



# Accelerating Life Sciences Research with HPC on AWS

**Patrick Combes,**  
Technical Leader, Healthcare and Life  
Sciences



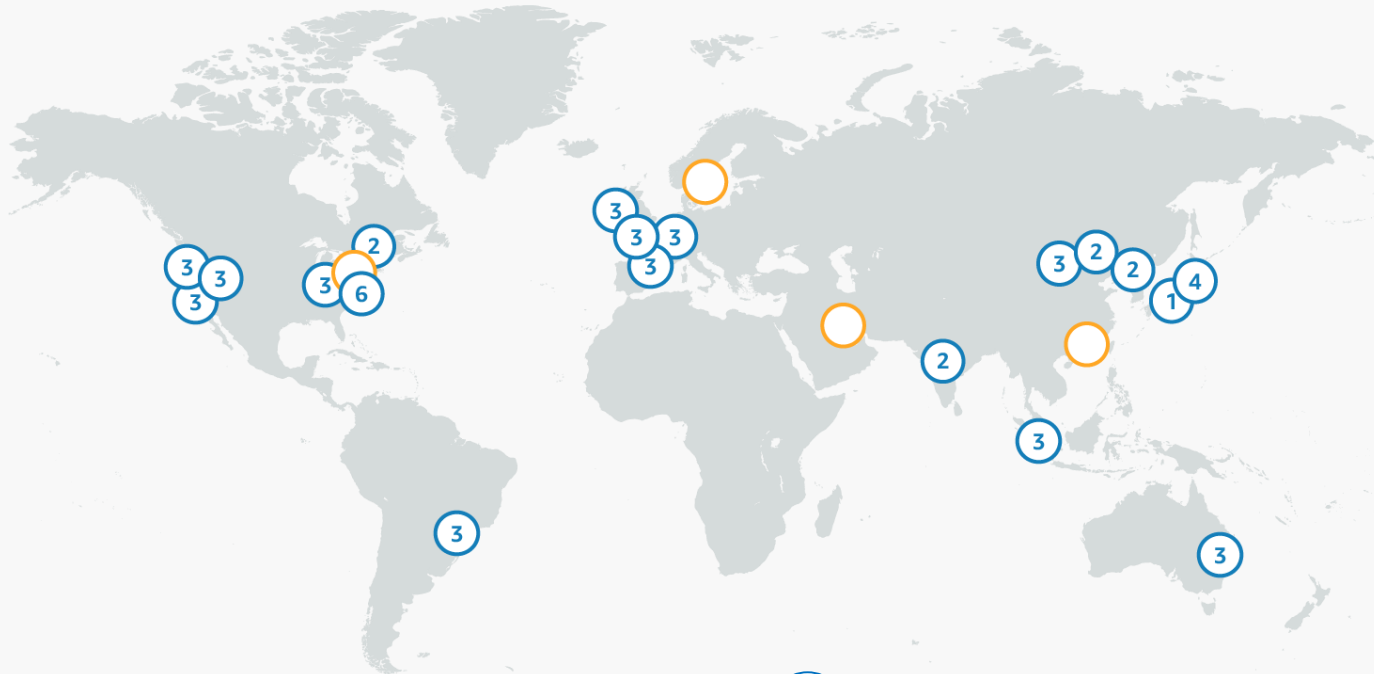
# Agenda

- Overview of AWS Infrastructure
- HPC Solution Components
- HPC Use Cases in Life Sciences & Common Tools
- Specific HPC Use Cases in Life Sciences & Customer Success Stories
- Best Practices
- Getting Started

# AWS Global Infrastructure

# AWS Global Infrastructure

18 Regions – 55 Availability Zones



The AWS Cloud spans **55 Availability Zones** within **18 geographic Regions** and 1 Local Region around the world,

Announced plans: 12 more Availability Zones and four more Regions in Bahrain, Hong Kong SAR, Sweden, and a second AWS GovCloud Region in the US.



