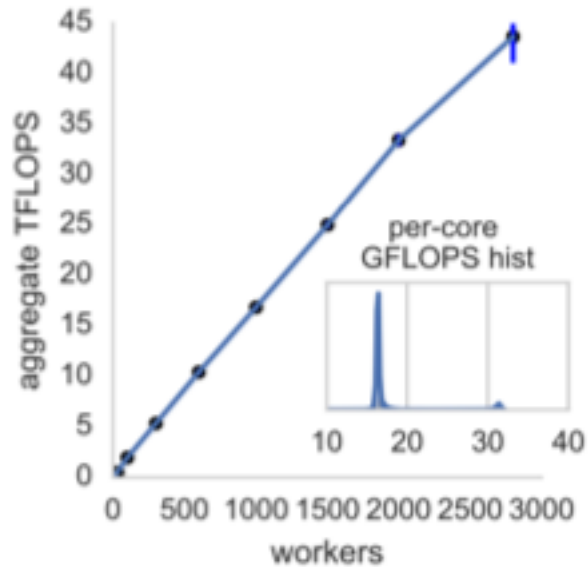


Figure 2: Running a matrix multiplication benchmark inside each worker, we see a linear scalability of FLOPs across 3000 workers.

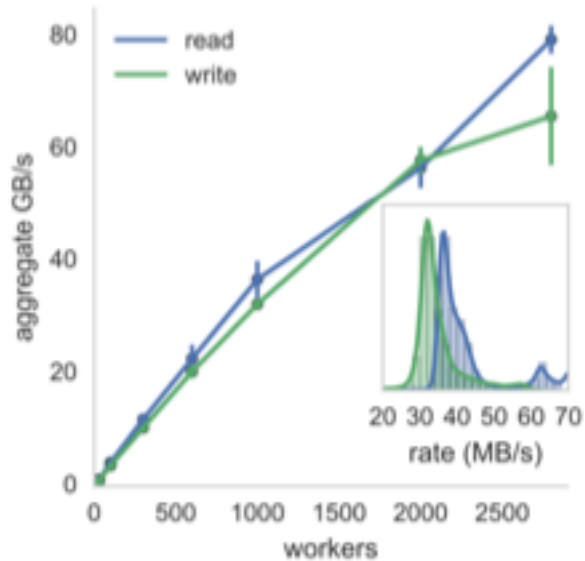


Figure 3: Remote storage on S3 linearly scales with each worker getting around 30 MB/s bandwidth (inset histogram).

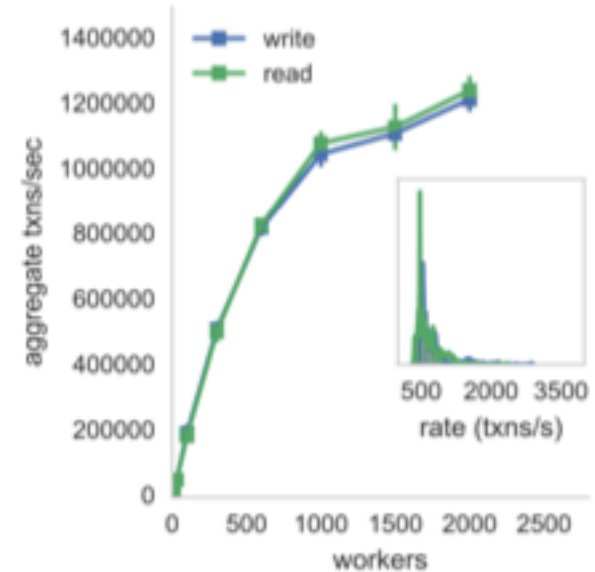


Figure 4: Remote key-value operations to Redis scales up to 1000 workers. Each worker gets around 700 synchronous transactions/sec.

Source: "Occupy the Cloud: Distributed Computing for the 99%"

<https://arxiv.org/pdf/1702.04024.pdf>



Pywren democratizes parallel scaling capabilities that used to be the sole preserve of large super-computing centers. Tutorial: <http://pywren.io/pages/gettingstarted.html> then <https://github.com/pywren/examples/>



CSIRO – Cloud-based CRISPR prediction

- CSIRO used **AWS Lambda functions to completely re-engineer a cluster HPC workload** to identify optimal gene editing sites for personalized treatment.
- “GTScan-2” job runtime varies from 1 second to 5 minutes, because the complexity of the targeted gene can vary dramatically.
- Rapid turn-around times are needed for real-time analysis.
- Server-based solutions can’t be provisioned efficiently to handle the variability and quick turn-around – either you have lots of servers sitting idle, or you have to wait minutes for new servers to spin up.
- Deployed using AWS Lambda, the GTScan-2 runtime is stable at a few minutes **per complete job**, no matter how many jobs are sent to it.
- Re-casting of the code took **only a few weeks**



ten seconds



Ranked choices

CSIRO – CRISPR search with AWS Lambda

GT-Scan2.0 is implemented as a microservices architecture using AWS Lambda

Serverless:

- Does not require users to have high-compute power

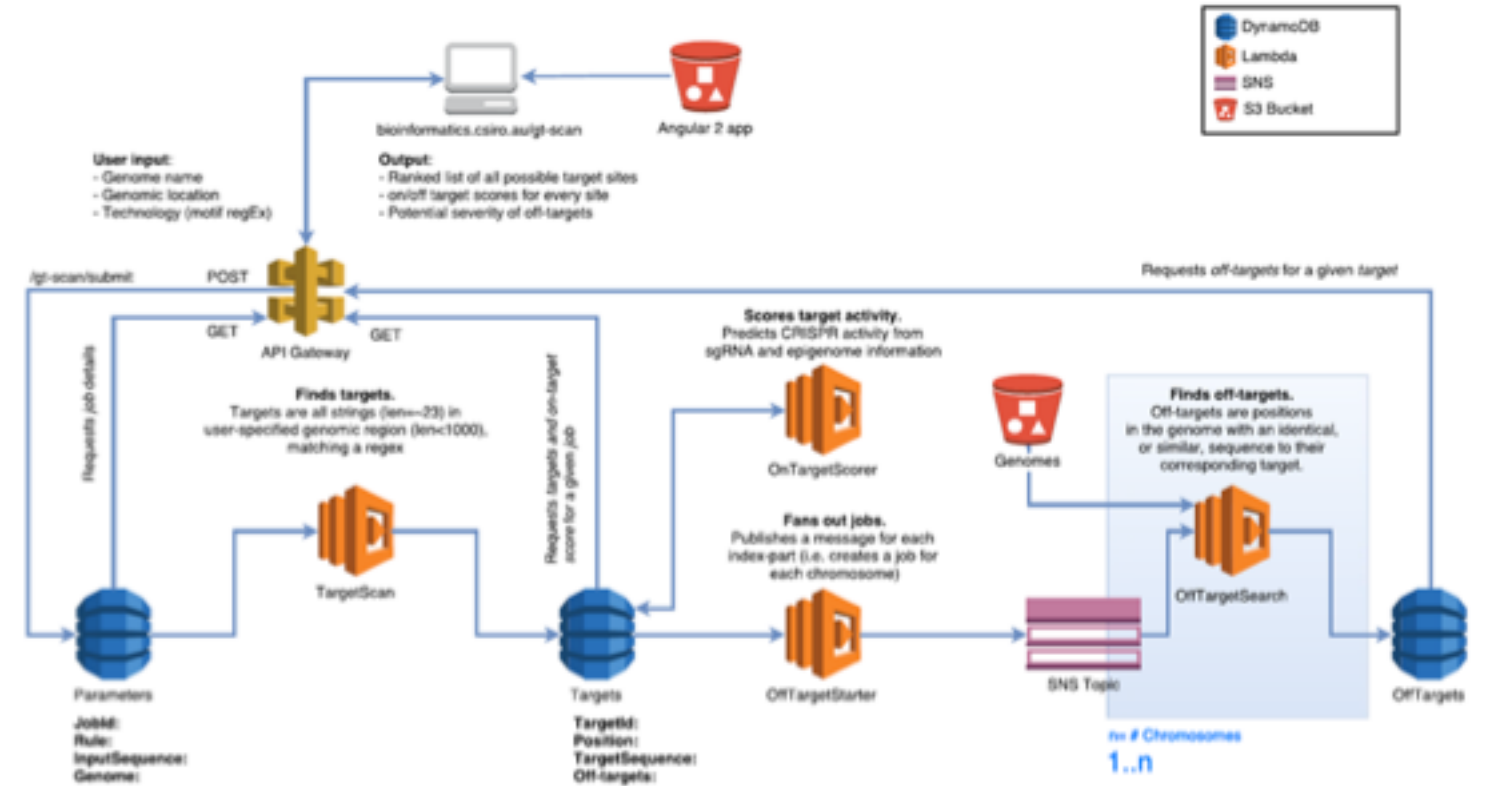
Scalable:

- Can be easily scaled to whole genome analysis

Also implement as a “stand-alone”

- Can be run on local servers
- Can incorporate your own CHIP-seq data rather than public data

GT-Scan2 Microservice-based target-finder for genome editing technologies



6

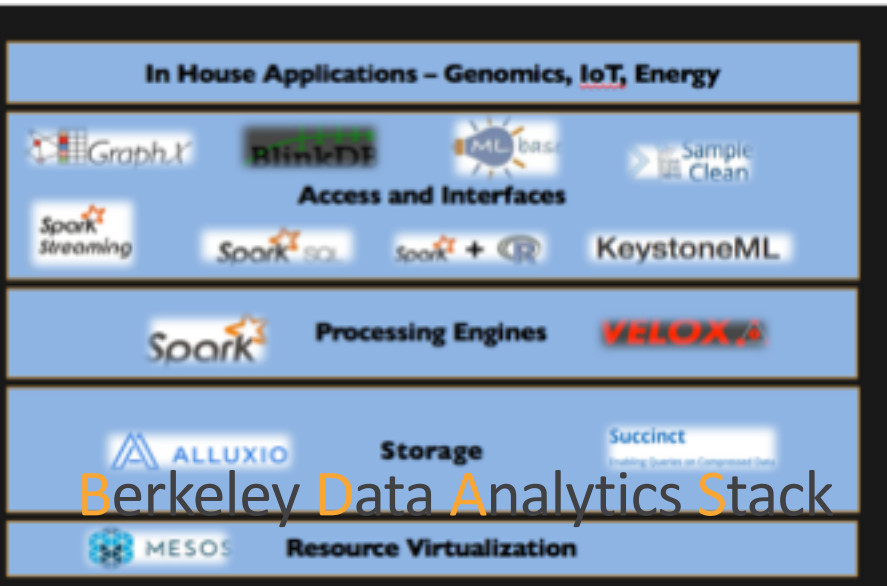
Machine Learning and Amazon SageMaker

RISELab (Real-time Intelligent Secure Execution)

