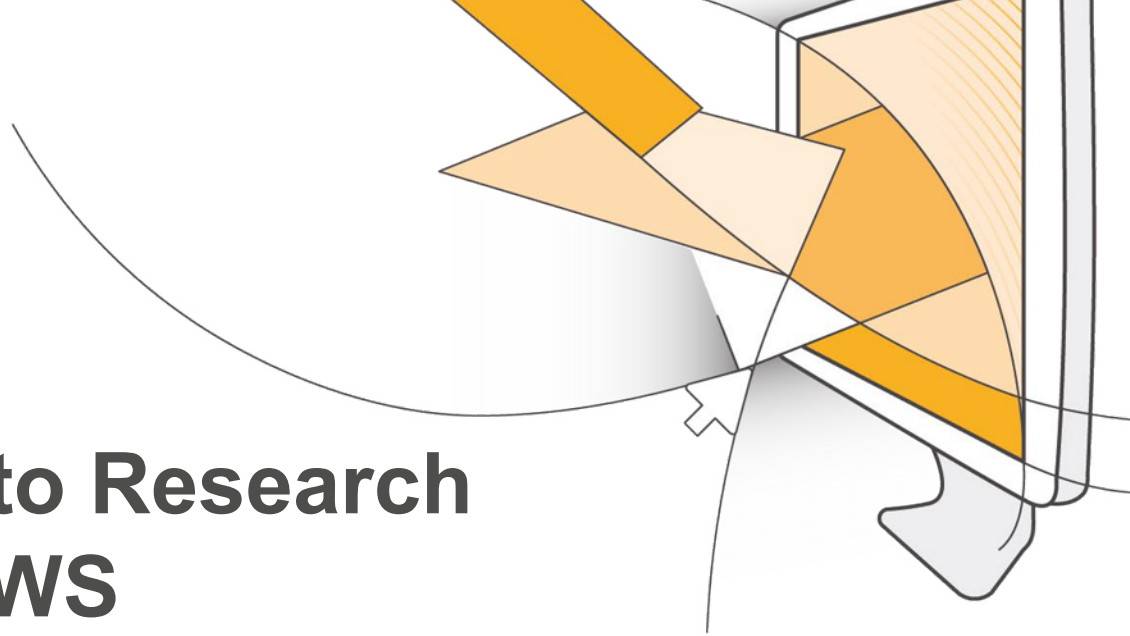


An Introduction to Research Computing on AWS

Adrian White, Research & Technical Computing
Amazon Web Services

November 2017

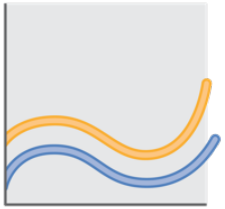
Agenda

10:00	Introduction and welcome	(30 mins)
10:30	AWS Overview	(30 mins)
11:00	Research & Technical Computingon AWS	(1 hour)
	Demo: Jupyter Notebooks	
12:00	**Lunch break**	**(1 hour)**
13:00	High performance computing	(30 mins)
13:30	**Demo: Alces Flight**	
13:50	**Short break**	**(10 mins)**
14:00	**Lab: Introduction to Research Computing**	**(1.5 hours)**
15:30	Wrap-up & Next Steps	(30 mins)
16:00	**Ronin – Unleash your Research**	**(1 hour)**
17:00	Finish	

AWS for Research & Technical Computing

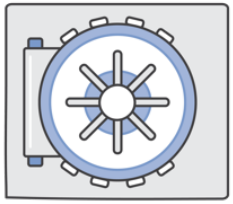
Time to Science
Access research infrastructure in minutes

Low Cost
Pay-as-you-go pricing

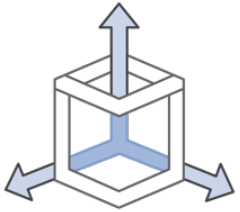


Elastic
Easily add or remove capacity

Globally Accessible
Easily Collaborate with researchers around the world

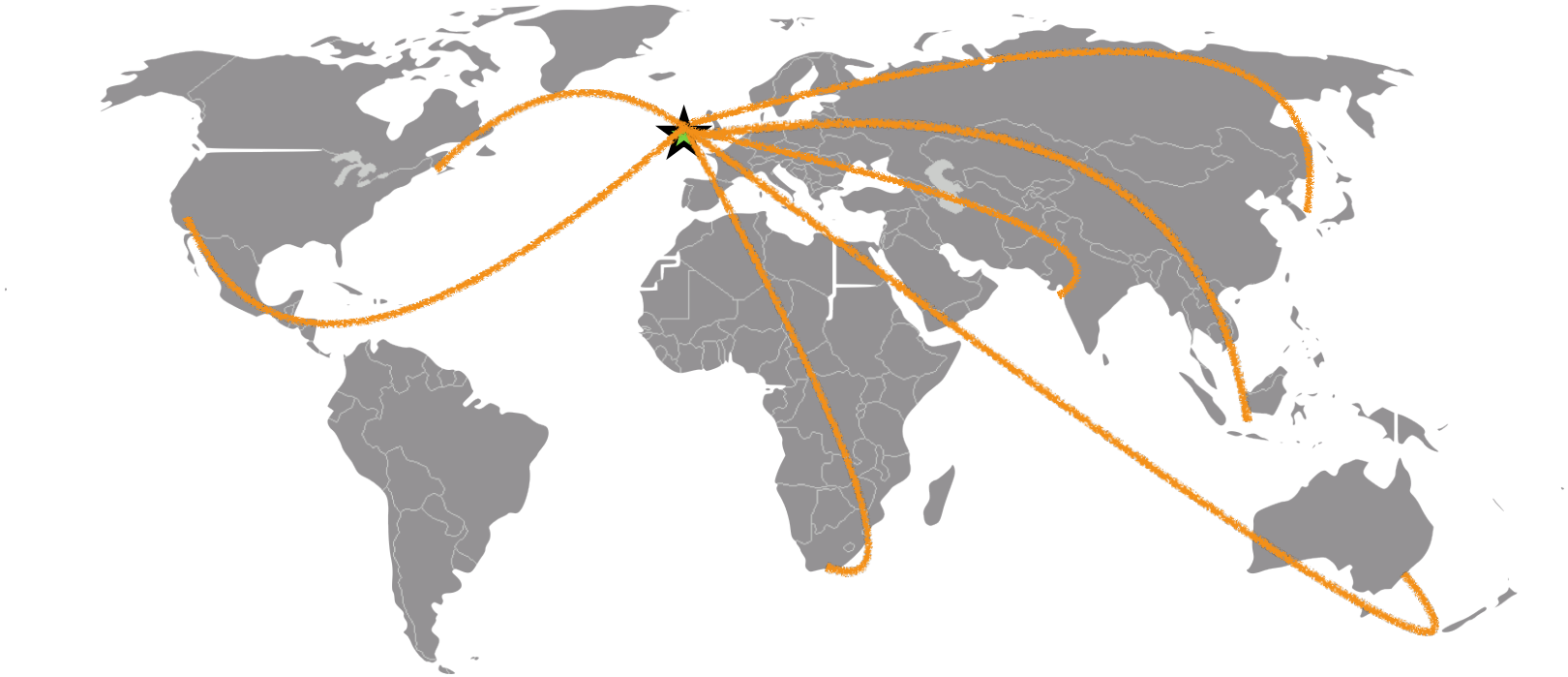


Secure
A collection of tools to protect data and privacy



Scalable
Access to effectively limitless capacity

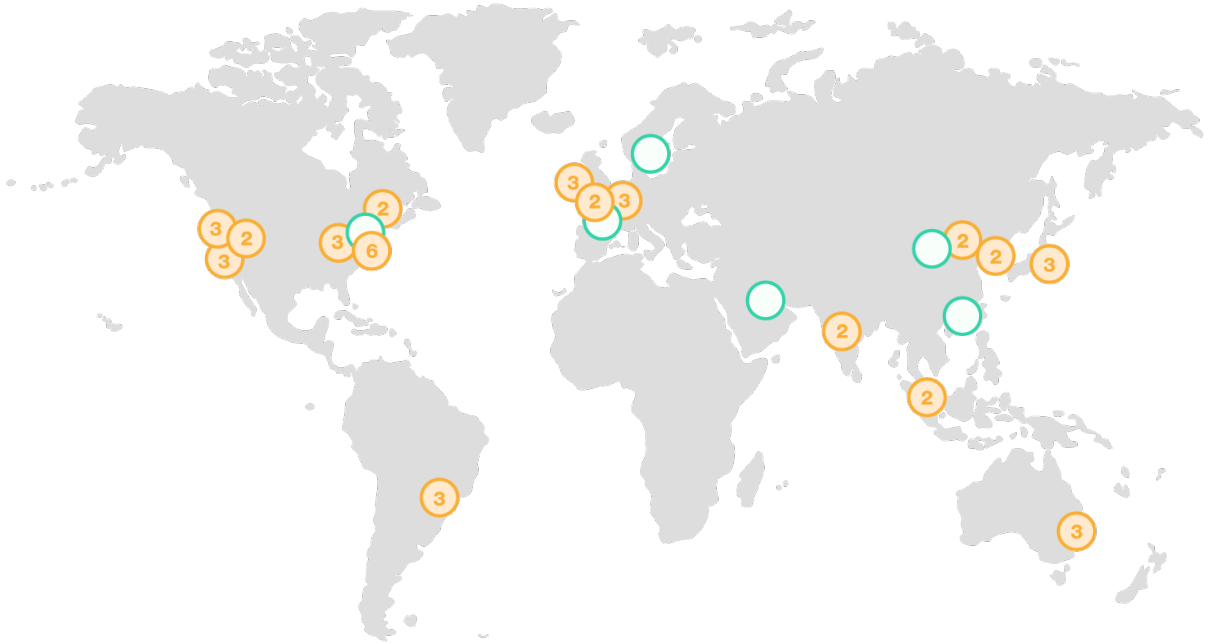
Collaboration is easier in the cloud



More time spent computing the data than moving the data.

Global AWS Regions

Current: 44 Availability Zones within 16 geographic Regions
Announced: 17 more Availability Zones and 6 new regions



Americas
- AWS GovCloud (2)
- **AWS GovCloud East**
- US West
- Oregon (3)
- Northern California (3)
- Northern Virginia (5)
- Ohio (3)
- Montreal (2)
- São Paulo (3)

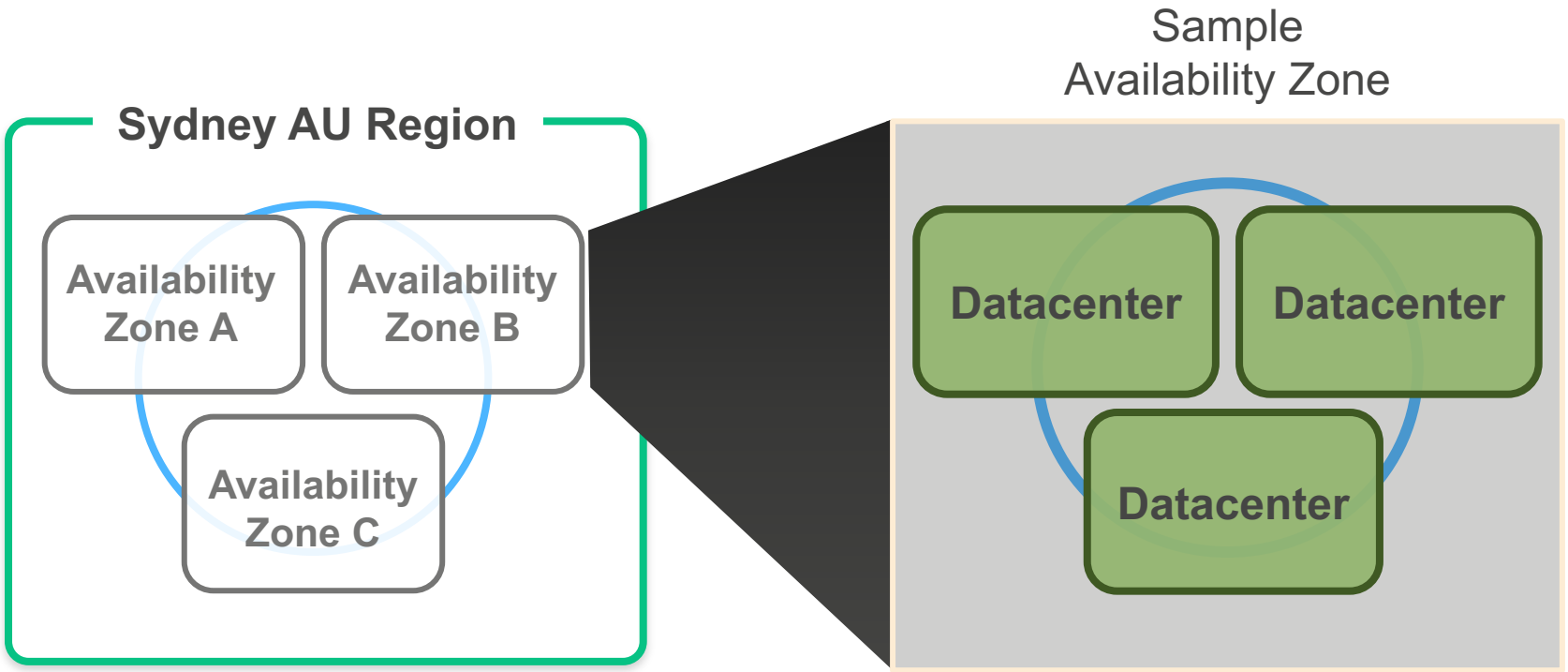
Europe / Middle East
- Ireland (3)
- Frankfurt (3)
- London (2)
- **Paris**
- **Stockholm**
- **Bahrain**

Asia Pacific
- Singapore (2)
- Sydney (3)
- Tokyo (3)
- Seoul (2)
- Mumbai (2)
- Beijing (2)
- **Ningxia**
- **Hong Kong(3)**

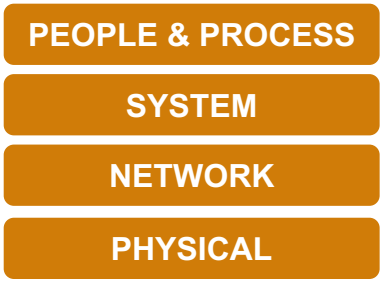
AWS Region = A cluster of Availability Zones
Availability Zone = A cluster of data centers