# AWS Global Infrastructure

Regions

## Amazon Global Network

- Redundant 100GbE network
- Redundant private capacity between all Regions except China

Over **100** Global CloudFront PoPs

# High Performance Computing on AWS

# High Performance Computing on AWS

- **Innovate faster** with virtually unlimited infrastructure enabling scaling and agility not attainable on-premises
- **Optimize cost** with flexible resource selection and pay per use
- **Increase collaboration** with secure access to clusters around the world

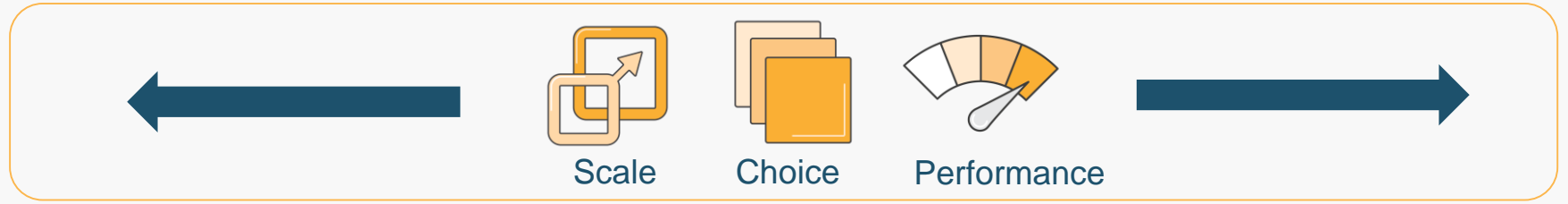


**Faster Time to Results**



**Better ROI**

# AWS Advantages for HPC Workload Types



**Tightly Coupled  
Parallel  