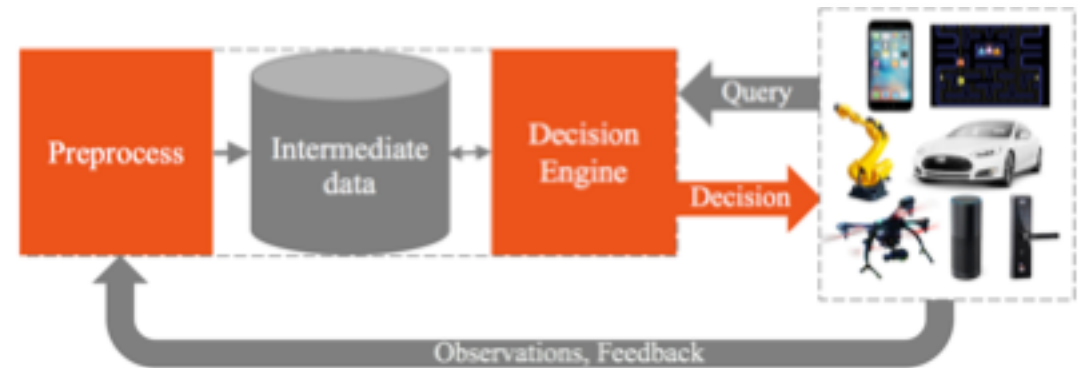
Collaborative 5-year effort between UC **Berkeley**, National Science Foundation, and industry partners. (2017-2021) – AWS is founding partner. <https://riselab.cs.berkeley.edu>

- Students and researchers at RISELab use AWS to **rapidly prototype and develop new systems at a scale and with a speed not possible before.**
- Resulted in Apache Spark, developed on AWS, and integrated with AWS core services.

GOAL: Develop **open source** platforms, tools, and algorithms for intelligent real-time decisions on live-data



From **live data** to **real-time decisions**



Deep Learning using clusters to improve accuracy

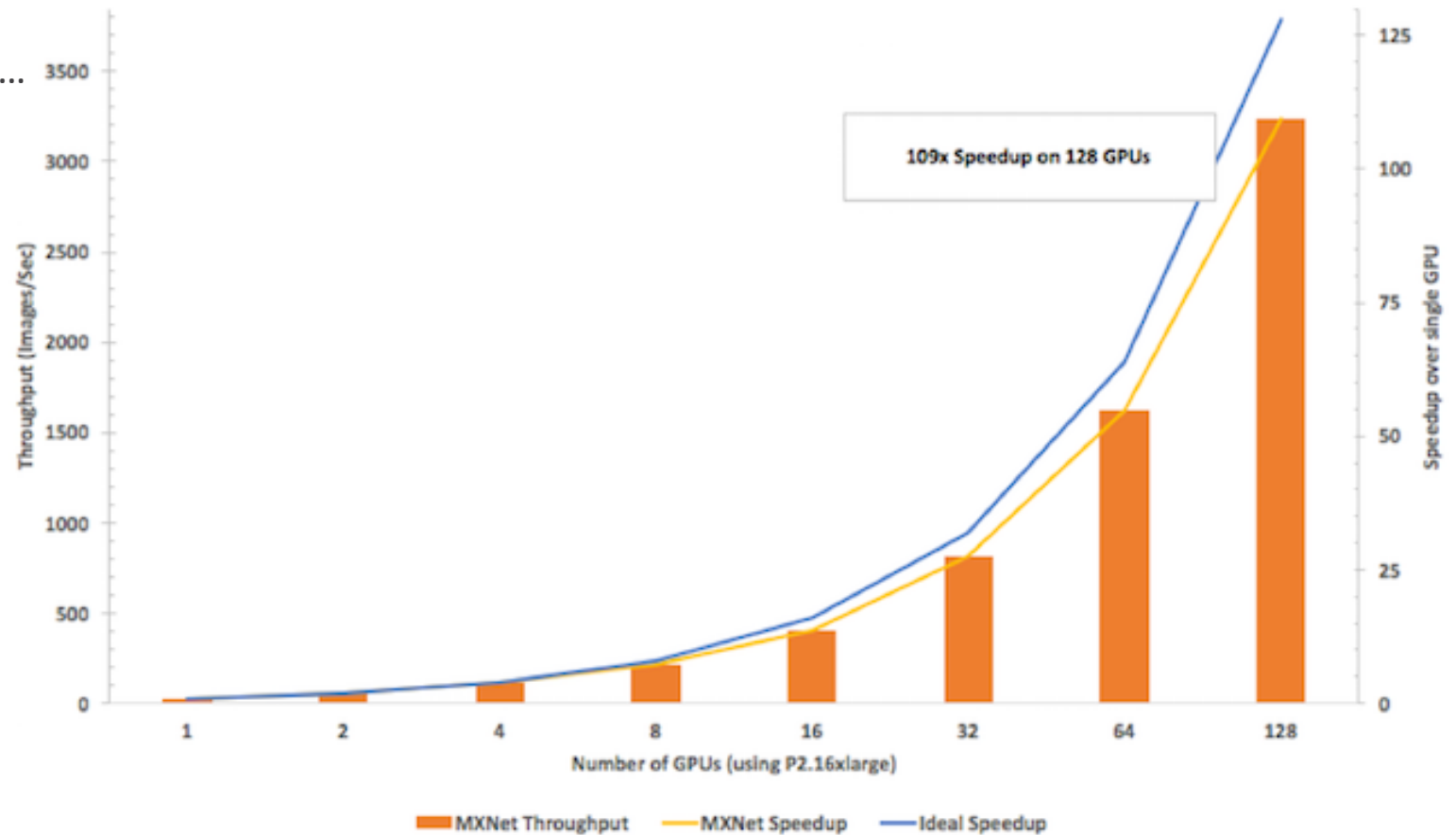
You can train a ML model on your laptop but ...