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

Anatomy of a region

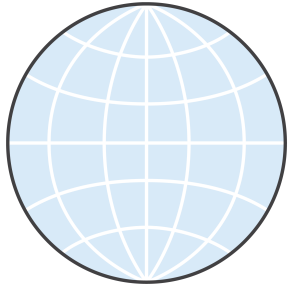


Security is Job Zero



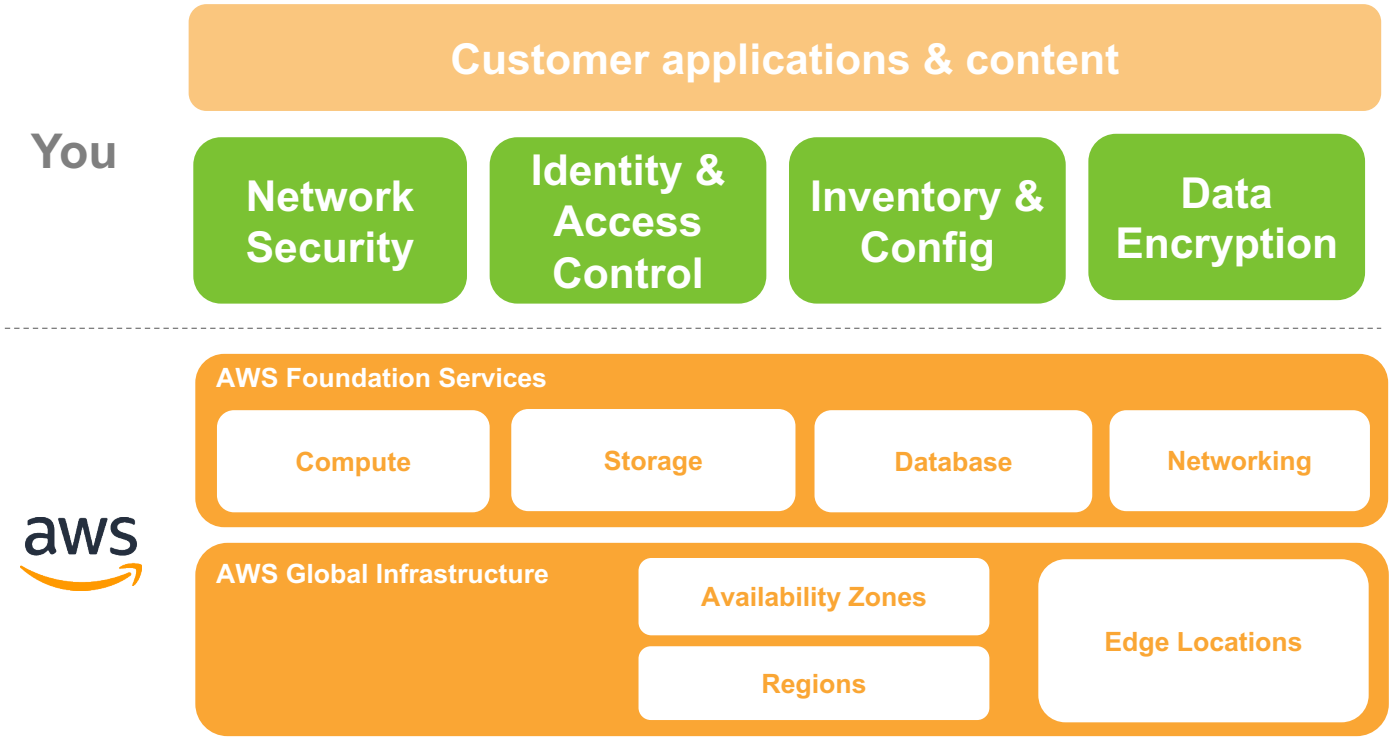
Familiar Security Model

Validated and driven by customers' security experts



Benefits all customers

AWS And You Share Responsibility for Security

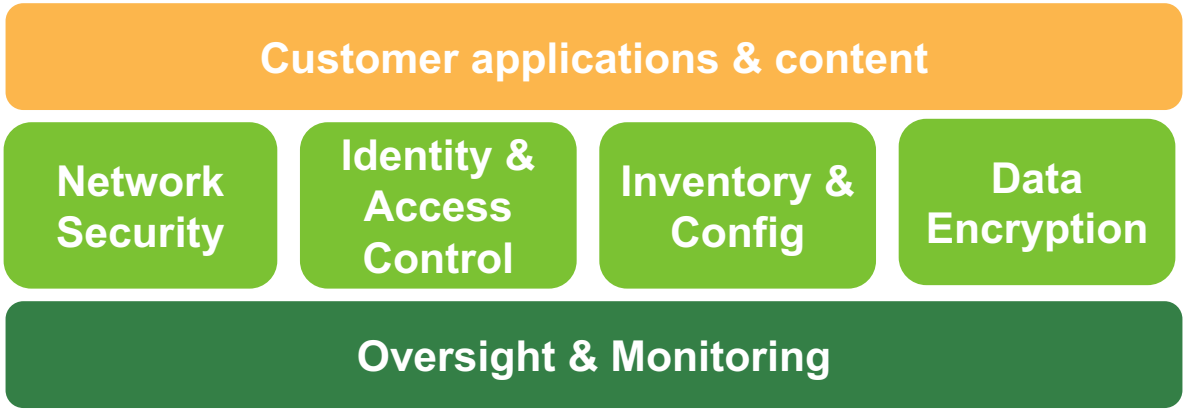


You get to define your controls **IN** the Cloud

AWS takes care of the security **OF** the Cloud

Key AWS Certifications and Assurance Programs

AWS Security Tools & Features



AWS and its partners offer over 700 security services, tools and features

Mirror the familiar controls you deploy within your on-prem environments

Encrypt your sensitive information

Native encryption across services for free

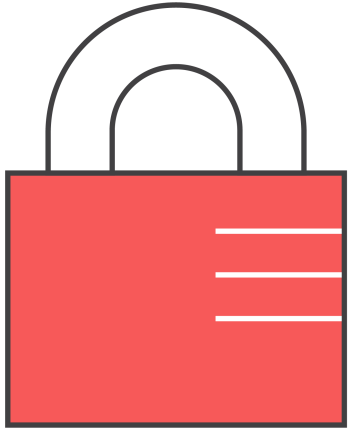
- S3, EBS, RDS, RedShift
- End to end SSL/TLS

Scalable Key Management

- AWS Key Management Services provides scalable, low cost key management
- CloudHSM provides hardware-based, high assurance key generation, storage and management

Third Party Encryption options

- Trend Micro, SafeNet, Vormetric, Hytrust, Sophos etc.

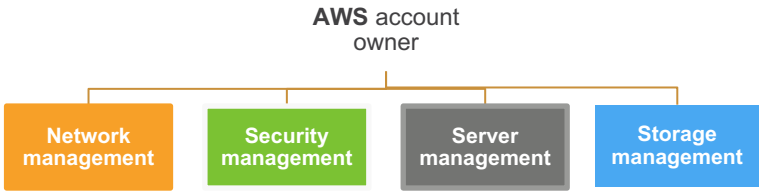


Control access and segregate duties everywhere

You get to control **who** can do **what** in your AWS environment **when** and from **where**

Fine-grained control of your AWS cloud with **multi-factor authentication**

Integrate with your existing Active directory using federation and single sign-on

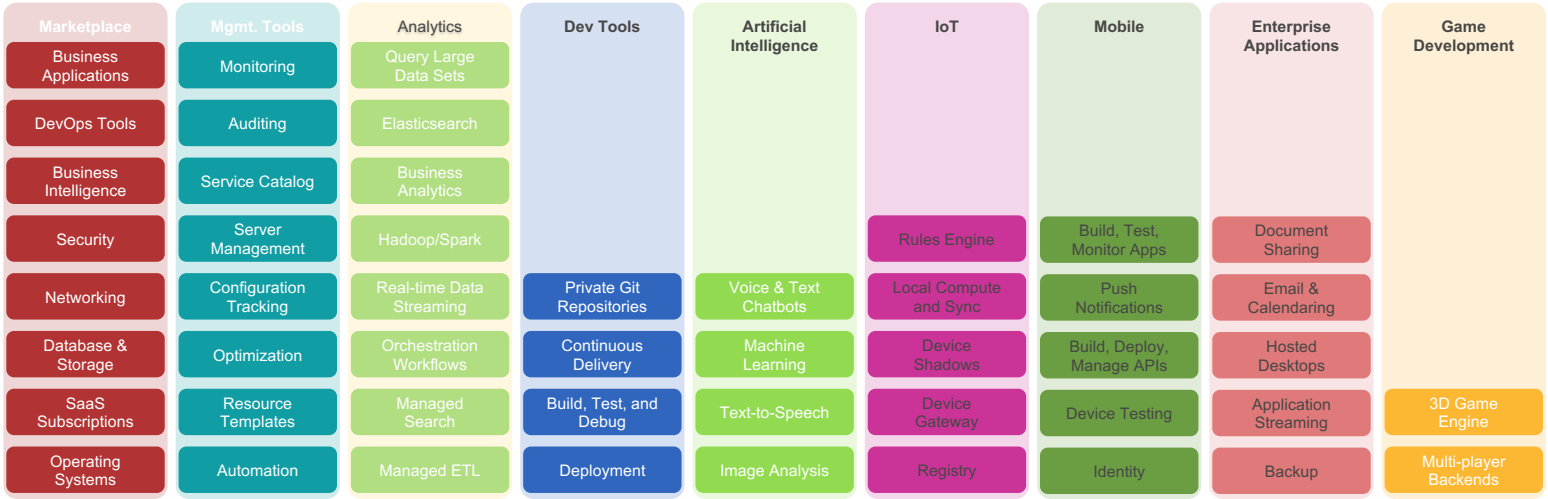


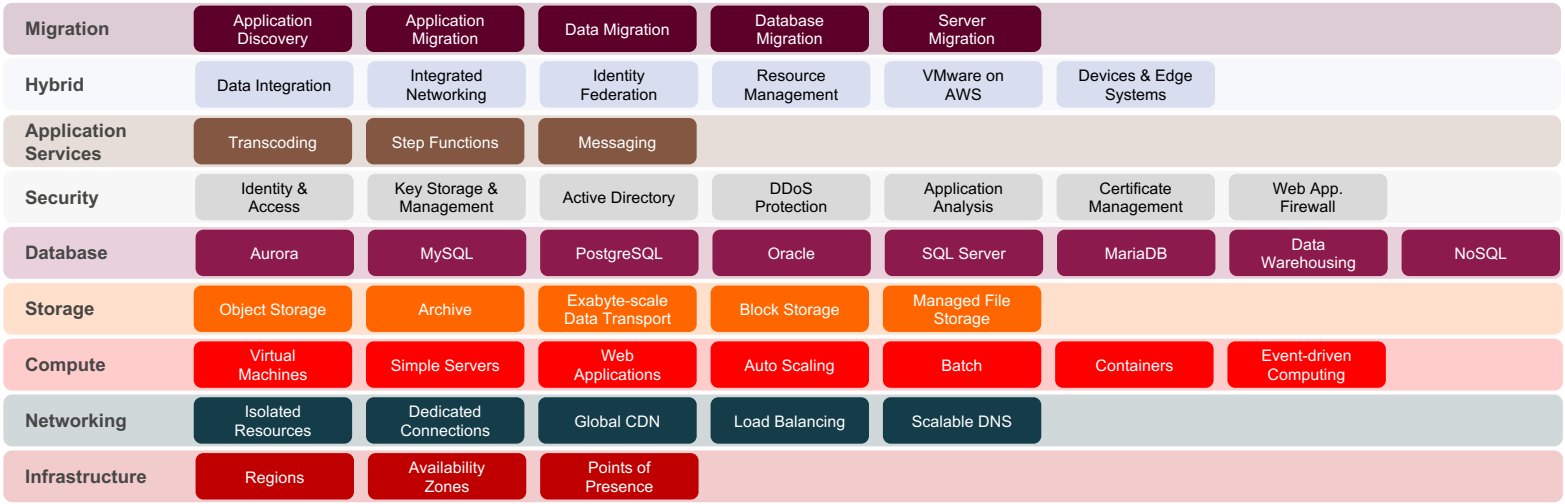
A tool for almost every job

A tool for almost every job

because you'll want tools you don't know about yet

The AWS Platform



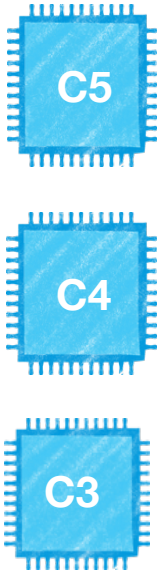


AWS Instance Types

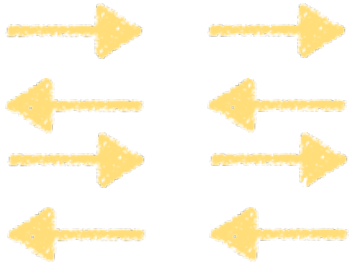
Broad Set of Compute Instance Types
for HPC and Deep Learning

General purpose

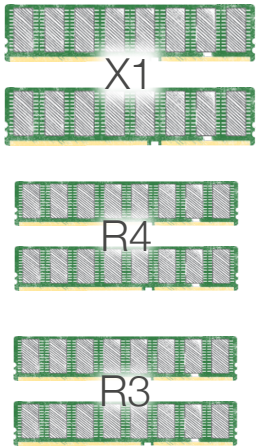
Compute optimized



Storage and I/O optimized



Memory optimized



GPU or FPGA enabled

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 c4.8xlarge 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

M4.16XL

Intel Xeon 2686v4, custom built for AWS.

Intel Broadwell, 16 FLOPS/tick

2.2 GHz, turbo to 3.60 GHz

Performance	
# of Cores	22
# of Threads	44
Processor Base Frequency	2.20 GHz
Max Turbo Frequency	3.60 GHz
Cache	55 MB SmartCache
Bus Speed	9.6 GT/s QPI
# of QPI Links	2
TDP	145 W
VID Voltage Range	0

Expanding the M4 Instance Type – New M4.16xlarge

by Jeff Barr | on 27 SEP 2016 | in Amazon EC2, Launch | Permalink | Comments

EC2's M4 instances offer a balance of compute, memory, and networking resources and are a good choice for many different types of applications.

We launched the M4 instances last year (read The New M4 Instance Type to learn more) and gave you a choice of five sizes, from **large** up to **10xlarge**. Today we are expanding the range with the introduction of a new **m4.16xlarge** with 64 vCPUs and 256 GiB of RAM. Here's the complete set of specs:

Instance Name	vCPU Count	RAM	Instance Storage	Network Performance	EBS-Optimized
m4.large	2	8 GiB	EBS Only	Moderate	450 Mbps
m4.xlarge	4	16 GiB	EBS Only	High	750 Mbps
m4.2xlarge	8	32 GiB	EBS Only	High	1,000 Mbps
m4.4xlarge	16	64 GiB	EBS Only	High	2,000 Mbps
m4.10xlarge	40	160 GiB	EBS Only	10 Gbps	4,000 Mbps
m4.16xlarge	64	256 GiB	EBS Only	20 Gbps	10,000 Mbps

The new instances are based on Intel Xeon E5-2686 v4 (Broadwell) processors that are optimized specifically for EC2. When used with Elastic Network Adapter (ENA) inside of a placement group, the instances can deliver up to 20 Gbps of low-latency network bandwidth. To learn more about the ENA, read my post, Elastic Network Adapter – High Performance Network Interface for Amazon EC2.

C5

Intel Xeon Skylake.

Supports AVX 512

Up to **72 vCPUs**
Up to **144 GiB** memory

Ideal for:

- Scientific computing
- 3D rendering
- Machine learning inference
- Distributed analytics

Coming Soon: Amazon EC2 C5 Instances, the next generation of Compute Optimized instances

Amazon EC2 C5 instances are the most powerful Compute Optimized instances and the best price to compute performance in EC2. C5 is the first AWS cloud instance based on Intel's next-generation Skylake Xeon® processors. They are ideal for compute-intensive workloads like ad serving, scientific modeling, 3D rendering, cluster computing, machine learning inference, and distributed analytics.

Coming in early 2017, C5 instances are powered by a new custom version CPU based on the next generation of Intel® Xeon® Processor family (code-named "Skylake") that feature a new microarchitecture and offer up to 72 vCPUs along with up to 144 GiB of memory. With support for new Intel AVX 512 advanced instruction, customers can more efficiently run vector processing workloads with single and double floating point precision, such as machine learning inference or video processing.

C5 instances use our next-generation Elastic Network Adapter (ENA), which has been optimized to deliver high packet per second (PPS) performance, low inter-instance latencies, and very low network jitter. C5 instances also feature new AWS hardware acceleration that delivers three times the Amazon EBS bandwidth of C4 instances, for workloads that require high amounts of IOPs. Overall, C5 instances are ideal for batch processing, distributed analytics, high performance science and engineering applications, ad serving, massively multiplayer online (MMO) gaming, and video encoding.

P2

- Up to 16 x Nvidia K80 GPUs in a single instance (Volta V100 is coming)
- Including peer-to-peer PCIe GPU interconnect
- Supporting a wide variety of use cases including deep learning, HPC simulations, and batch rendering

Instance Size	GPUs	GPU Peer to Peer	vCPUs	Memory (GiB)	Network Bandwidth*
p2.xlarge	1	-	4	61	1.25 Gbps
p2.8xlarge	8	Y	32	488	10 Gbps
p2.16xlarge	16	Y	64	732	25 Gbps

F1

- Up to 8 Xilinx Virtex UltraScale Plus VU9p FPGAs in a single instance with four high-speed DDR-4 per FPGA
- Largest size includes high performance FPGA interconnects via PCIe Gen3 (FPGA Direct), and bidirectional ring (FPGA Link)
- Designed for hardware-accelerated applications including financial computing, genomics, accelerated search, and image processing

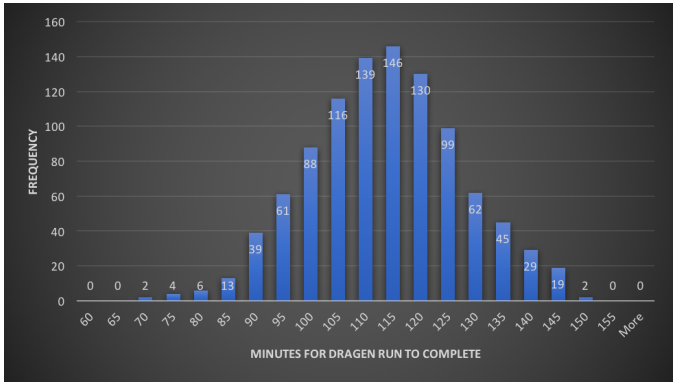
Instance Size	FPGAs	FPGA Link	FPGA Direct	vCPUs	Memory (GiB)	NVMe Instance Storage	Network Bandwidth*
f1.2xlarge	1	-		8	122	1 x 480	5 Gbps
f1.16xlarge	8	Y	Y	64	976	4 x 960	30 Gbps

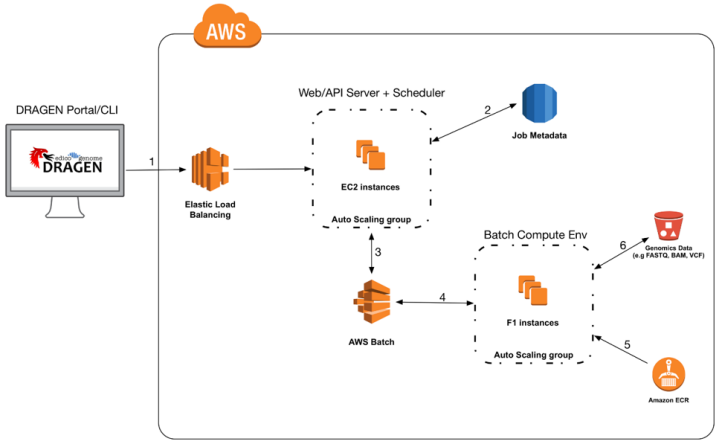
Accelerating Precision Medicine at Scale