Computing**

**Loosely Coupled  
Parallel  
Computing**

**Accelerated  
Computing**

**Visualization and  
Interpretation**

**High Performance  
Data Storage and  
Analytics**



Skip the Queue



EC2 Spot  
Pricing



Early Access to  
Technology



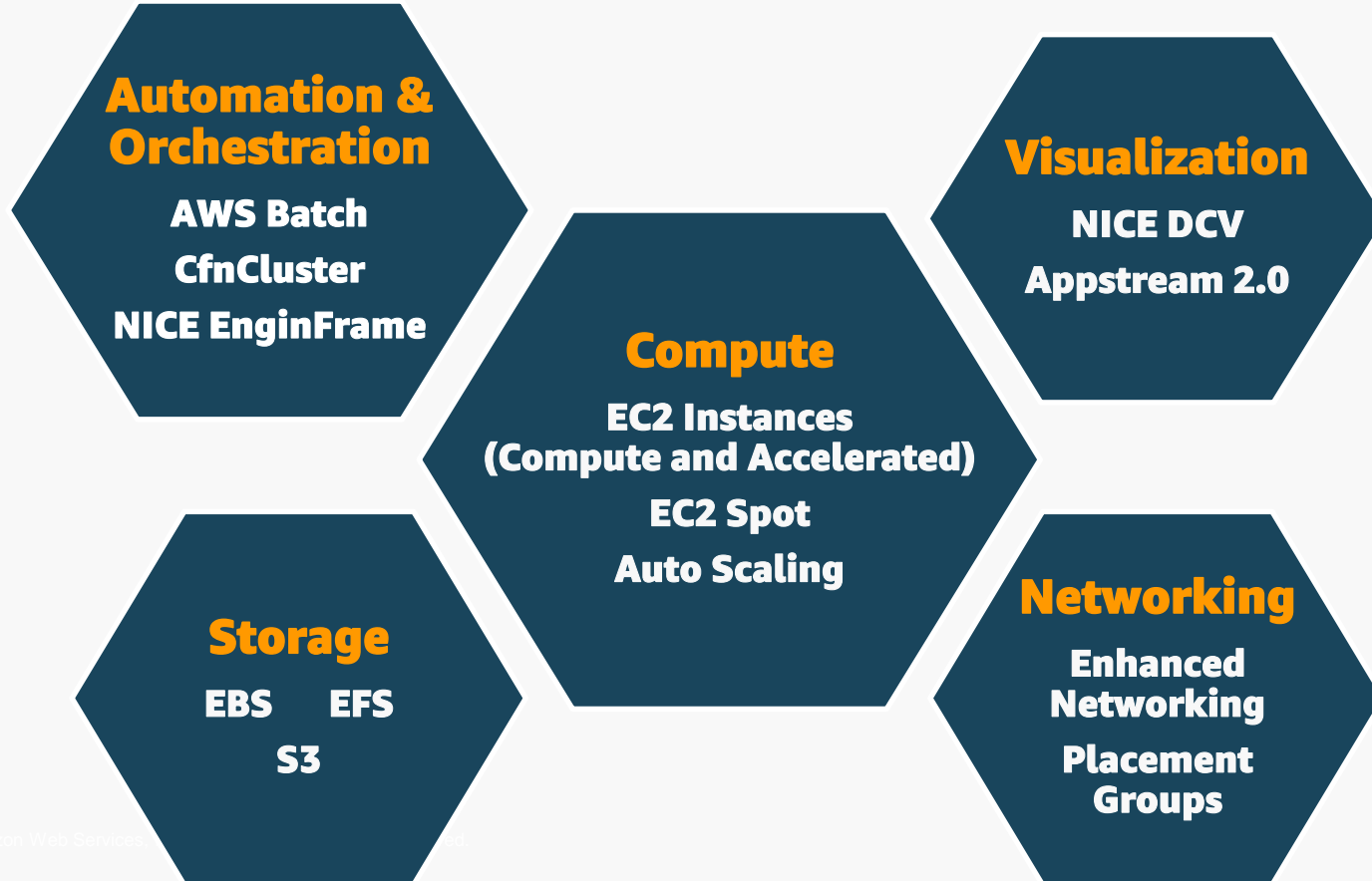
View results  
instantly



Derive unique  
insights with AI/ML

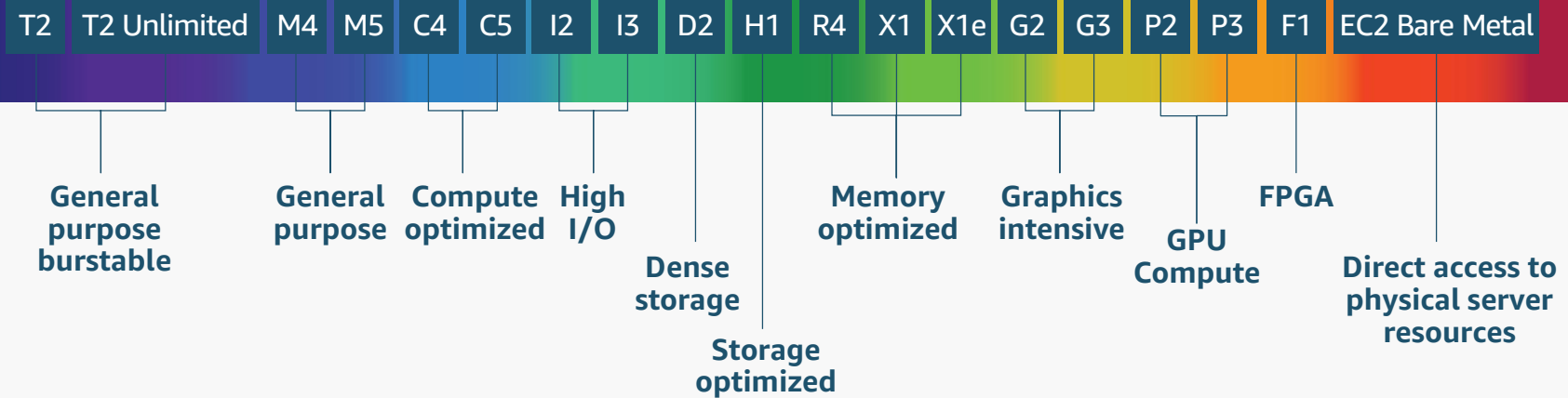


# AWS HPC Solution Components

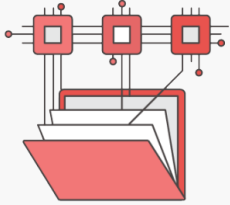


# Amazon EC2 Instances

Optimize the price/performance of your HPC Workloads with the widest range of compute instances