In order to train a very ACCURATE model that can make real-world predictions that is publishable and cutting-edge ... You will need a much LARGER training dataset and training job that you can ONLY run on a cluster, possibly using GPU servers.

A job for EXPERTS?

⇒ AMAZON SAGEMAKER

<go to SageMaker deck>



7

Research Cloud Program and **Getting Started**

AWS Research Cloud Program



Science first, not servers.

Researchers are not professional IT people (nor do they wish to be).



Simple and easily explained

procedures to get set up with cloud access.



Budget management tools to ensure that over-spends do not happen.



Best practices to ensure both data and research budgets are safe and privacy is protected.



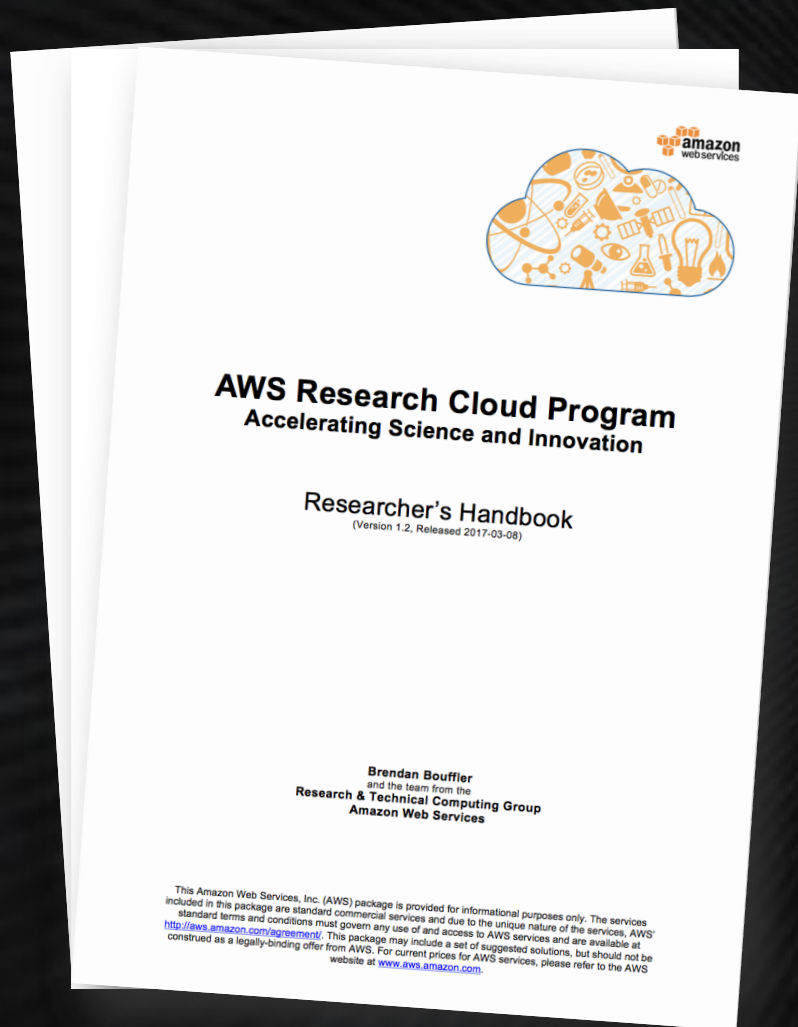
Fast track to invoice-backed billing & Egress Waiver.



Large catalog of scientific solutions from partners, including instant clusters from AWS Marketplace.

AWS Researcher's Handbook

The 150-page “**missing manual**” for science in the cloud.



Written by Amazon's Research Computing community **for scientists**.

- **Explains** foundational concepts about how AWS can accelerate time-to-science in the cloud.
- **Step-by-step best practices** for securing your environment to ensure your research data is safe and your privacy is protected.
- **Tools for budget management** that will help you control your spending and limit costs (and preventing any over-runs).
- **Catalogue of scientific solutions** from partners chosen for their outstanding work with scientists.