- Children's Hospital of Philadelphia (CHOP) analysed 1000 genomes in 2.5 hours using **Edico DRAGEN**

- Sourced data from the Center for Applied Genomics Biobank
- Averaged approximately $3 per genome (AWS cost)

- Used 1000 x f1.2xlarge instances in a single AWS region
- Orchestrated with **AWS Batch** with all DRAGEN binaries in Docker

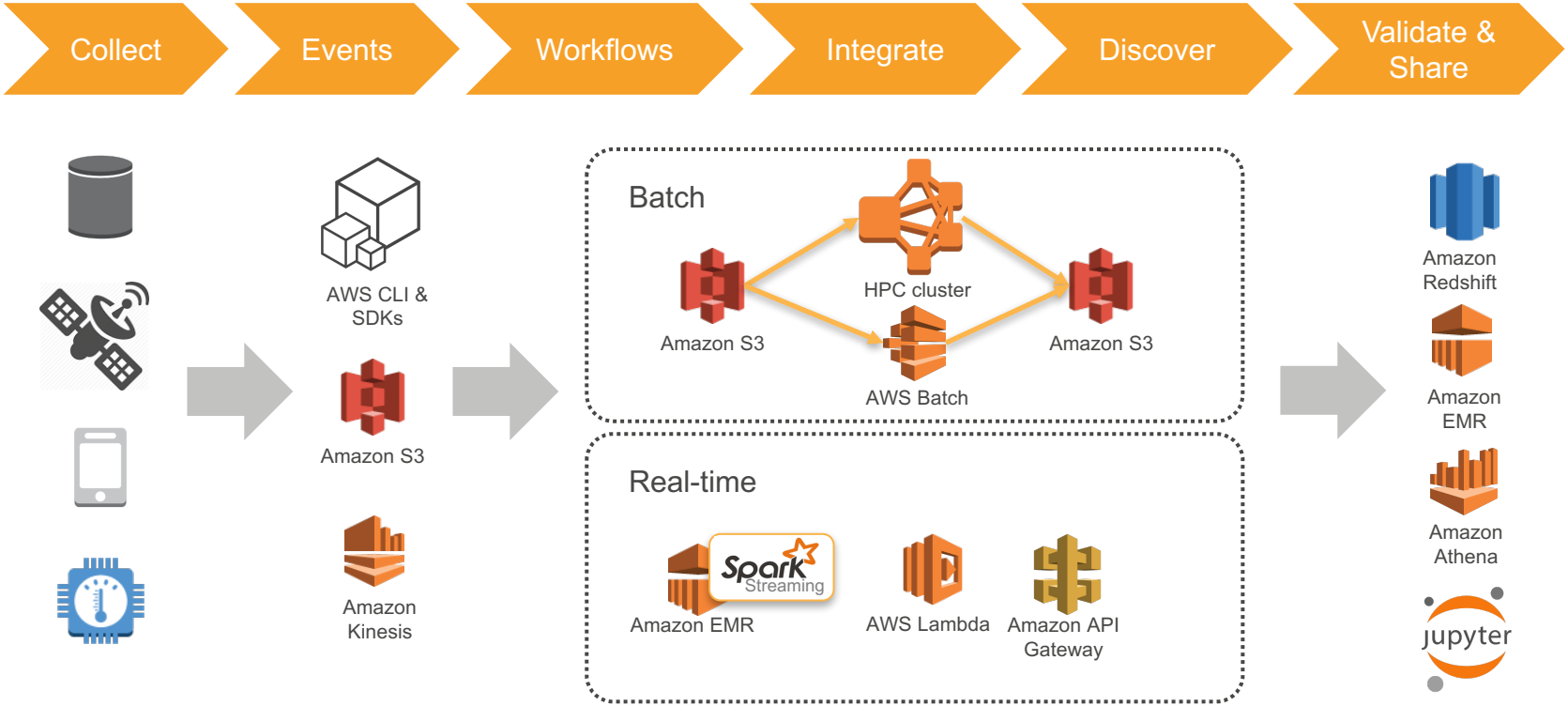




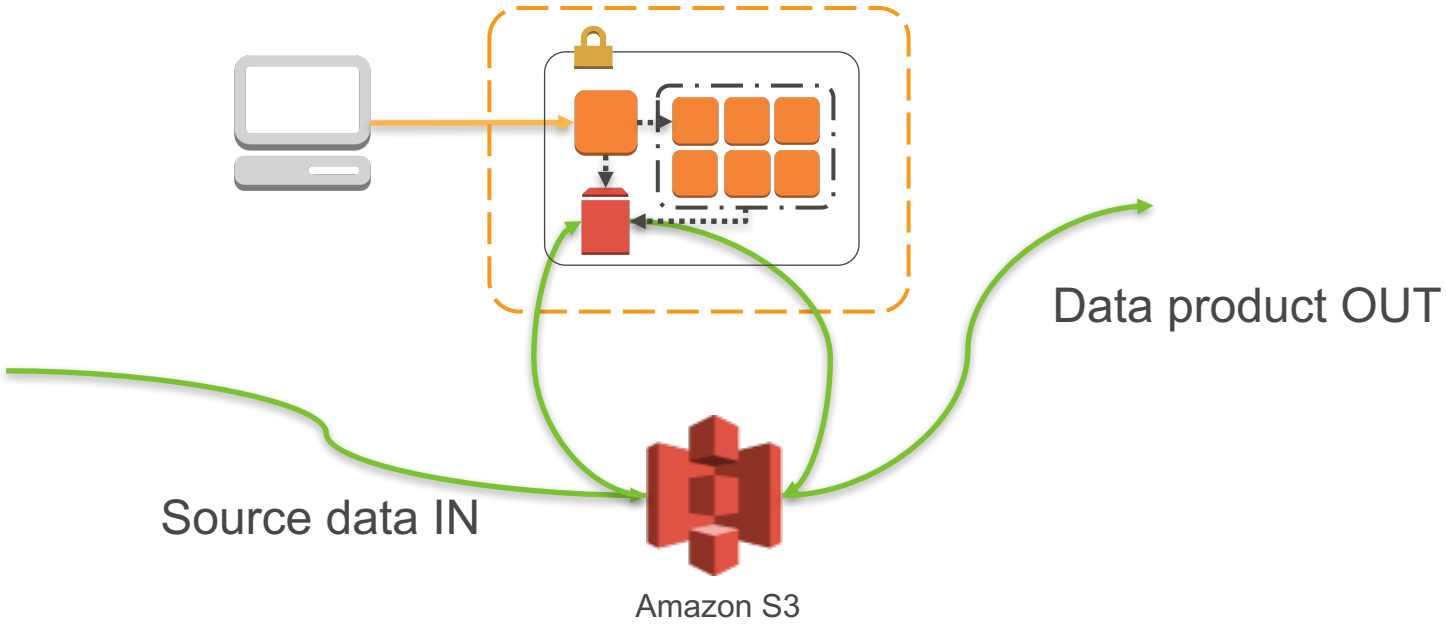
Source: https://aws.amazon.com/blogs/compute/accelerating-precision-medicine-at-scale/

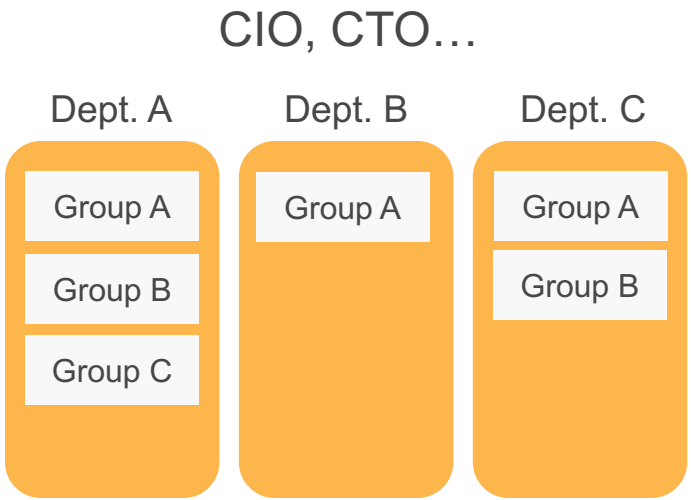
Research as workflow

Model Capability based on Research Workflows

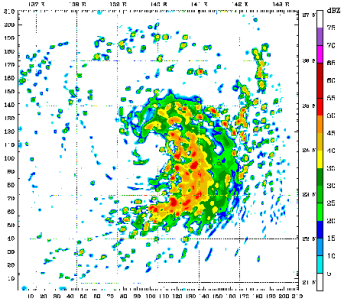


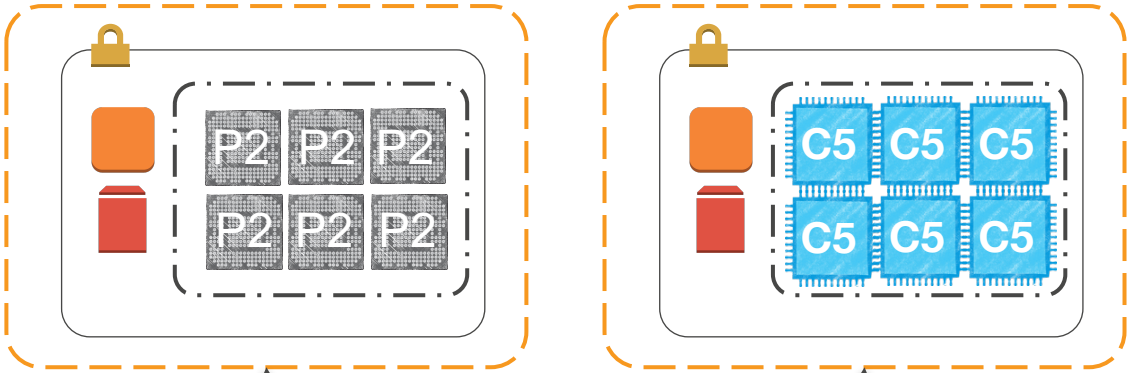
E.g. a cluster in the cloud is an ephemeral tool

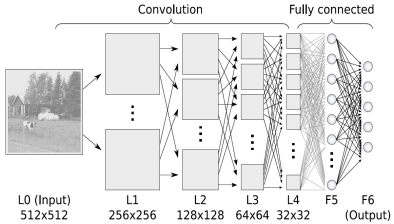



CIO, CTO…
Dept. A
Dept. B
Dept. C
Group A
Group B
Group C
Group A
Group A
Group B

Clusters in the cloud are fit for purpose



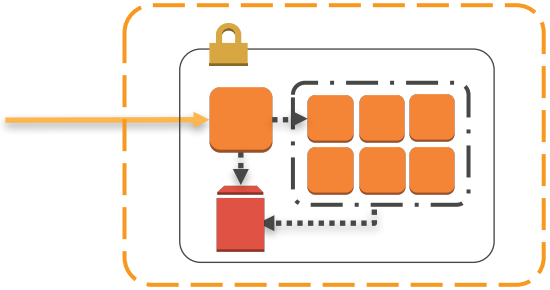


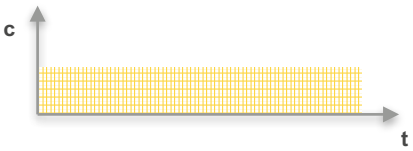


Amazon S3

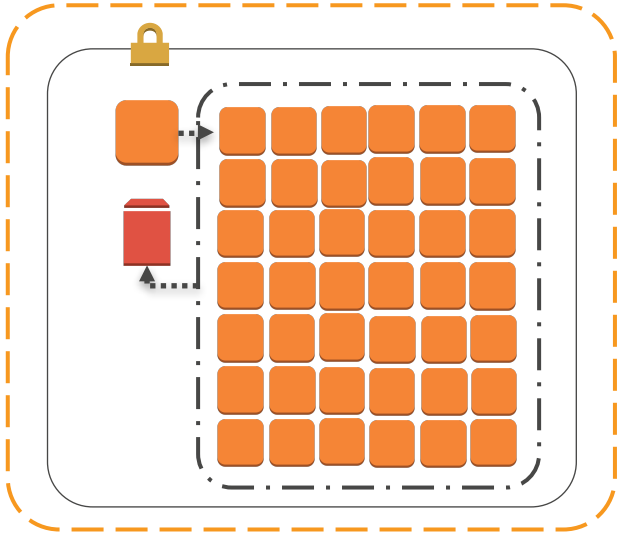
Clusters can scale and are elastic

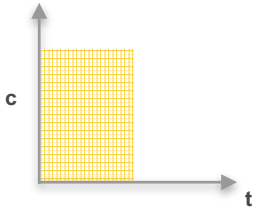
W = 1, C =1



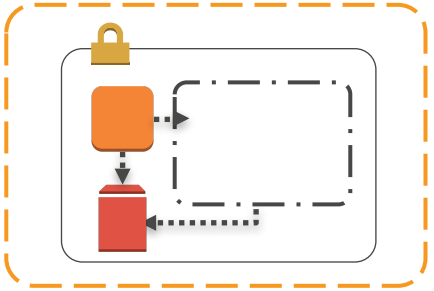


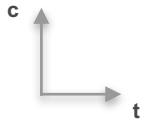
W = n, C = n





W = 0, C ~ 0





Ready-made HPC+HTC Tools

HPC and HTC tools on AWS

CfnCluster

CfnCluster is provided by AWS to quickly provision configurable HPC and HTC cluster environments

Alces Flight is available in the AWS Marketplace and bundles 1000+ commonly used scientific applications https://aws.amazon.com/marketplace/

+

AWS Batch

AWS Batch provides compute resources via Docker containers with user-definable queues and an optimised job scheduler

Amazon EMR

Amazon EMR provides a managed Hadoop framework supporting Apache Spark, HBase, Presto, and Flink, Spark MLLib on Amazon EC2 and EC2 Spot

‘Self-service’ HPC in 2017

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.

1,150+ popular scientific applications

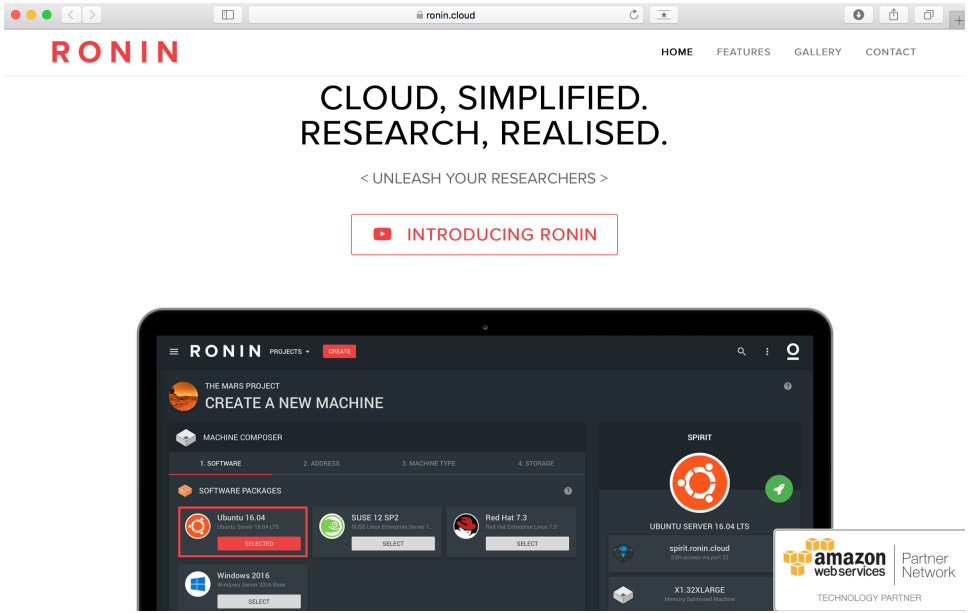
- Multiple versions, complete with libraries and various compiler optimizations, ready to run
- Supports Docker and Singularity
- Slurm default scheduler (also PBS Pro, SGE etc)

Available via the AWS Marketplace

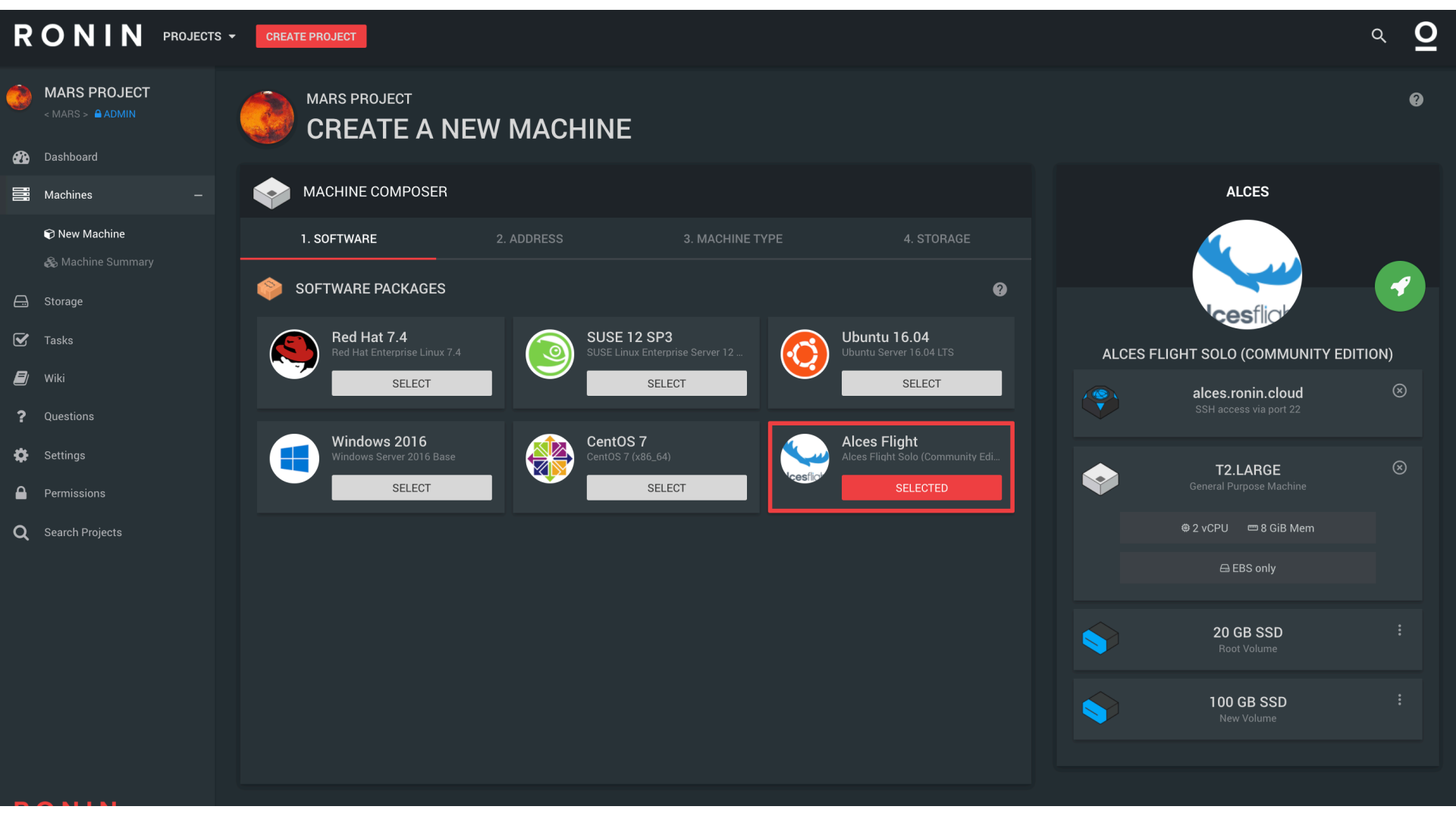
http://alces-flight.com/ for more information

‘ResearchTechs’ on AWS

http://alces-flight.com



http://ronin.cloud


RONIN
PROJECTS
CREATE PROJECT
MARS PROJECT
< MARS > ADMIN
Dashboard
Machines
New Machine
Machine Summary
Storage
Tasks
Wiki
Questions
Settings
Permissions
Search Projects
MARS PROJECT
CREATE A NEW MACHINE
MACHINE COMPOSER
1. SOFTWARE
2. ADDRESS
3. MACHINE TYPE
4. STORAGE
SOFTWARE PACKAGES
Red Hat 7.4
Red Hat Enterprise Linux 7.4
SELECT
SUSE 12 SP3
SUSE Linux Enterprise Server 12 ...
SELECT
Ubuntu 16.04
Ubuntu Server 16.04 LTS
SELECT
Windows 2016
Windows Server 2016 Base
SELECT
CentOS 7
CentOS 7 (x86_64)
SELECT
Alces Flight
Alces Flight Solo (Community Edi...
SELECTED
ALCES
ALCES FLIGHT SOLO (COMMUNITY EDITION)
alces.ronin.cloud
SSH access via port 22
T2.LARGE
General Purpose Machine
2 vCPU
8 GiB Mem
EBS only
20 GB SSD
Root Volume
100 GB SSD
New Volume

Demo: Alces Flight

Remote Visualisation

Graphics and Collaboration

Cloud can be used for pre and post processing as well as HPC

- GPUs in the cloud for remote rendering
- Amazon WorkSpaces for managed Windows desktops
- NICE DCV and other protocols for remote graphics and WAN optimization
- Commercial and Open Source Protocols such as RDP, VNC, NoMachine (NX), Exceed onDemand, etc

Cloud is more secure for collaboration

- Encrypt the data in flight and at rest
- Manage your own keys and credentials
- Deliver pixels to your collaborators, not the actual data

G3 Instances – “Graphics”

- Powered by NVIDIA Tesla M60 GPUs & 2.7Ghz Intel Broadwell Processors
- Each GPU supports 8 GiB of GPU memory, 2048 parallel processing cores
- Supports OpenGL 4.5, DirectX 12.0, CUDA 8.0, and OpenCL 1.2.
- Parallel encoding of 10x H.265 (HEVC) 1080p30 streams and up to 18x H.264 1080p30 streams

Model	GPUs	GPU Memory	Cores	Main Memory	EBS Bandwidth
g3.4xlarge	1	8 GiB	8 (16 vCPUs)	122 GiB	3.5 Gbps
g3.8xlarge	2	16 GiB	16 (32 vCPUs)	244 GiB	7 Gbps
g3.16xlarge	4	32 GiB	32 (64 vCPUs)	488 GiB	14 Gbps

NICE DCV Streaming Protocol

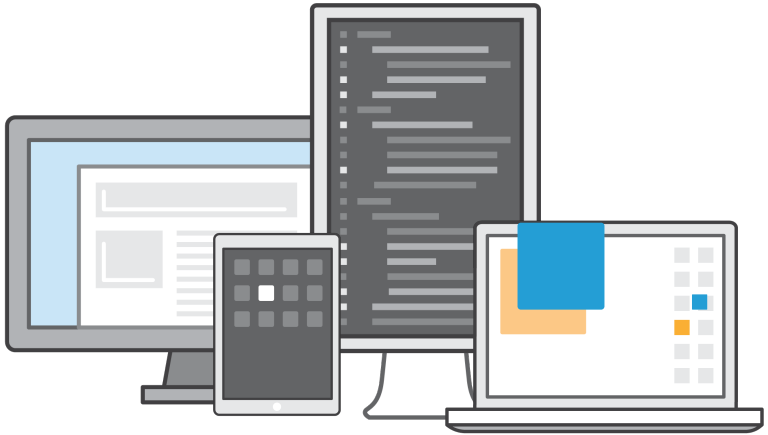
- High fidelity visualization delivered to browsers
- HTTPS access via streaming gateways
- Adaptive and responsive streaming
- AES-256 encrypted
- Supports both 3D and non-graphics applications

Desktop Application Streaming

amazon AppStream 2.0

Run Desktop Apps in a Web Browser

Secure apps & data

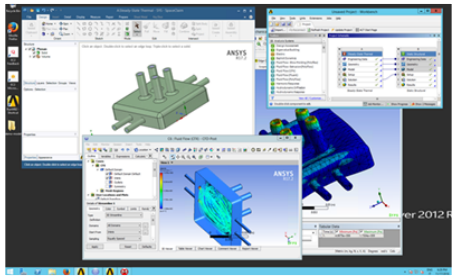


Stream desktop applications securely to any web browser

Pay-as-you-go

Scale globally

Elastic GPU



Workstation-Class Graphics Performance

Elastic GPUs are capable of running a variety of graphics workloads, such as 3D modeling and rendering, with similar workstation performance compared to direct-attached GPUs.

Optimized Performance and Cost

Elastic GPUs come in multiple frame buffer sizes up to 8GB, allowing you to achieve the optimal graphics performance for your workload for the lowest possible cost.

Application Support

Elastic GPUs plan to support Open Graphics Library (OpenGL), a cross-language, cross-platform API for rendering 2D and 3D vector graphics, with a roadmap for certification of additional APIs.

Graphics Certification Program

AWS now offers a Graphics Certification Program for software vendors and developers to help make sure that their applications take full advantage of Elastic GPUs and our other GPU-based offerings. To accelerate availability of the applications being certified, the program is supported with AWS credits and co-marketing funds for qualifying applications.

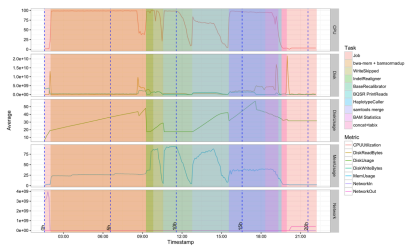
Name	GPU Memory
eg1.medium	1 GiB
eg1.large	2 GiB
eg1.xlarge	4 GiB
eg1.2xlarge	8 GiB

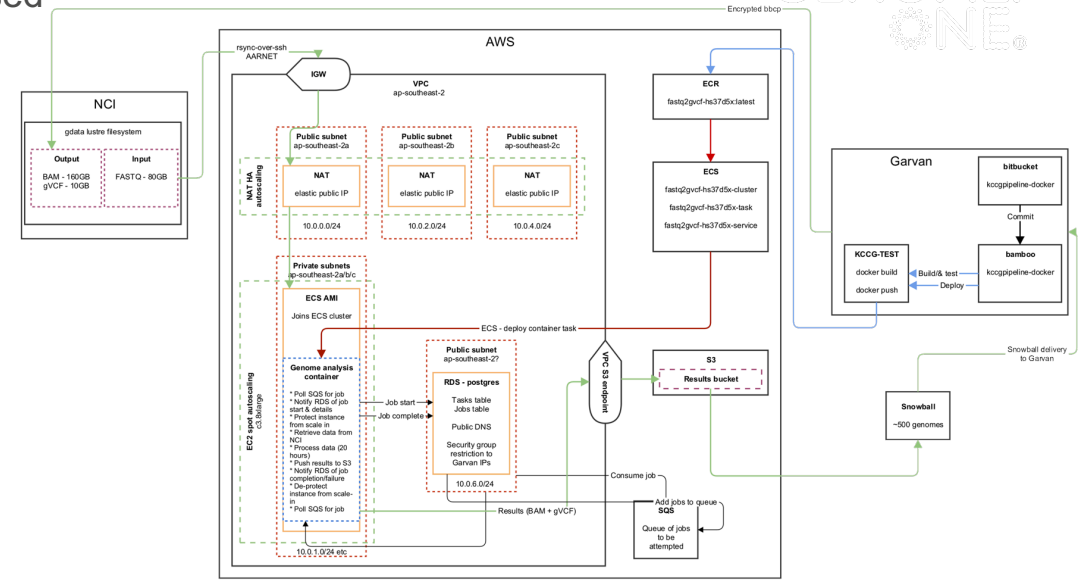
A familiar HPC architecture: moved to AWS

Garvan Institute – Containerized Bioinformatics

- The Garvan sequences **18,000** human genomes per year
- Have processed 4,500 genomes using 150 c3.8xlarge **spot instances**
- Used 3 million CPU-hours and processed 1.1 PB of data
- Use Docker with AWS ECS and ECR

- Realized a **20x cost saving** on AWS
- Have used AWS to containerize their genome analysis and run at scale - leveraging ECS, SQS and RDS





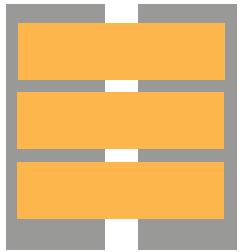
Source: https://www.slideshare.net/AmazonWebServices/containerised-bioinformatics-pipeline-on-aws

But it’s not (just) about servers…

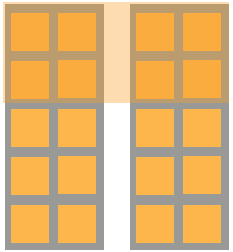
Evolving Compute Abstractions

Physical

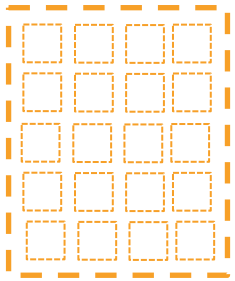
Virtualisation



Containerization



Serverless



AWS Lambda – How it Works

Bring your own code

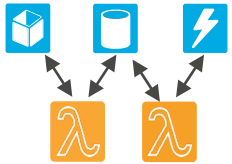
Node.JS, Java, Python

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

Event or RequestResponse invoke options

Existing integrations with various AWS services

Fine grained permissions

- Uses IAM role for Lambda execution permissions
- Uses Resource policy for AWS event sources

Lambda in the context of HPC & HTC

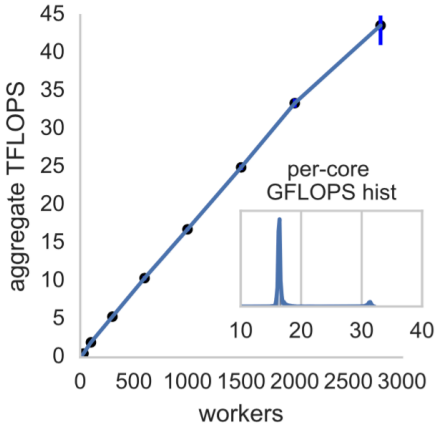


Figure 2: **Running a matrix multiplication benchmark inside each worker, we see a linear scalability of FLOPs across 3000 workers.**

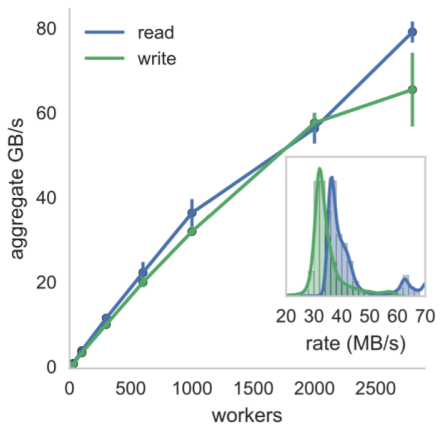


Figure 3: **Remote storage on S3 linearly scales with each worker getting around 30 MB/s bandwidth (inset histogram).**

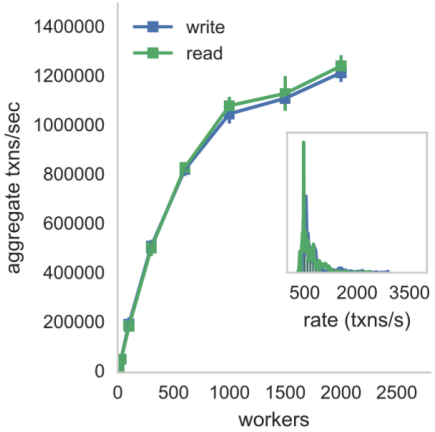


Figure 4: **Remote key-value operations to Redis scales up to 1000 workers. Each worker gets around 700 synchronous transactions/sec.**

Source: “Occupy the Cloud: Distributed Computing for the 99%”
https://arxiv.org/pdf/1702.04024.pdf

riselab
UC Berkeley

An exemplar cloud-native architecture

CSIRO – CRISPR search with AWS Lambda

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

Serverless:

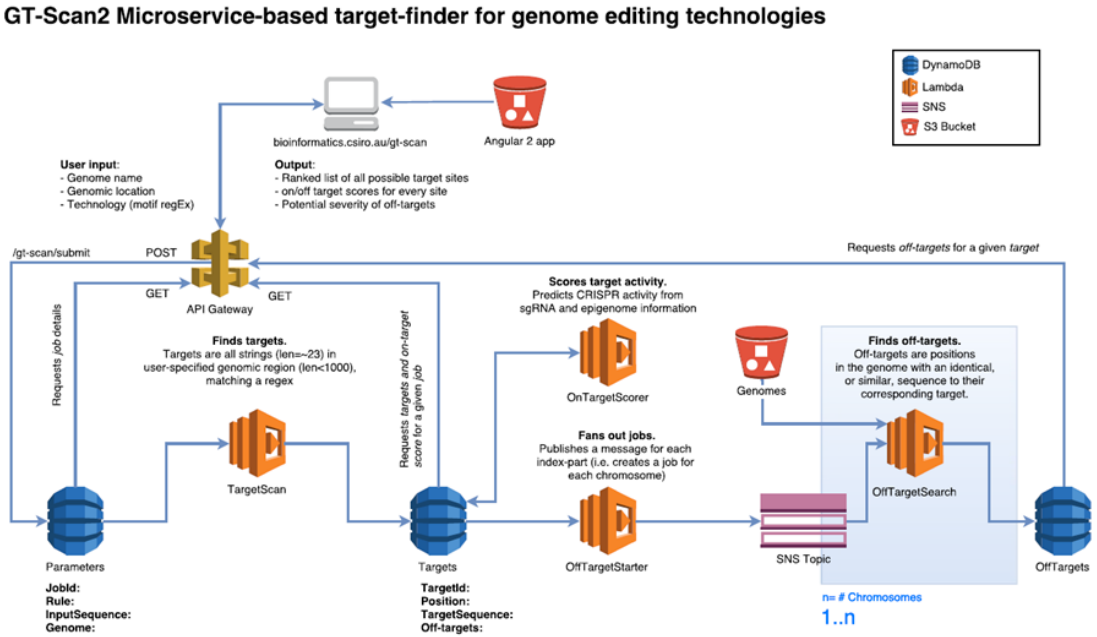
- Does not require users to have high-compute power