aws.amazon.com/rcp

AWS Researcher's Handbook

Contents

| | |
|--|-----------|
| 1 INTRODUCTION TO THE AWS RESEARCHER'S HANDBOOK | 1 |
| 1.1 WHAT IS CLOUD COMPUTING? | 2 |
| 1.2 WHAT IS THE AWS CLOUD?..... | 3 |
| 1.3 WHY AWS FOR SCIENCE AND RESEARCH? | 6 |
| 1.4 AWS COMPUTE AND STORAGE | 11 |
| 1.5 AWS MARKETPLACE | 13 |
| 1.6 SCIENCE AS A SERVICE..... | 14 |
| 1.7 SECURITY | 15 |
| 1.8 DATA PRIVACY | 16 |
| 2 CONFIGURING YOUR AWS ACCOUNT | 18 |
| 2.1 CREATING AN AWS ACCOUNT AND REGISTERING FOR RCP BENEFITS | 18 |
| 2.2 GETTING A NEW AWS ACCOUNT DIRECT WITH AWS..... | 21 |
| 2.3 THE ROOT ACCOUNT | 23 |
| 2.4 CREATING IAM LOGINS..... | 24 |
| 2.5 AWS ORGANIZATIONS | 31 |
| 2.6 SETTING UP SSH ACCESS KEYS | 31 |
| 3 BUDGETING | 36 |
| 3.1 SETTING BUDGETS IN YOUR AWS ACCOUNT | 36 |
| 3.2 OPTIONAL: THE BUDGET SAFETY SWITCH..... | 41 |
| 3.3 AMAZON EC2 LAUNCH LIMITS | 51 |
| 3.4 NEXT STEPS..... | 52 |
| 4 WORKING WITH DATA | 53 |
| 4.1 RANGE OF STORAGE TYPES | 53 |
| 4.2 STORING YOUR RESEARCH DATA | 54 |
| 4.3 SHARING THE DATA IN YOUR S3 BUCKET. | 58 |
| 4.4 GLOBAL DATA EGRESS WAIVER FOR RESEARCH | 59 |
| 4.5 VERY LARGE DATA TRANSFERS TO S3 | 60 |
| 4.6 DATA STREAM INGESTION | 61 |
| 4.7 OPEN DATA MEANS MORE SCIENTIFIC IMPACT..... | 61 |
| 5 WORKING WITH SENSITIVE AND CONTROLLED-ACCESS DATA | 65 |
| 5.1 THE SHARED RESPONSIBILITY SECURITY MODEL | 65 |
| 5.2 MEETING SECURITY AND COMPLIANCE | 65 |
| 6 WORKING WITH COMPUTE | 65 |
| 6.1 THE AMAZON ELASTIC COMPUTE CLOUD | 65 |
| 6.2 AMAZON EC2 COMPUTE INSTANCE | 65 |
| 6.3 HOW ARE EC2 COMPUTE INSTANCES USED | 65 |
| 6.4 OTHER COMPUTE SERVICES..... | 65 |
| 6.5 DATABASE SERVICES | 65 |
| 6.6 TUTORIALS | 65 |
| 7 HPC AND CLUSTERS | 81 |
| 7.1 EASY-LAUNCH TEMPLATE-BASED HPC CLUSTERS | 81 |
| 7.2 MARKETPLACE HPC SOLUTIONS..... | 86 |
| 7.3 PARTNER AND SaaS HPC SOLUTIONS | 87 |
| 7.4 CONTAINERS, MICROSERVICES, AND AWS BATCH | 87 |
| 7.5 SERVERLESS COMPUTE FUNCTIONS: AWS LAMBDA..... | 88 |
| 7.6 ELASTIC MAPREDUCE | 88 |

| | |
|---|------------|
| 7.7 TUTORIALS | 89 |
| 8 MACHINE LEARNING AND OTHER ADVANCED SERVICES | 90 |
| 8.1 MACHINE LEARNING AND PREDICTIVE ANALYTICS | 90 |
| 8.2 AMAZON MACHINE LEARNING (AML) | 91 |
| 8.3 DEEP LEARNING ON AWS | 91 |
| 8.4 ARTIFICIAL INTELLIGENCE AS A SERVICE (AMAZON REKOGNITION, AMAZON LEX AND AMAZON POLLY)..... | 92 |
| 8.5 AMAZON ATHENA..... | 93 |
| 8.6 AWS INTERNET OF THINGS (IoT)..... | 94 |
| 8.7 AMAZON WORKSPACES | 95 |
| 9 JUPYTER AND ZEPPELIN NOTEBOOKS ON AWS | 96 |
| 9.1 JUPYTER ON AWS..... | 96 |
| 9.2 JUPYTERHUB AND ZEPPELIN WITH AMAZON EMR | 96 |
| 9.3 TRAIN A MACHINE LEARNING MODEL ON AWS THROUGH A JUPYTER NOTEBOOK | 97 |
| 10 LEARNING MORE ABOUT AWS | 108 |
| 10.1 HANDS-ON AND FACE-TO-FACE..... | 108 |
| 10.2 ONLINE | 109 |
| 10.3 AWS GLOBAL SUMMIT SERIES..... | 110 |
| 10.4 BLOGS AND PAPERS ON RESEARCH DONE IN THE AWS CLOUD | 110 |
| 11 FINDING AND BUILDING SOLUTIONS | 115 |
| 11.1 BUILD IT YOURSELF | 115 |
| 11.2 THE CLOUD CREDITS FOR RESEARCH PROGRAM | 116 |
| 11.3 THIRD PARTY SOLUTIONS | 116 |
| 12 APN TECHNOLOGY PARTNERS LISTING | 118 |
| 12.1 AEWACS B.V. | 119 |
| 12.2 ACECLOUD, BY ACELLERA | 122 |
| 12.3 ALCES FLIGHT | 125 |
| 12.4 AWS DEEP LEARNING | 130 |
| 12.5 BEEGFS FROM FRAUNHOFER ITWM | 132 |
| 12.6 CFD DIRECT LIMITED | 135 |
| 12.7 DNANEXUS | 141 |
| 12.8 FIGSHARE | 144 |
| | 149 |
| | 152 |
| | 154 |
| | 157 |
| | 162 |
| | 167 |
| | 171 |
| 13 APN CONSULTING PARTNERS LISTING | 174 |
| 13.1 ACELLERA LTD | 175 |
| 13.2 ALCES SOFTWARE LTD. | 178 |
| 13.3 ARCUS GLOBAL | 182 |
| 13.4 INQOO B.V. | 184 |
| 13.5 PIRONET/ CANCOM | 188 |
| 13.6 STERLING GEO..... | 190 |
| 13.7 THE SERVER LABS LTD..... | 192 |

The topics that matter to researchers
One-stop tutorial: no more getting lost on the AWS website

AWS Researcher's Handbook



definitely saving money by actively monitoring jobs to catch problems early and reduce rework," explains Andrew McComas, Engineering Manager at TLG. "We can also use it to reduce unnecessary cost in larger jobs that may otherwise run longer than required."

- <https://www.top500.org/news/sponsored/why-customers-are-moving-high-performance-computing-workloads-to-amazon-web-services-1/>

1.3.7 Medical Imaging: National Database for Autism Research (NDAR)

Collaborative Big Data research. The National Institute of Mental Health Data Archive (NDA) makes research data available for reuse. Data collected across projects can be aggregated and made available, including clinical data, and the results of imaging, genomic, and other experimental data collected from the same participants. In this way, separate experiments on genotypes and brain volumes can inform the research community on the over one hundred thousand subjects now in the NDA.

The NDA holds rich datasets (fMRI, brain imaging) in object-based storage (Amazon S3). It supports the deployment of packages (created through the NDA Query tools) to an Amazon Web Service Oracle database. The NDA envisions real-time computation against rich datasets that can be initiated without the need to download full packages. Furthermore, a new category of data structure has been created called "evaluated data." This allows researchers using NDA cloud capabilities and computational pipelines to write any analyzed data directly back to the mNDAR database. Databases can also be populated with your own raw or evaluated data and uploaded directly back into the NDA for a streamlined data submission directly from a hosted database.

- https://ndar.nih.gov/cloud_overview.html

1.3.8 Genomics: GT-Scan2 from CSIRO in Australia

New HPC paradigms. In 2016, the Commonwealth Scientific and Industrial Research Organisation (CSIRO - a federal government agency for scientific research in Australia) used AWS Lambda functions to completely re-engineer their HPC workflow for GT-Scan2 that had been developed in a traditional cluster environment. The workflow for CRISPR gene editing sites through simulation. The workflow is now running on AWS Lambda, and other "serverless" functions in AWS, allowing developers at CSIRO.

AWS Lambda (see chapter 7.5) is a service that allows you to run code in the cloud of languages into the cloud natively, triggered directly from your code. The infrastructure (hardware, operating system) is managed by AWS and scales rapidly. This is ideal for personalized treatment, because the complexity of the analysis grows dramatically. A typical GTScan-2 job takes less than a minute, but the variation between jobs ranges from 1 second to 5 minutes. This fast fluctuation in load over minutes rather than hours, and the need for rapid turn-around times meant that large amounts of server hardware could end up idle simply waiting for a job to arrive. A naive EC2-based solution would also be limited, since new instances - which may take minutes to deploy - would come online too slowly to keep the runtime stable. But with AWS Lambda, the



these models on GPU instances in the cloud. This is just the beginning! You can run Jupyter notebooks on any instance types available in EC2. You can use Jupyter to run big data analytics using Amazon EMR (a managed Hadoop platform on AWS), and you can even control HPC clusters from the comfort of your Jupyter Notebook as well.

We'd also love to hear about your use case and what you're looking to do with AWS.

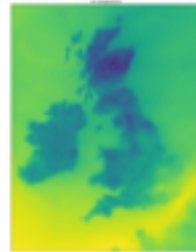
Don't forget to turn your AWS resources off when you are finished!

9.3.4 Further tutorials

Run Jupyter Notebook and JupyterHub with an Amazon EMR cluster:

<https://aws.amazon.com/blogs/big-data/running-jupyter-notebook-and-jupyterhub-on-amazon-emr/>

UK Met Office has made 80TB of MOGREPS dataset with meteorological (weather) data available on S3 through the AWS Open Data program. They've published 2 tutorial notebooks showing you how to pull data from the data set, manipulate it, and visualize it. It uses the Iris python library for much of this. See http://data.informaticslab.co.uk/mogreps_data_basics.html and http://data.informaticslab.co.uk/mogreps_data_intermediate.html



[The Awesome Data Science](#) YouTube video tutorial is an excellent series of tutorials about using Jupyter for the basics of Data Science in Python. All the tutorials mentioned in this video tutorial series are available in their [GitHub repository](#) as well.