Scalable:

- Can be easily scaled to whole genome analysis

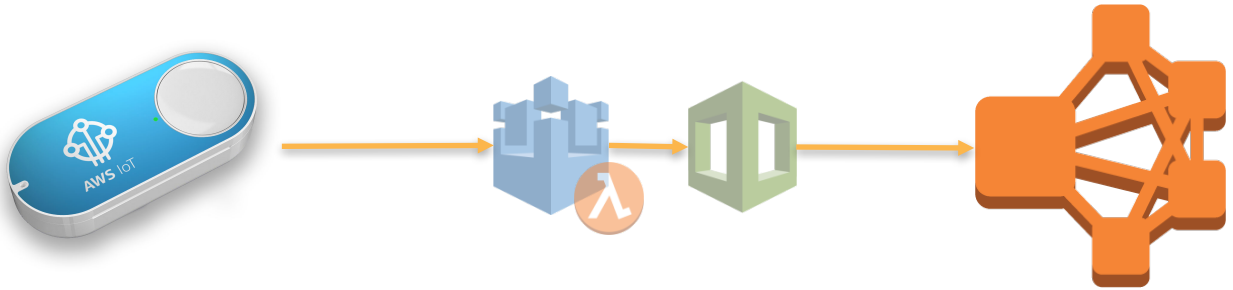
Also implement as a “stand-alone”

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



Source: https://aws.amazon.com/blogs/aws/genome-engineering-applications-early-adopters-of-the-cloud/

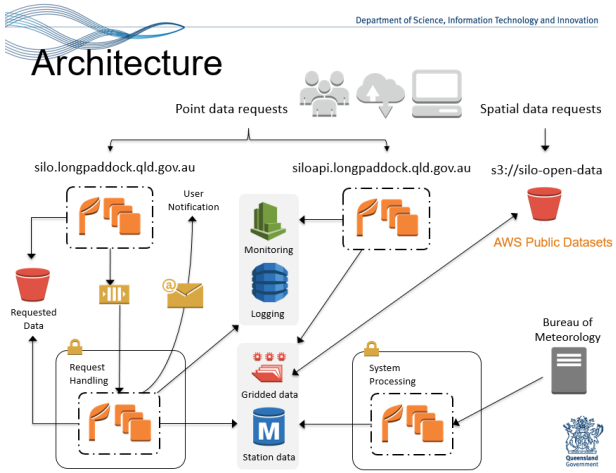
Demo: Alces Flight (v2)



Open Data at AWS

Open Data at AWS

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



The Big Data Challenge

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

Accessing data at scale is often a prohibitive challenge.

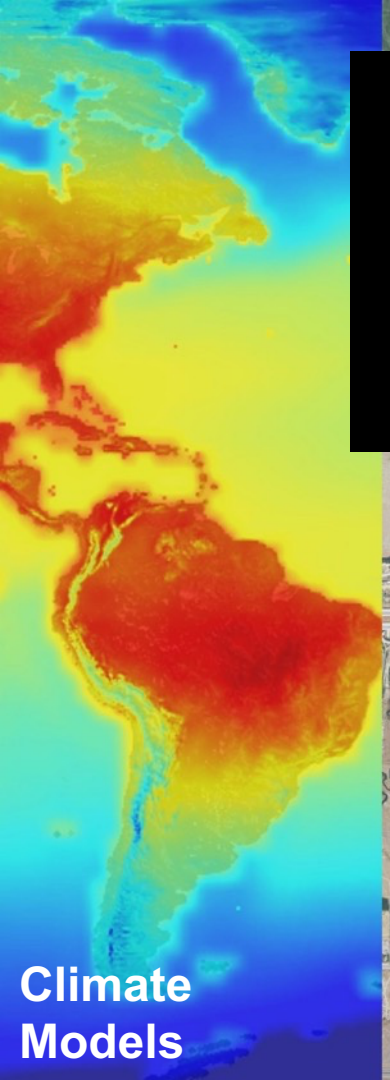
Our Solution – Shared Open Data on AWS

AWS global footprint makes it a powerful platform for scientific collaboration. Users and compute can be brought to the data. AWS offers many advanced big data related services.

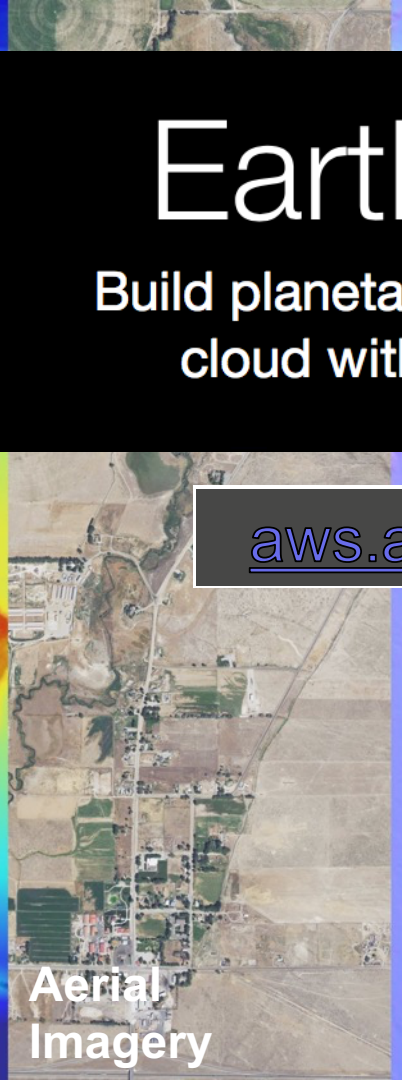
Earth on AWS

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

aws.amazon.com/earth

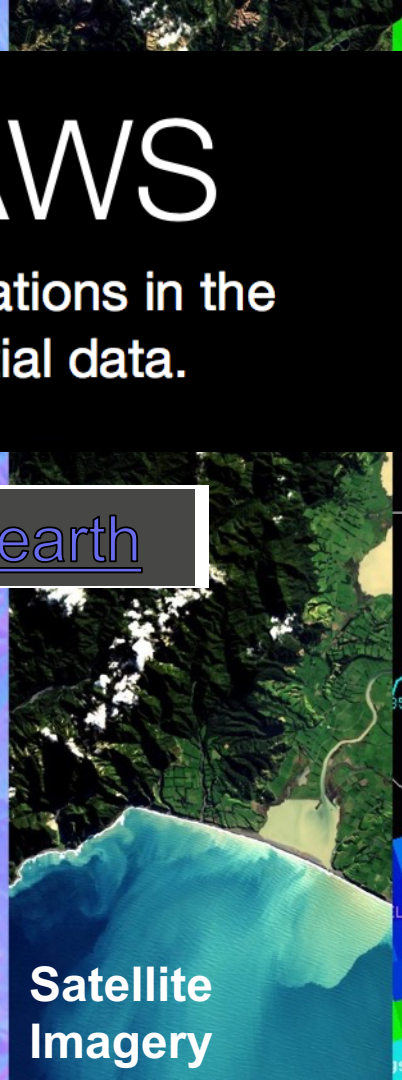


Climate Models

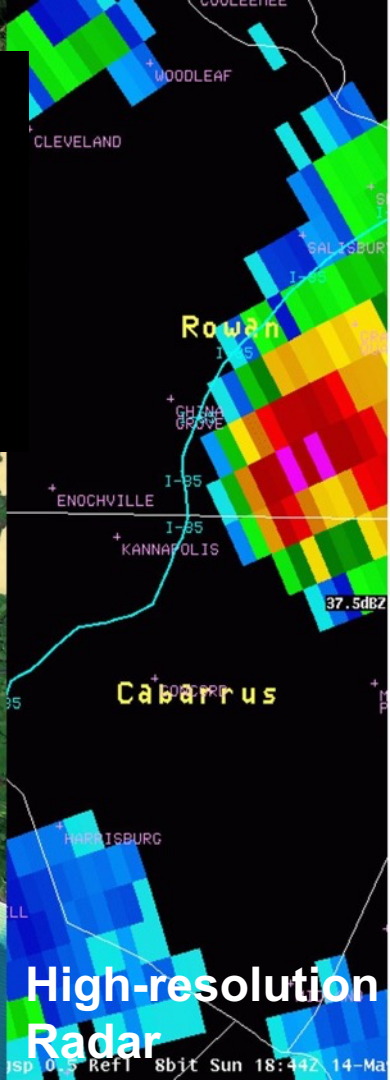


Aerial Imagery

Elevation Models



Satellite Imagery



High-resolution Radar

Landsat 8 Lives on AWS

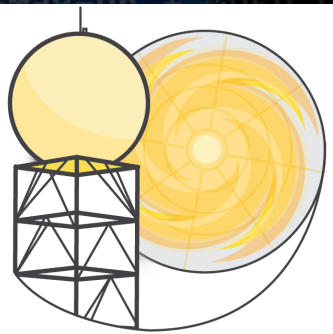
Within the first year of Landsat on AWS, the data has been requested over 1 billion times, globally. Over 400,000 scenes are now available

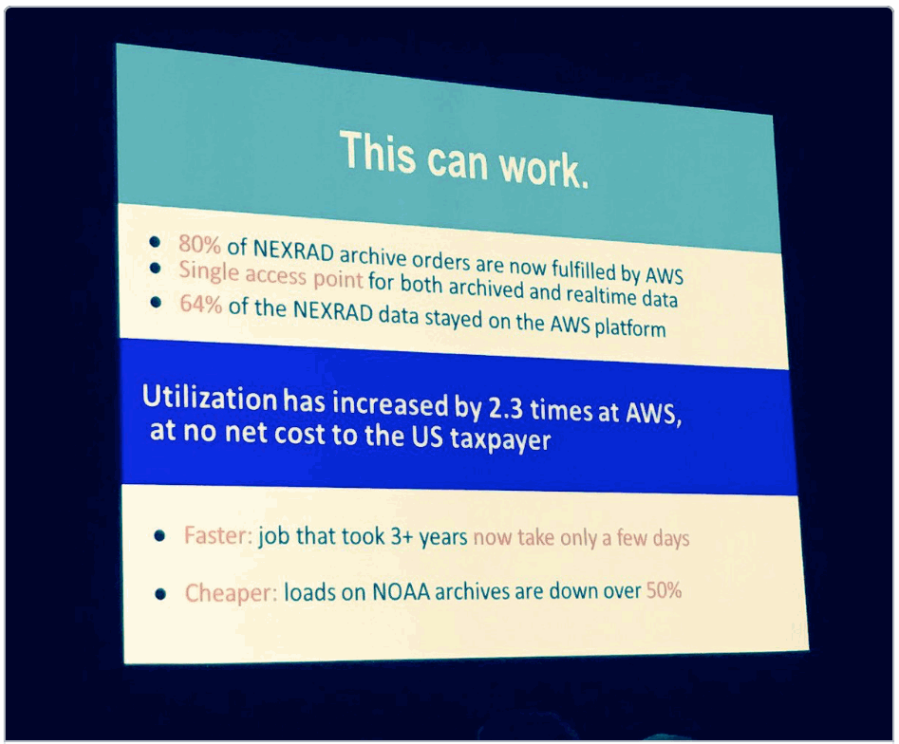
NEXRAD on AWS

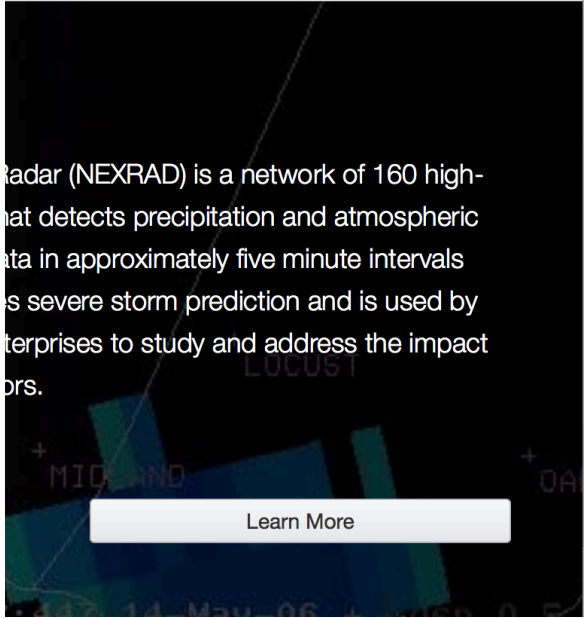
The Next Generation Weather Radar (NEXRAD) is a network of 160 high-resolution Doppler radar sites that detects precipitation and atmospheric movement and disseminates data in approximately five minute intervals from each site. NEXRAD enables severe storm prediction and is used by researchers and commercial enterprises to study and address the impact of weather across multiple sectors.

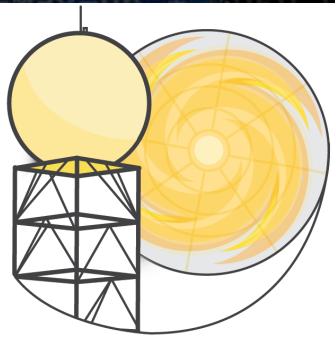
Image from NEXRAD data

Learn More









DJ Patil
@DJ44

Follow

Wow. What happens when @NOAA puts their data on the cloud. #WHOpenData 👊👍🙌🇺🇸

4:04 PM - 28 Sep 2016

92 124

Research Programs at AWS

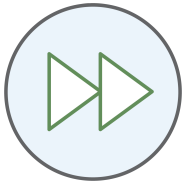
AWS Research Cloud Program

Science first, not servers. Researchers are not professional IT people (nor do they wish to be).

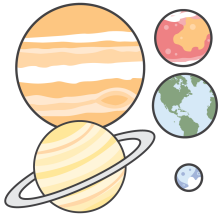
Simple and easily explained procedures to get set up with cloud access.

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.

IT'S ABOUT SCIENCE, NOT SERVERS.

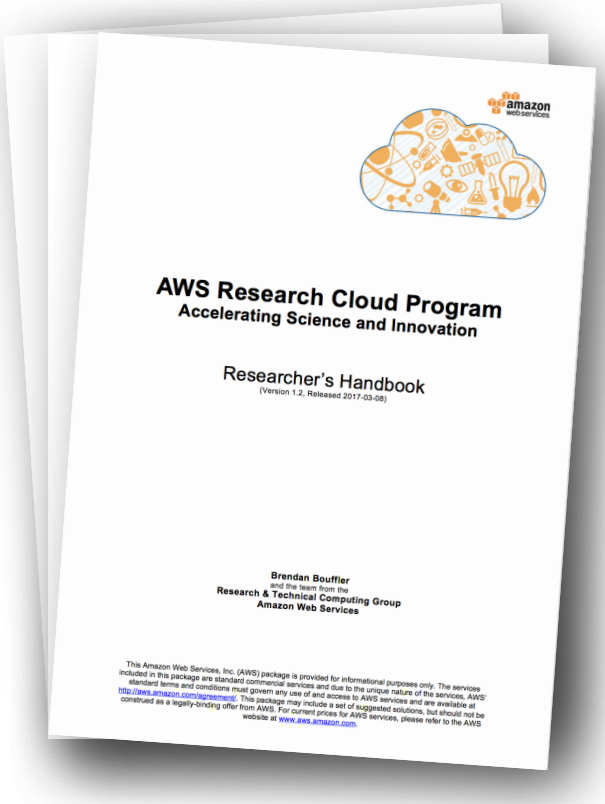
We recognize that whilst research is often a compute-intensive activity, most researchers are not IT experts.

We want to simplify research in the cloud with easy-to-use tools for researchers and their students, and share the catalogue of "researcher-obsessed" products and services created by many of our partners.

aws.amazon.com/rcp

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

Global Data Egress Waiver

Why?

Researchers need predictable budgets

Who?

Available to Degree-granting / Research Institutions in APAC (and elsewhere)

What?

Waives data egress charges from Qualified Accounts (capped at 15% of total spend)

How?

Contract Addendum Required.

Talk to your AWS account team.

All qualifying research customers should use this!

Cloud Credits for Research

The Cloud Credits for Research program aims to support:

1. Building tools to facilitate future research
2. Performing proof of concept for research or open data workloads
3. Training the research community on the usage of the cloud

Also available to incubate ”centers of excellence” in research on the AWS cloud.