Science use cases
Step-by-step tutorials and links to further tutorials
Written in the right tone for researchers

AWS Researcher's Handbook



| | | |
|---------------------------|--|--|
| Technology Partner | Product or Company Alces Flight Ltd. | Home Page http://www.alces-flight.com |
| | Vendor Country of Origin United Kingdom | Delivery method AWS Marketplace |
| Domains | Benchmarks, Biochemistry, Bioinformatics, Bio-physics, Chemistry, Complex, Databases, Electronics, Engineering, Geography, Graphics and Imaging, Languages, Libraries, Mathematics, Medicine, MPIs, Physics, Statistics, Tools, Visualization (more information available at http://docs.alces-flight.com/) | |

Regional availability – All AWS Regions are covered.

| | | | | | | | | | | | | | | | | |
|---------|-----------|--------|-------------|--------|-------------|---------------|--------|------|-----------|-----------|-------|--------|-------|--------|---------|--------------|
| IRE | DE | UK | FR | CA | US | US | US | US | BR | SG | JP | AU | KR | IN | CN | CN |
| Ireland | Frankfurt | London | Paris (IAD) | Mumbai | A. Virginia | A. California | Prague | Ohio | Sao Paulo | Singapore | Tokyo | Sydney | Seoul | Mumbai | Beijing | Beijing (CN) |

12.3 Alces Flight

Alces Flight Compute provides a fully-featured, scalable **High Performance Computing (HPC)** environment for research and scientific computing. Compatible with both on-demand and spot instances, Flight rapidly delivers a whole HPC cluster, ready to go and complete with job scheduler and applications.

Clusters are deployed in a Virtual Private Cloud (VPC) environment for security, with SSH and graphical-desktop connectivity for users. Data management tools for POSIX and S3 object storage are also included to help users transfer files and manage storage resources.



12.3.1 How is it accessed?

Alces Flight Compute is available in a solo user experience in the Marketplace, with multi-user and companion appliances available by contacting your AWS Account Manager or Alces Flight directly. Read on to see which version is right for you:

| Flight Compute feature | Available in AWS Marketplace | | |
|--|------------------------------|---------------------------|--------------------|
| | Solo Community Edition | Solo Professional Edition | Enterprise Edition |
| HPC job scheduler (Open Grid Scheduler/SGE) | ✓ | ✓ | ✓ |
| Alces Golden software application library | ✓ | ✓ | ✓ |
| Interactive graphical desktop sessions | ✓ | ✓ | ✓ |
| Secure VPN access | ✓ | ✓ | ✓ |
| Community support | ✓ | ✓ | ✓ |



Who is it for?

Alces Flight Solo is designed for use by end-users - that's the scientists, researchers, engineers and software developers who actively run compute workloads and process data. Flight provides tools that enable self-service - it's very configurable, and can be expanded by individual users to deliver a scalable platform for computational workloads.

Prerequisites

To get started you need three things:

- Access to some computers
- A client device
- Access to cloud resources (AWS Account)

Check out our full prerequisites list here: <http://docs.alces-flight.com/en/stable/overview/whatist.html#prerequisites>

Where can I get help?

The online documentation (<http://docs.alces-flight.com/>) is designed to walk users

Detailed 2-4p description of Partner solutions: Shows the scope of our ecosystem. Many 3rd parties invest in HPC on AWS.

Flight clusters, report any bugs with the software, and share knowledge to help everyone work more effectively. There is no payment for using this service except for the general requirement to be nice to each other.

Creating your Cluster

The simplest method of launching a cluster in the AWS Marketplace is by following these steps:

1. Create (if you haven't already) and sign-in to your AWS account and navigate to the [AWS Marketplace](#).
2. Search for **Alces Flight** in the search box provided to find the Alces Flight Solo Community Edition (or Solo Professional if you wish to use our paid service) and click to select it.
3. Read the product information and click on the **Continue** button to view details on how to launch.
4. After clicking the **Continue** button from the main product page select the **Custom Launch** tab in your browser.
5. Scroll down the page and select your local AWS region in the **Select a Region** section.

At the INSTITUTIONAL level

Build institutional cloud competency

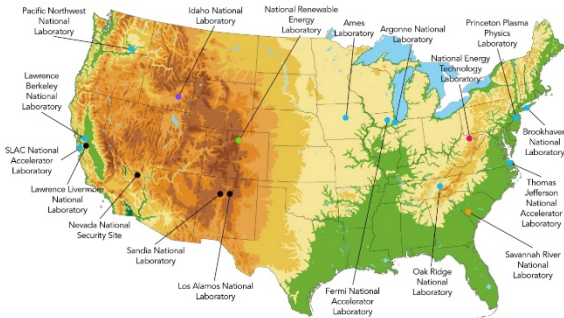
Support for domain specialists who want to use cloud

Platforms to help end users consume cloud services (credentials, security, compliance, billing, ...)



Pacific Northwest National Lab (PNNL): Supporting research

The national laboratory system



- Office of Science Laboratory
- National Nuclear Security Administration Laboratory
- Office of Fossil Energy Laboratory
- Office of Energy Efficiency & Renewable Energy Laboratory
- Office of Nuclear Energy, Science & Technology Laboratory
- Office of Environmental Management Laboratory



PNNL at a glance



- \$920.4 M In R&D expenditures
- 104 U.S. and foreign patents granted
- 1,058 Peer-reviewed publications
- 2 FLC Awards
5 R&D 100 Awards
- 4,400 Scientists, engineers and non-technical staff



SRV 318
AWS re:INVENT
 Research at PNNL: Powered by AWS
 Mike Giardinelli and Ralph Perko
 Pacific Northwest National Laboratory
 November 28, 2017