Data Analytics

Evolution of Data Analytics

Batch

Amazon EMR

Amazon Redshift

AWS Batch

Real time

Amazon Kinesis

Amazon Kinesis Analytics

Amazon SNS

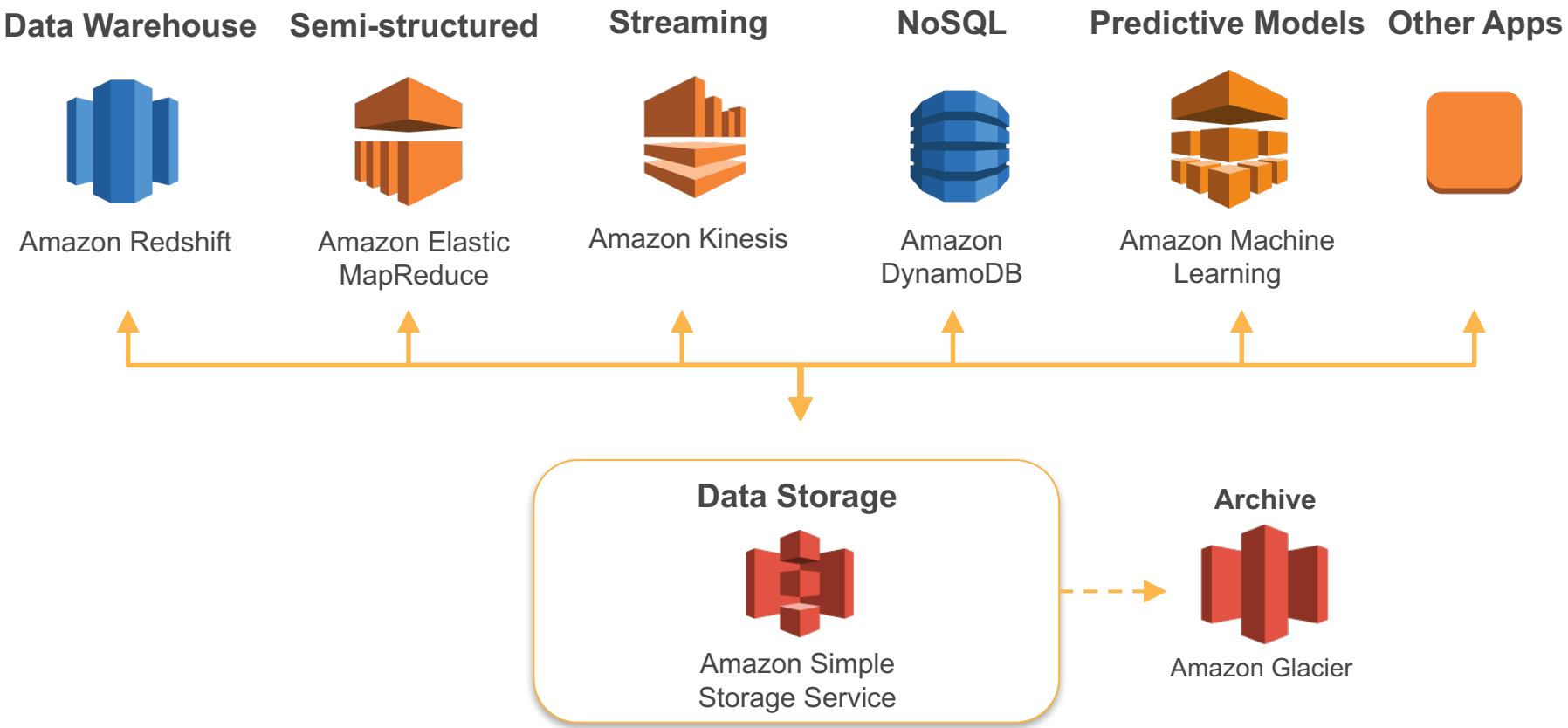
AWS IoT

Prediction

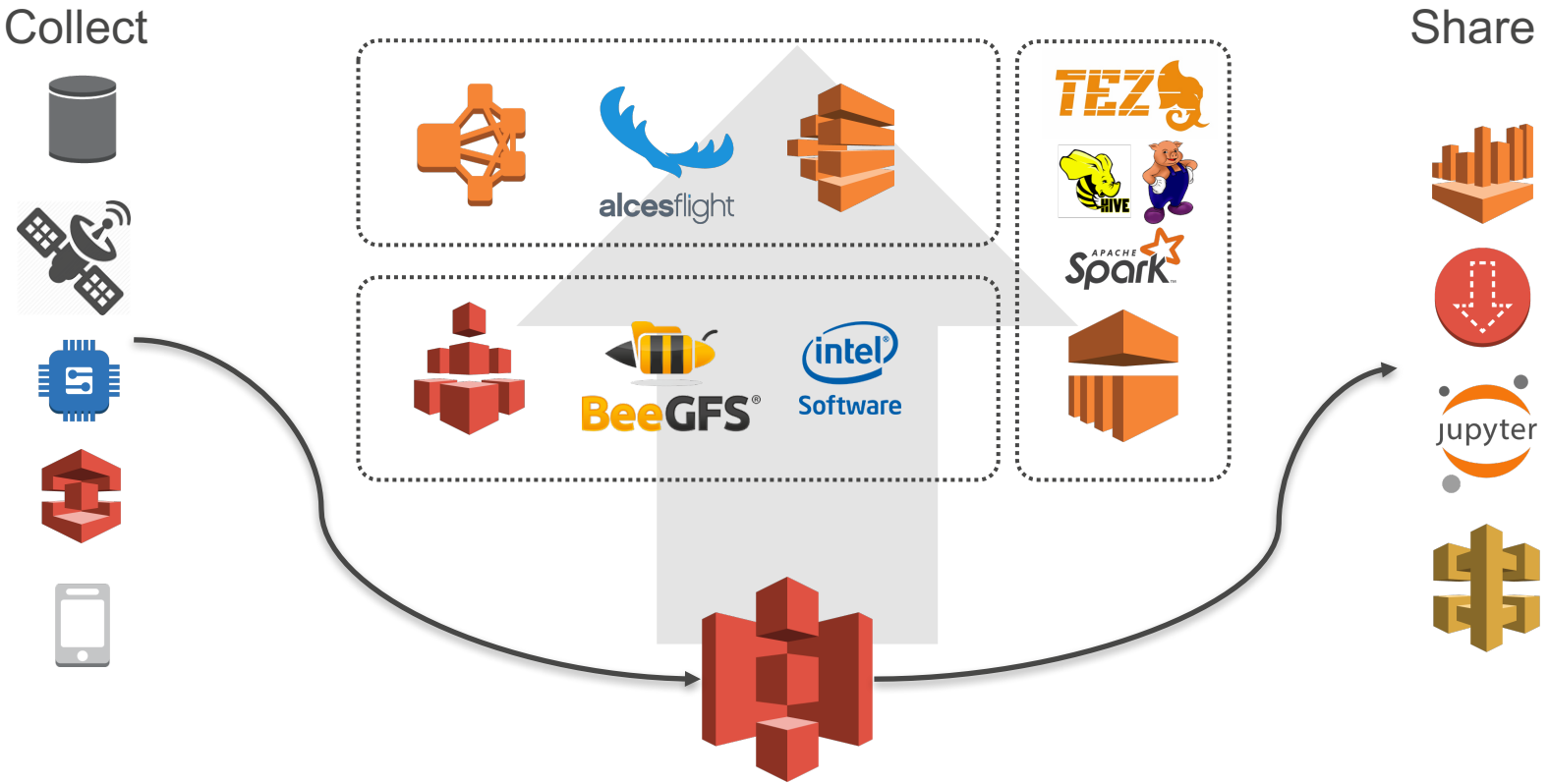
Amazon Machine Learning

Amazon Rekognition

Use an optimal combination of highly interoperable services

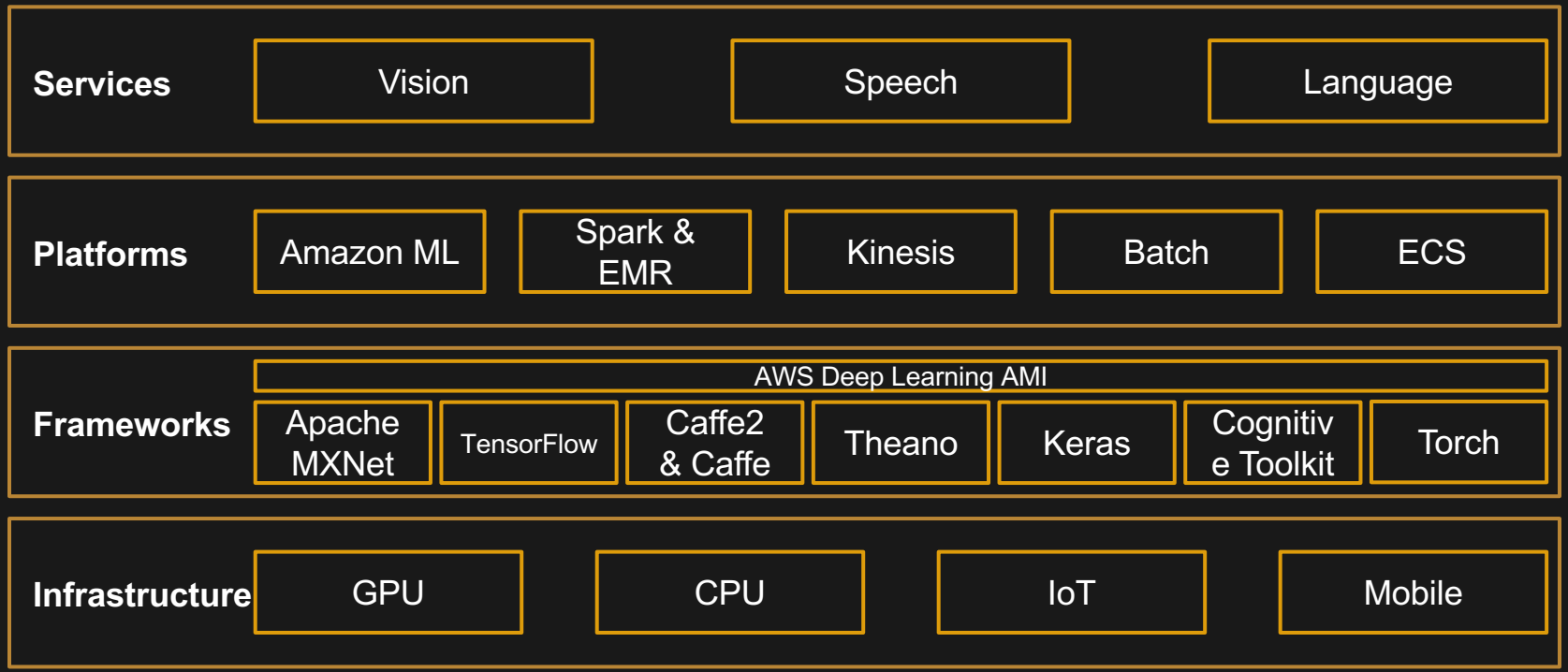


S3 as a Data Lake

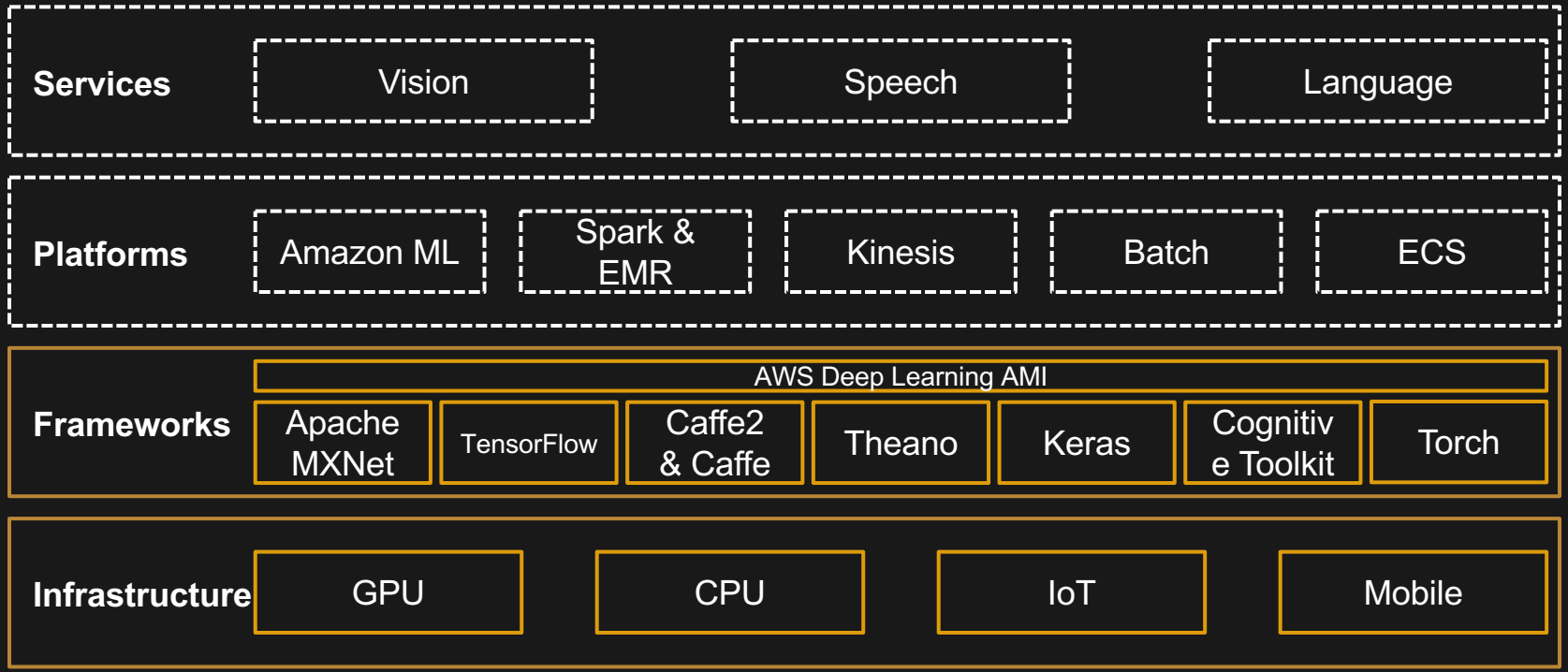


Machine Learning

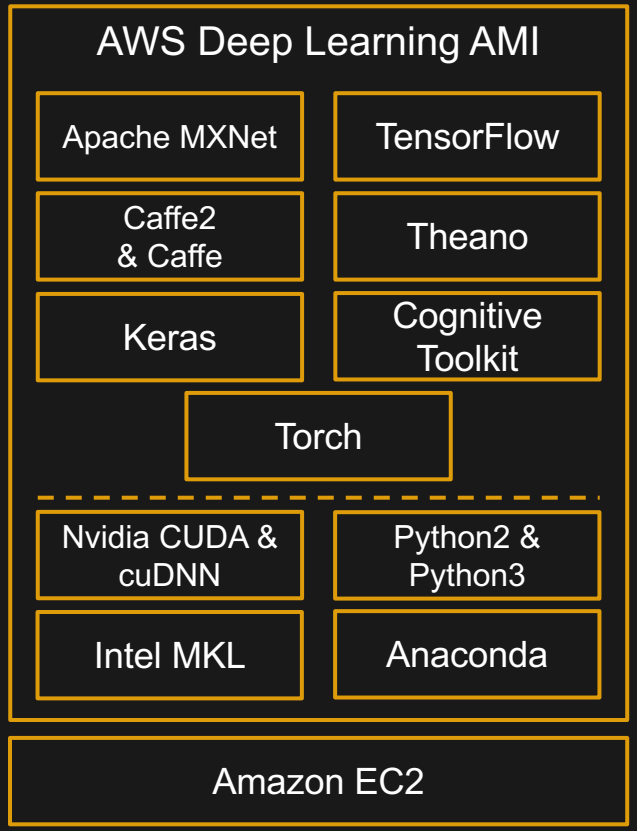
The Amazon ML Stack



Frameworks & Infrastructure

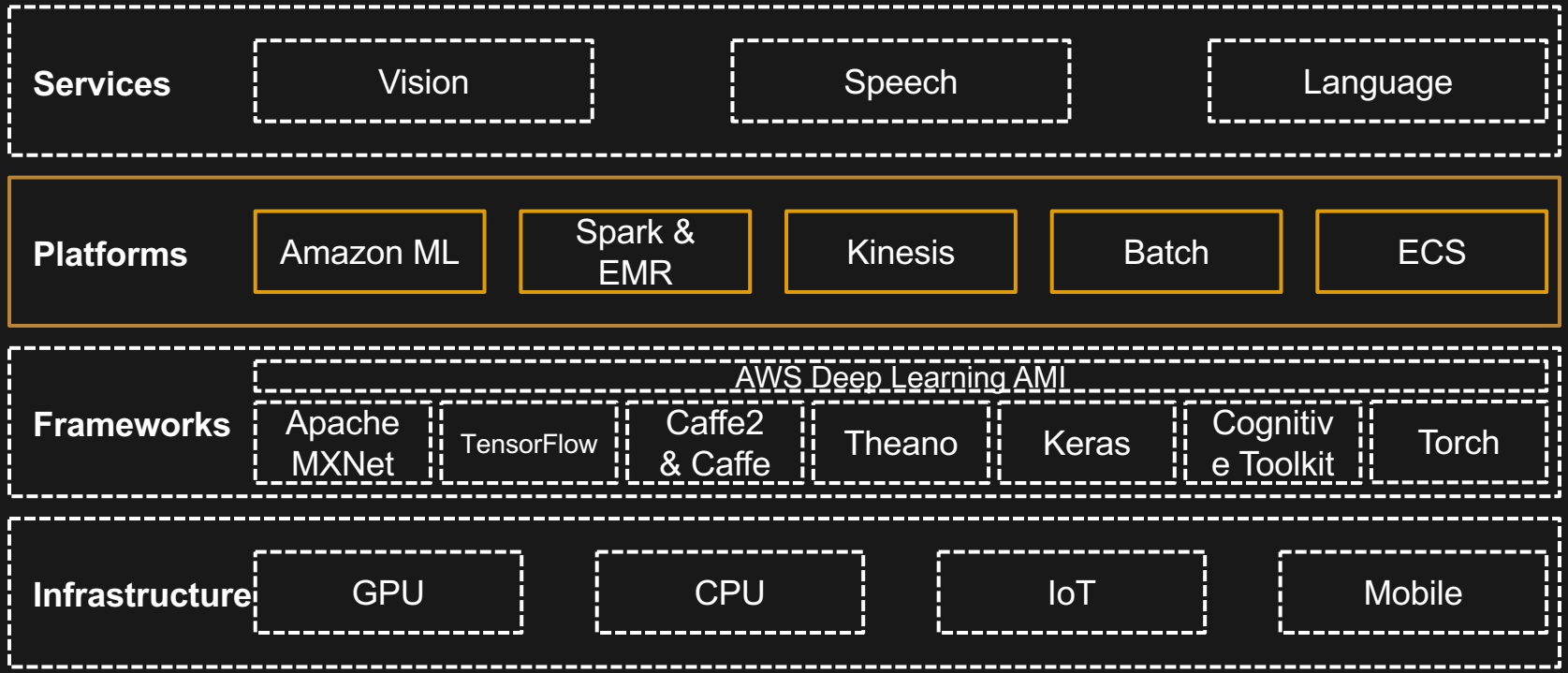


Support for all major frameworks

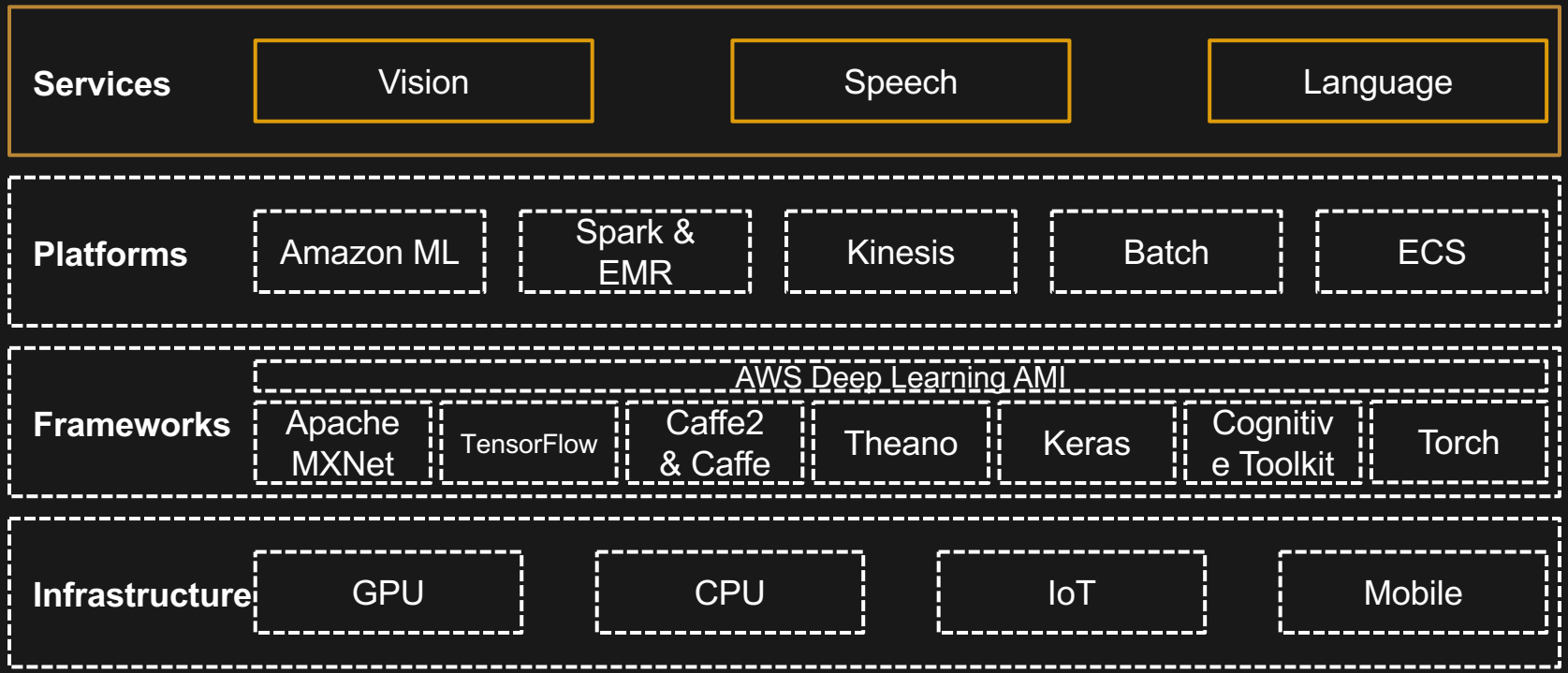


Expedia
Zillow
Mapillary
BUILD FAX PROPERTY HISTORY
OpenAI
move
SCDM
Fraud.net
NVIDIA
JAM CITY
tu Simple
clarifai
Carnegie Mellon
ADBRAIN
NETFLIX
SIGOPT
Horizon Robotics
TwoSense
edmunds
CLAIRE
Pinterest
UNIVERSITY of WASHINGTON
Missing Link
WOLFRAM
julia
UCLA
zendesk
intel

Machine Learning Platforms



API-driven Services



Amazon Rekognition

Potential Use Cases

Searchable Image Library

Face-based User Verification

Facial Recognition

Detect Inappropriate Content in Images

Sentiment Analysis

Celebrity Identification

C-SPAN

influicity

realtimes

SmugMug

Kaliber
ENABLING INFORMATION ASSURANCE

Openinfluence

Mobilink

GO GIRL

THETAKE

zmags

MOTOROLA SOLUTIONS

Artfinder

bynder

SHERIFF
SHERIFF'S OFFICE
WASHINGTON COUNTY
OR

Sturdy

SOCIAL SOUP

witlee

Amazon Polly

Potential Use Cases

Content Creation

Mobile & Desktop Applications

Internet of Things (IoT)

Education & E-learning

Customer Contact Center

Accessibility

SAMSUNG SmartThings

GoAnimate

duolingo

amazon RAPIDS

mapbox

The Washington Post

BeeLiked SOCIAL POLLINATION

Storybulbs

iTranslate

inhealthcare

Aloud

Wizkids

RNIB Supporting people with sight loss

aculabcloud A true cloud telecoms platform

Amazon Lex

Potential Use Cases

Appointment Booking

Informational Services

Internet of Things (IoT)

Customer Support

Access Enterprise Data

infor

Liberty Mutual

CapitalOne

TRD TOYOTA RACING DEVELOPMENT

NASA

KELLEY BLUE BOOK OFFICIAL GUIDE

HubSpot

Vonage

freshdesk

American Heart Association

OhioHealth

twilio

slack

Jupyter Notebooks on AWS

Research customers are increasingly doing exploratory data science and analytics work using notebooks.

Jupyter on AWS allows researchers to take advantage of any AWS compute node type:

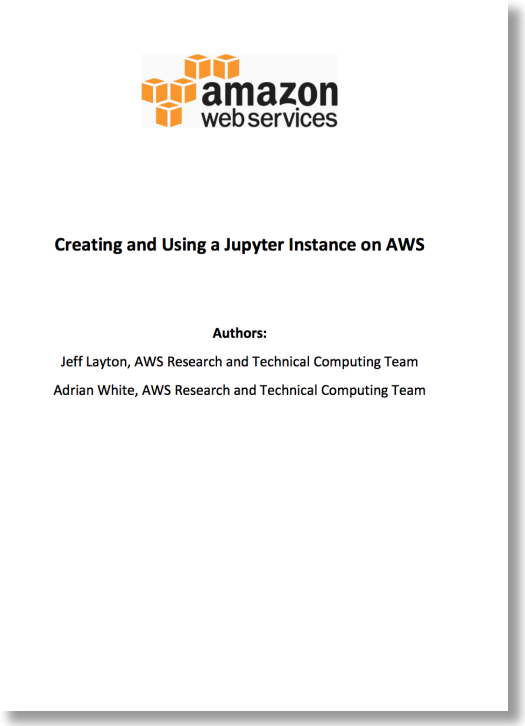
- Large memory, CPU optimized, IO optimized
- GPU nodes (e.g. multiple K80 GPUs)