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

<https://www.youtube.com/watch?v=hcnhdwnSY94>

Enabling research with AWS



Environments



Staff

- Research is the life blood of the organization
- Researchers should not be troubled with environment configurations, optimizations, etc.
- Software engineers provide expertise needed to build applied solutions
- Utilizing AWS has been a turning point.
- AWS has dramatically helped to improve collaboration.
- AWS fits better with our Agile software processes



Moving to the cloud



- Drivers**
 - Lack of resources internally (hardware and people)
 - Customer deliverables and demands / deadlines
- Concerns**
 - Cost
 - Vendor lock-in
- Initial Approach**
 - Fork-lift model
 - Missed out on AWS services
 - Still had operational headaches
- Current Approach**
 - Serverless wherever possible



As a result, researchers can focus on the problem

Our progression to AWS

Emory-specific AWS landing page

EMORY | AWS Products Documentation [user name if known] Support AWS Console VPCP Console Create Account/VPC

Introducing the Emory HIPAA compliant AWS research service

Explore the Emory AWS Research Service

| Products | Security & Compliance | Support | Billing | Professional Services |
|---|--|---|---|---|
| <p>To create a secure platform for research innovation in the Cloud, Emory has implemented the following:</p> <ol style="list-style-type: none">1. Researcher-managed Amazon Web Services (AWS) Accounts with security risk assessments and countermeasures to meet Emory security and compliance requirements2. Centrally-managed Virtual Private Cloud (VPC) structures, security policies, and service control policies with provisioning and administration automation3. Cloud competency center staffed by Emory IT, AWS enterprise support, and preferred cloud consultants | <p>Emory AWS Research Service accounts can be used for both non-sensitive data and sensitive data that is subject to HIPAA and [list other compliance regimes here] regimes. Emory performs a risk assessment of each AWS service and implements controls and countermeasures prescribed by Emory information security and compliance policies.</p> <p>Emory users of the service must agree to use best practices and the terms and conditions defined in the Emory Rules of Behavior for the Emory AWS Research Service. [Emory AWS Research Service should be a link to these rules for Box, presumably we will have one for AWS too]</p> | <p>Frontline support for the Emory AWS Research Service is provided by AWS Enterprise Support. LITS provides support for LITS-delivered solutions for AWS, security and compliance controls, and provisioning. Whenever users working in AWS need support, the best approach is to file an AWS support ticket from the upper right-hand corner of the AWS console. If you know you are working with a solution stack provided by the LITS Helpdesk, you may use the Emory VPCP application or ServiceNow to file a support request. Support links in the VPCP application for managing LITS provisioning and access management submit tickets to the LITS Helpdesk.</p> | <p>Emory AWS Research Service charges are passed through to the customer via and integration with Emory's financial accounting system. When you sign up for an account to provide an account number and users must maintain a valid account number of each of their accounts. The AWS bill view and Cost Explorer is available in each Emory AWS Account to help manage user's AWS spend.</p> | <p>Emory has existing agreements with preferred vendors to provide the Emory community with professional services for AWS. These vendors are vetted by Emory and familiar with Emory's security and compliance controls. They can provide on-shore resources at competitive rates. Request professional services for AWS.</p> |

<https://edscoop.com/emory-university-research-aws-cloud-rich-mendola>

Thank You & Homework

Sign up for the **Researchers' Handbook for AWS** at aws.amazon.com/rcp . **Browse data** at <https://registry.opendata.aws>

Tutorials:

- If you have your own AWS account, use that.
- The Alces Flight demo will run in an Alces account, but you won't have to worry about it.
- You can run the SageMaker demos in this account (today only):

<https://001868661679.signin.aws.amazon.com/console> user: unidata ; p/w: unidata2018

1. **Alces Flight compute cluster - NAMD tutorial:** Launch "Performance Compute (SGE)" cluster at <https://launch.alces-flight.com/default/launch> , wait for e-mail confirmation, then tutorial from <http://docs.alces-flight.com/en/stable/getting-started/environment-usage/using-openfoam-with-alces-flight-compute.html>
2. **WRF4.0 on AWS:** http://www2.mmm.ucar.edu/wrf/OnLineTutorial/wrf_in_cloud_aws_tutorial.php
3. **GEOS-CHEM on AWS:** http://cloud-gc.readthedocs.io/en/latest/chapter02_beginner-tutorial/quick-start.html
4. **Containers + AWS Batch for DNA sequencing:** <https://aws.amazon.com/blogs/compute/building-high-throughput-genomics-batch-workflows-on-aws-introduction-part-1-of-4/>
5. **Containers – WRF Big Weather Web:** www.bigweatherweb.org
6. **Serverless Computing – PyWren:** <http://pywren.io/pages/gettingstarted.html>
then <https://github.com/pywren/examples/>
7. **SageMaker Machine Learning labs:** files from <https://bit.ly/2HhD2SG> ; instructions at <https://github.com/wleepang/sagemaker4research-workshop> ; further labs at <https://developmentseed.org/blog/2018/01/19/sagemaker-label-maker-case/> and <https://aws.amazon.com/blogs/machine-learning/simulate-quantum-systems-on-amazon-sagemaker/>

Alces Launch - tokens

- nice-azure-stallion
- vainly-mysterious-parakeet
- magenta-barbarous-pig
- bitterly-cooing-wolverine
- generously-handsome-thermometer
- triumphantly-old-ring
- repeatedly-please-paint-brush
- painfully-dark-deer
- optimistically-worry-alligator
- loud-stallion-parakeet