Researchers can also access Batch, HPC and Spark/Mllib clusters with Jupyter

How to:

[Run Jupyter Notebook and JupyterHub on Amazon EMR]

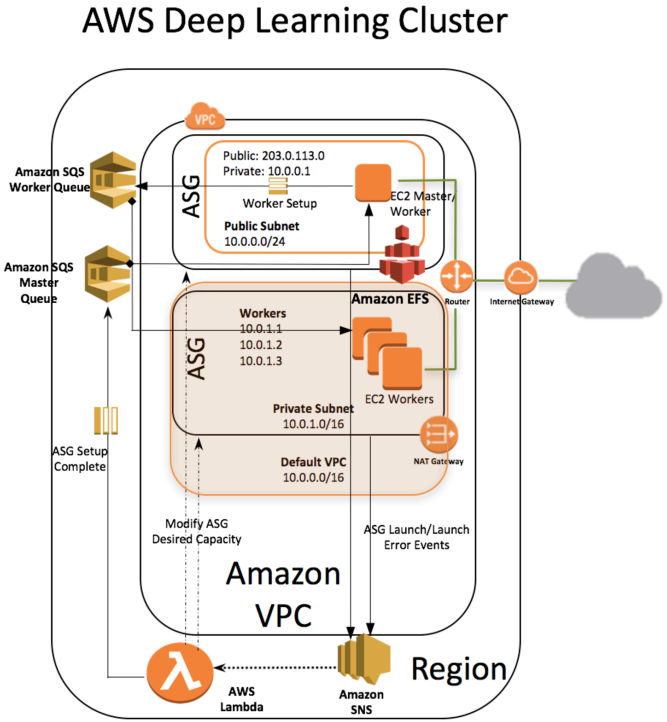
[Creating and Using a Jupyter Instance on AWS]



Distributed Deep Learning on AWS

- Distributed training across GPUs or CPUs using MXNet
- Spin up a cluster in minutes
- Automatically add or remove cluster nodes
- Supports Amazon EFS share filesystem
- Available on GitHub
  https://github.com/awslabs/deeplearning-cfn

```
INFO:root:Epoch[18] Time cost=6.020
INFO:root:Epoch[18] Validation-accuracy=0.992237
INFO:root:Epoch[19] Batch [100] Speed: 9960.58 samples/sec      Train-accuracy=1.000000
INFO:root:Epoch[19] Batch [200] Speed: 9830.53 samples/sec      Train-accuracy=1.000000
INFO:root:Epoch[19] Batch [300] Speed: 10233.37 samples/sec     Train-accuracy=1.000000
INFO:root:Epoch[19] Batch [400] Speed: 9747.18 samples/sec      Train-accuracy=1.000000
INFO:root:Epoch[19] Batch [500] Speed: 9913.81 samples/sec      Train-accuracy=1.000000
INFO:root:Epoch[19] Batch [600] Speed: 10109.95 samples/sec     Train-accuracy=1.000000
INFO:root:Epoch[19] Batch [700] Speed: 9935.70 samples/sec      Train-accuracy=1.000000
INFO:root:Epoch[19] Batch [800] Speed: 10210.48 samples/sec     Train-accuracy=1.000000
INFO:root:Epoch[19] Batch [900] Speed: 9824.65 samples/sec      Train-accuracy=1.000000
INFO:root:Epoch[19] Train-accuracy=1.000000
INFO:root:Epoch[19] Time cost=6.021
INFO:root:Epoch[19] Validation-accuracy=0.992237
[ec2-user@ip-10-0-0-224 image-classification]$
```



Case Study: Detecting Meteors

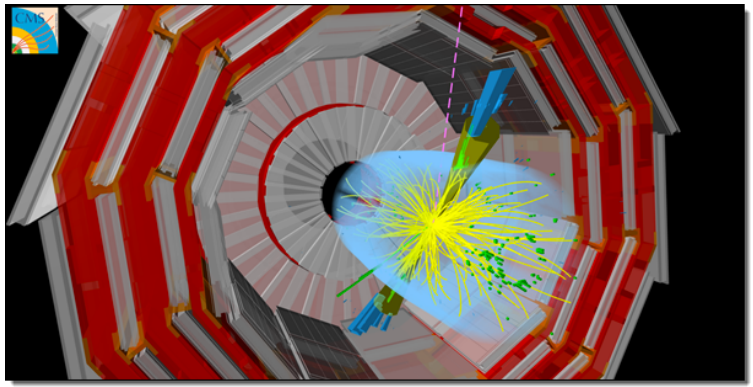
Lab: Deep Learning on AWS with Jupyter and MXNet

https://github.com/scicolabs/data-science-ml

Customer stories

High Energy Physics with Fermilab

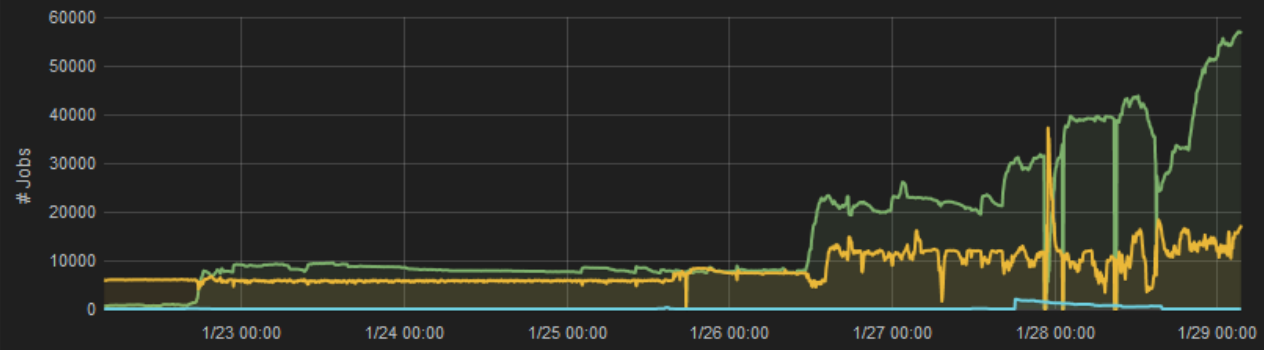
- Fermilab is one of the Tier 1 data centers for the CMS experiment (at CERN)
- Participated in finding the Higgs Boson to understand mass
- Launched the High Energy Physics Cloud Project in June, 2015
- Recently **added 58,000 cores** (or 4x increase in Fermilab capacity) to simulate 500 million events over 10 days
- AWS allowed FermiLab to burst capacity for large-scale data analysis, which on-prem systems were unable to do



Source: https://aws.amazon.com/blogs/aws/experiment-that-discovered-the-higgs-boson-uses-aws-to-probe-nature/

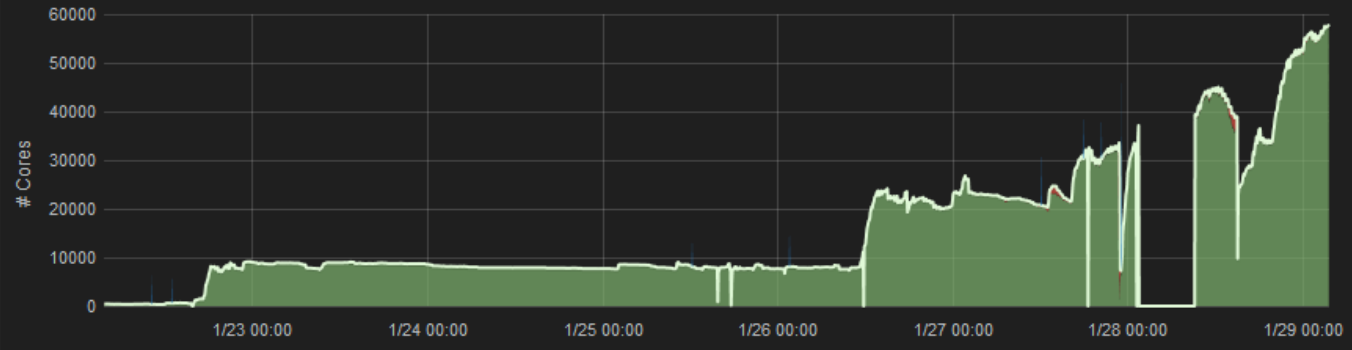
AWS VM Status
GCloud VM Status
HEP Cloud HTCondor Status
HEP Cloud Slots

Job Status



	min	max	avg	current
Running	0	56994	15861	56624
Idle (total)	0	37228	8082	17250
Idle (Fifebatch / nova)	0	2000	136	0

Slots Summary



	min	max	avg	current
01_cpu	0	57535	14214	56991
04_cpu	0	0	0	0
08_cpu	0	480	50	0
Unclaimed	0	32546	171	449
Total	0	57708	14435	57440
Idle	0	37454	46	0

https://fifemon.fnal.gov/hcf/dashboard/db/aws-vm-status-by-account

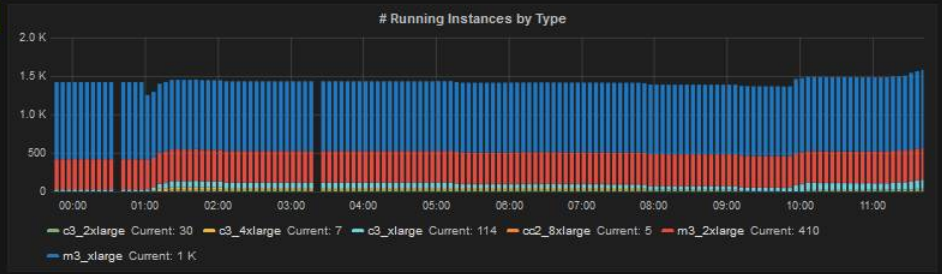
Most Visited Getting Started

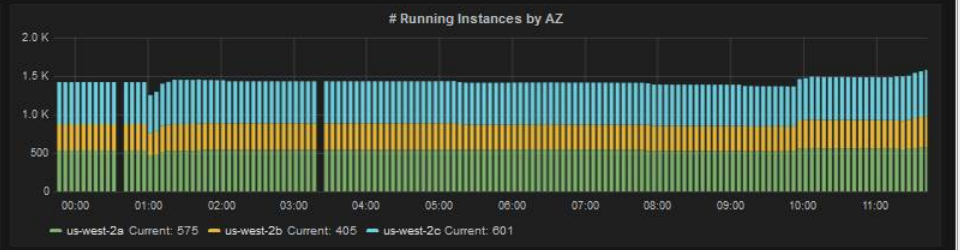
AWS VM Status by Account

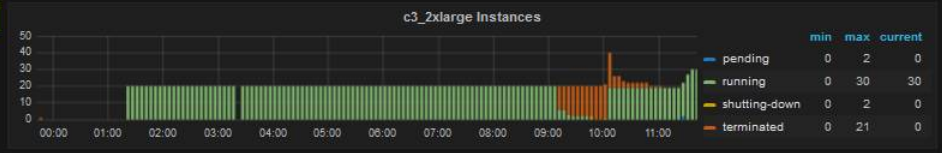
Zoom Out Last 12 hours

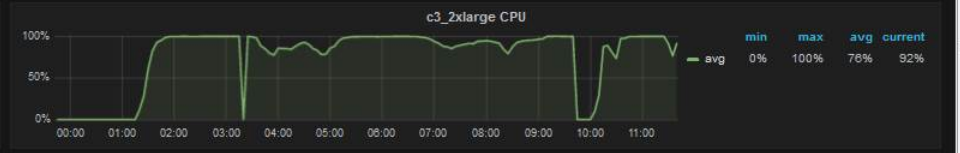
account: cms Region: us-west-2 VM Type: c3_2xlarge + c3_xlarge + m3_2xlarge

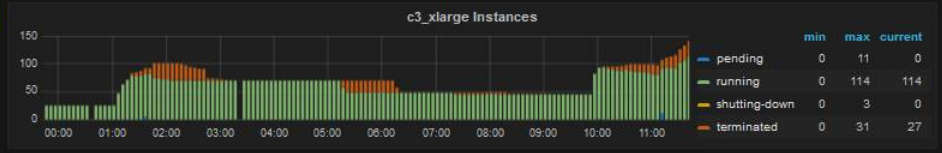
AWS VM Status | GCloud VM Status | HEP Cloud HTCondor Status | HEP Cloud Slots | HEP Cloud Summary

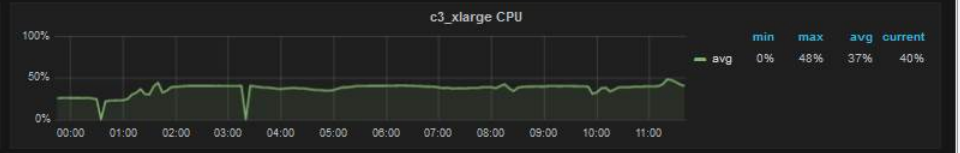


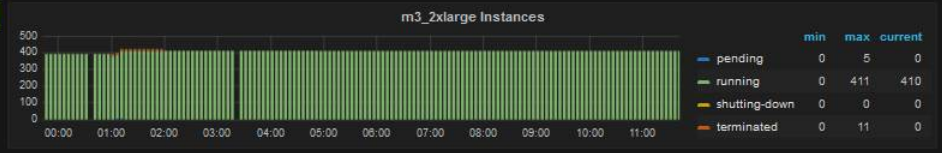


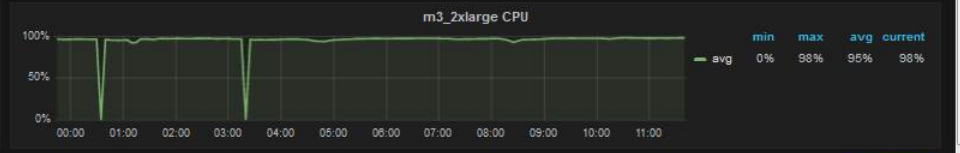












+ ADD ROW

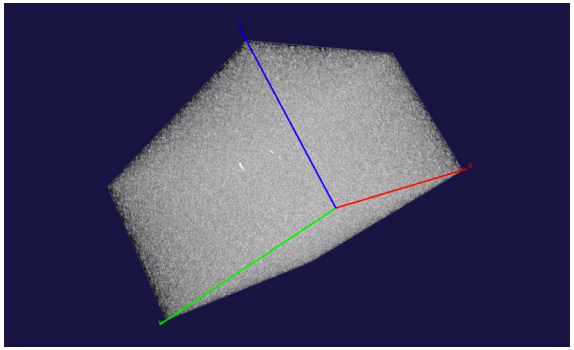
Australian Ocean Data Network (and IMOS)

- AODN hosts more than 160TB of ocean and marine data for research (including **IMOS**, **BOM**, **GA**, **CSIRO**)
- IMOS initially migrated to AWS for **scalability**, data **durability** and ability to **innovate** faster
- Thousands of publications have been based on IMOS data
- Enhanced Open Source software e.g. THREDDS / S3

Dr Peter Blain peter.blain@utas.edu.au

Looking Deep into our Universe with ICRAR

- International Center for Radio Astronomy Research (ICRAR) processes 10s to 100s TB of data from the VLA using AWS
- ICRAR has built an interactive ‘data cube’ from raw observation data which streams to Astronomer’s desktops
- Has recently supported the discovery of hydrogen emissions in a galaxy 5 billion light years away
- ICRAR is prototyping new data processing approaches on AWS



Source: https://aws.amazon.com/blogs/publicsector/looking-deep-into-our-universe-with-aws/

Real-time Flood Mapping with PetaBencana.id

Critical Web Services for Emergency Management

Custom interface for Emergency Control Room

Real time flood data entered into system via web interface and sourced from Twitter

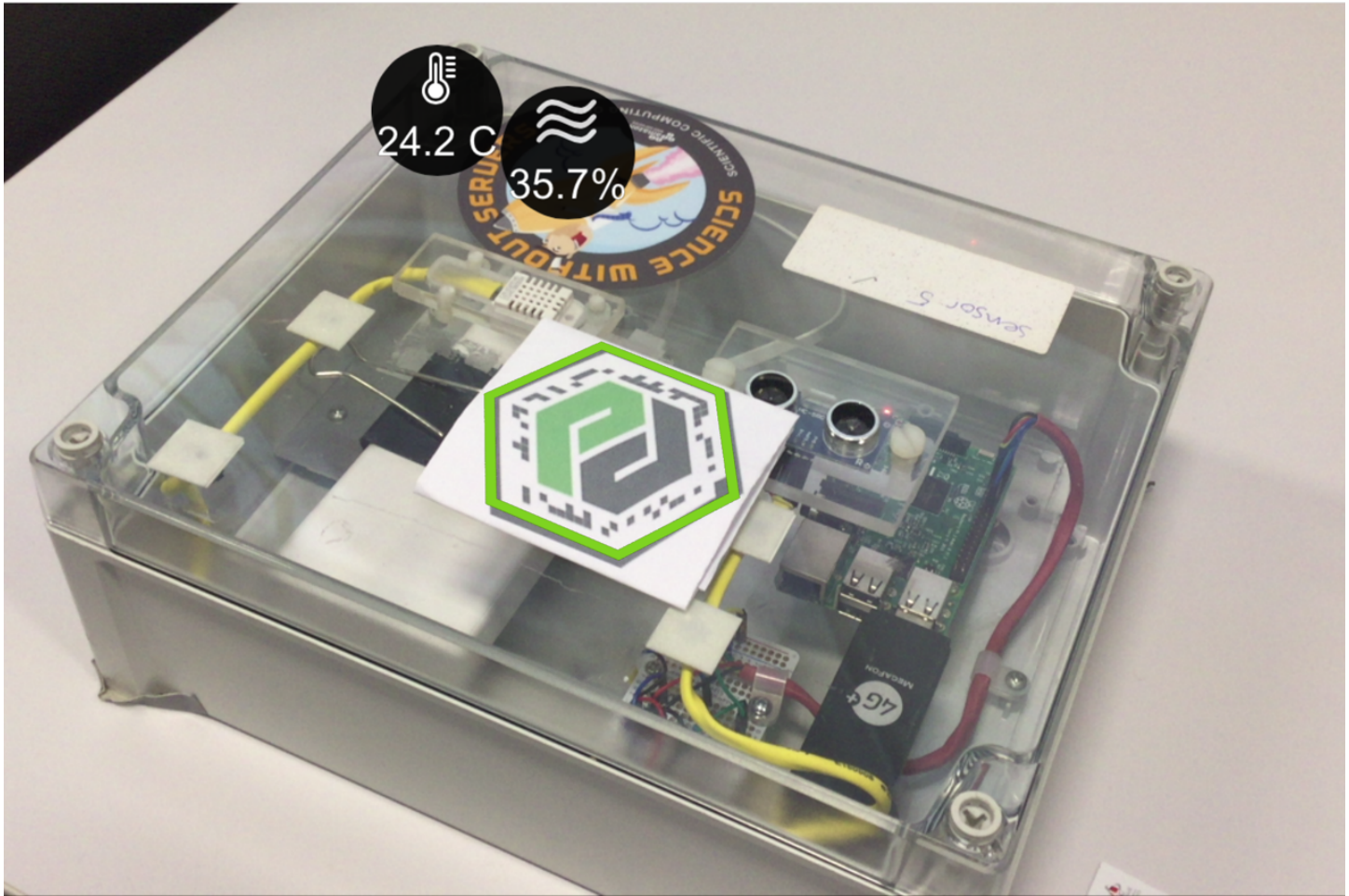
IoT water level sensing devices, to cheaply increase the monitoring across the waterway network in Jakarta

AWS Internet of Things

Prototype flood sensor

Idea: build IoT water level sensing devices, to cheaply increase the monitoring across the waterway network in Jakarta.
Use ultrasound sonar devices to measure distance.
Transmit data securely across the Internet to the AWS cloud and store the data in the database.

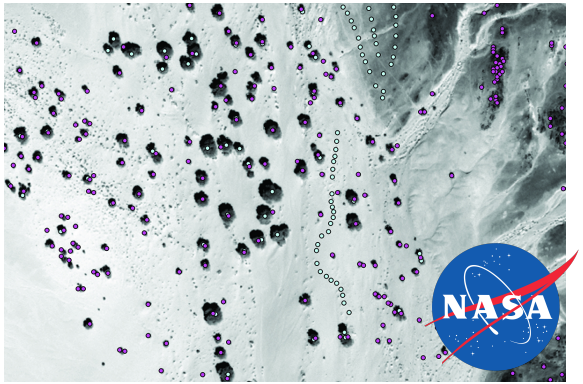
powered by amazon web services

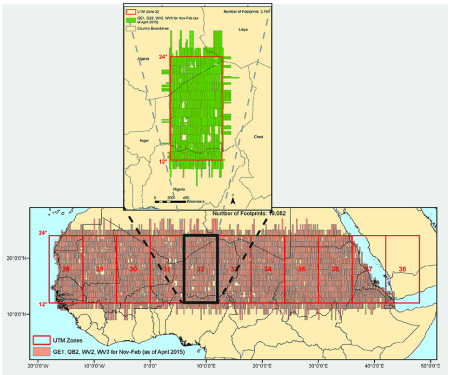

24.2 C
35.7%
SCIENCE WITHOUT
Sensor 5

NASA & Cycle Computing – 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 to less than 1 month from 10 months





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

CYCLECOMPUTING

AMPLab & RISELab (Algorithms, Machines, People)

- Collaborative 5-year effort between UC Berkeley, National Science Foundation, and industry partners (2012-2016) – AWS is founding partner
- Students and researchers AMPLab leveraged 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

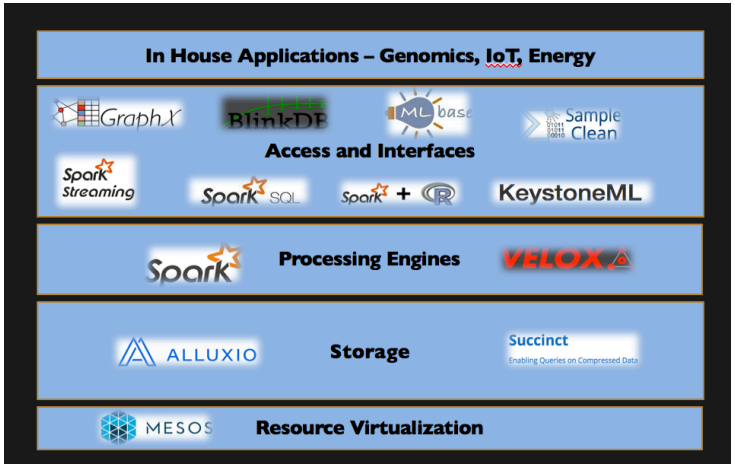
From batch data to advanced analytics

Algorithms
- Machine Learning, Statistical Methods
- Prediction, Business Intelligence

Machines
- Clusters and Clouds
- Warehouse Scale Computing

People
- Crowdsourcing, Human Computation
- Data Scientists, Analysts

Berkeley Data Analytics Stack



https://amplab.cs.berkeley.edu

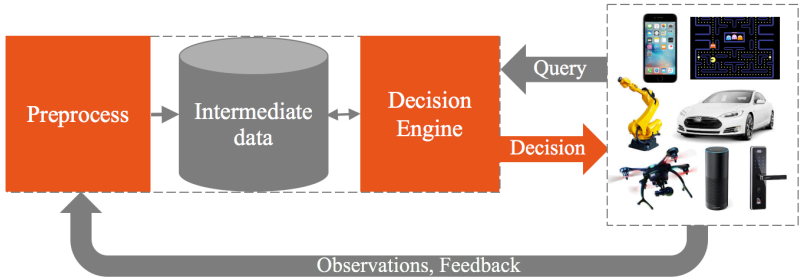
AMPLab & RISELab (Real-time Intelligent Secure Execution)

- Collaborative 5-year effort between UC Berkeley, National Science Foundation, and industry partners. (2017-2021) – AWS is founding partner

Data only as valuable as the decisions it enables

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

Typical decision system



From live data to real-time decisions

AMPLab & RISELab (Real-time Intelligent Secure Execution)

Challenges	RISE Lab
Automated decisions on live data are hard	Real-time, complex decisions that guarantee worst-case behavior on noisy and unforseen live data
Poor security: exploits are daily occurrences	Ensure privacy and integrity without impacting functionality
One-off solutions, expensive and slow to build	General platform: **Secure Real-time Decision Stack**

16

Some Proposed Research

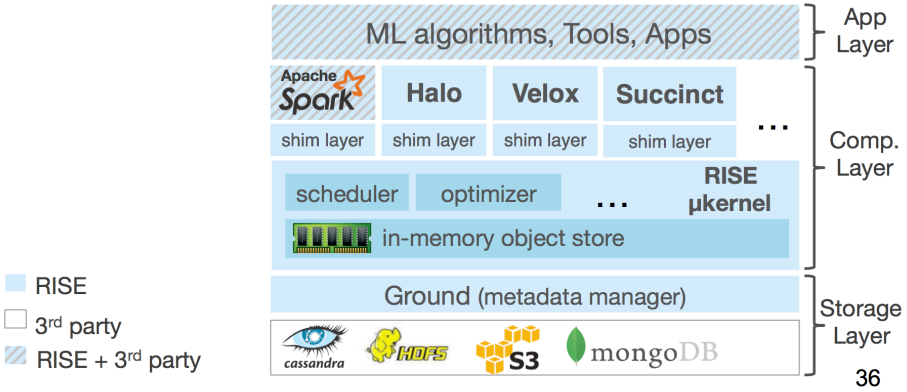
Secure Real-time Decisions Stack (SRDS)

- Open source platform to develop of RISE like apps
- Reinforcement Learning (RL) as one of key app patterns
- Secure from ground up

Learning control hierarchies: speedup learning, training

Shared learning: learn over confidential data

Preliminary software stack



36