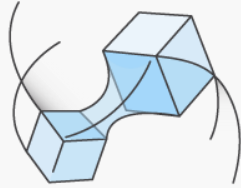


# AWS Storage is a Platform



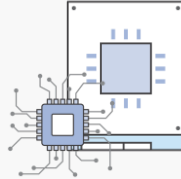
Amazon EFS

File



Amazon EBS

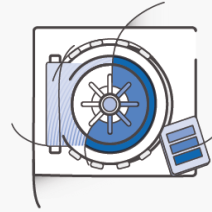
Block



Amazon EC2  
Instance Store



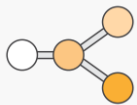
Amazon  
S3 / S3-IA



Amazon Glacier

Object

## Data Transfer



Internet/  
VPN



AWS Direct  
Connect



Amazon  
CloudFront



S3 Transfer  
Acceleration



ISV  
Connectors



Storage  
Gateway



AWS  
Snowball



Amazon  
Kinesis  
Firehose

# EC2 Purchasing Options

## On-Demand

Pay for compute capacity **by the second** with no long-term commitments

Spiky workloads, to define needs



## Reserved

Make a **1 or 3 Year commitment** and receive a significant discount off On-Demand prices

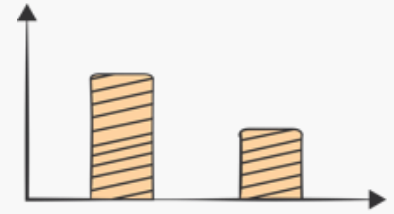
Committed, steady-state usage



## Spot

Spare EC2 capacity at savings of **up to 90%** off On-Demand prices


Fault-tolerant, dev/test, time-flexible, stateless workloads

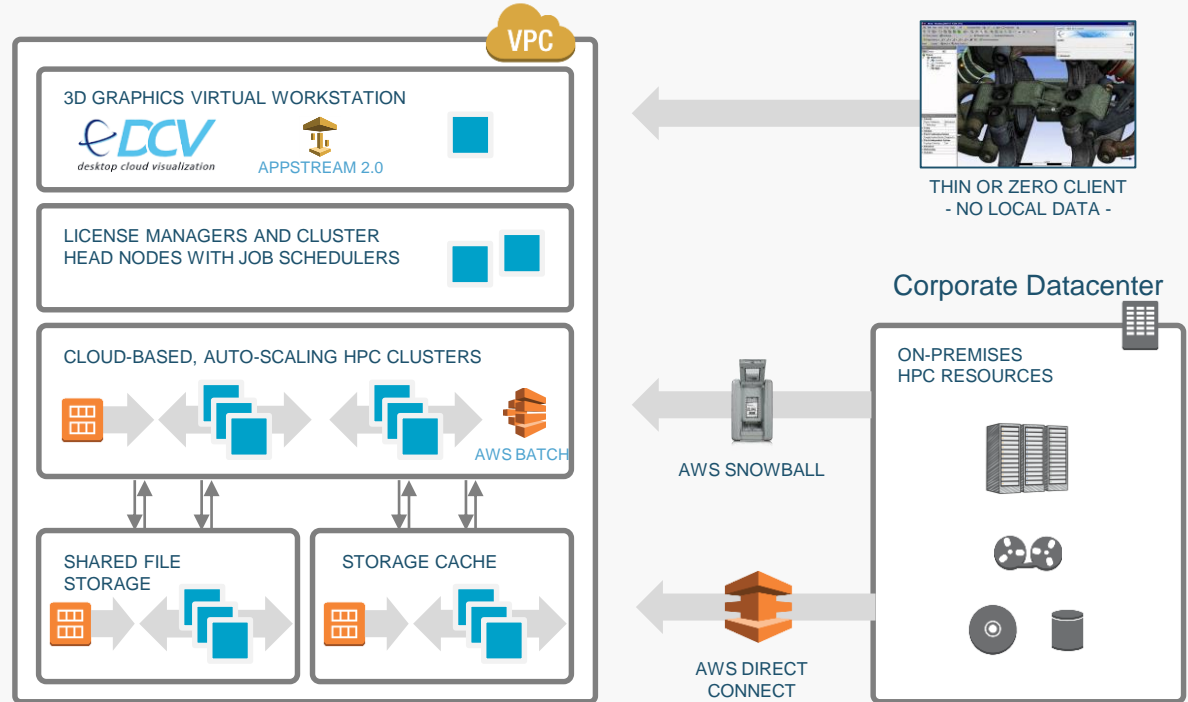


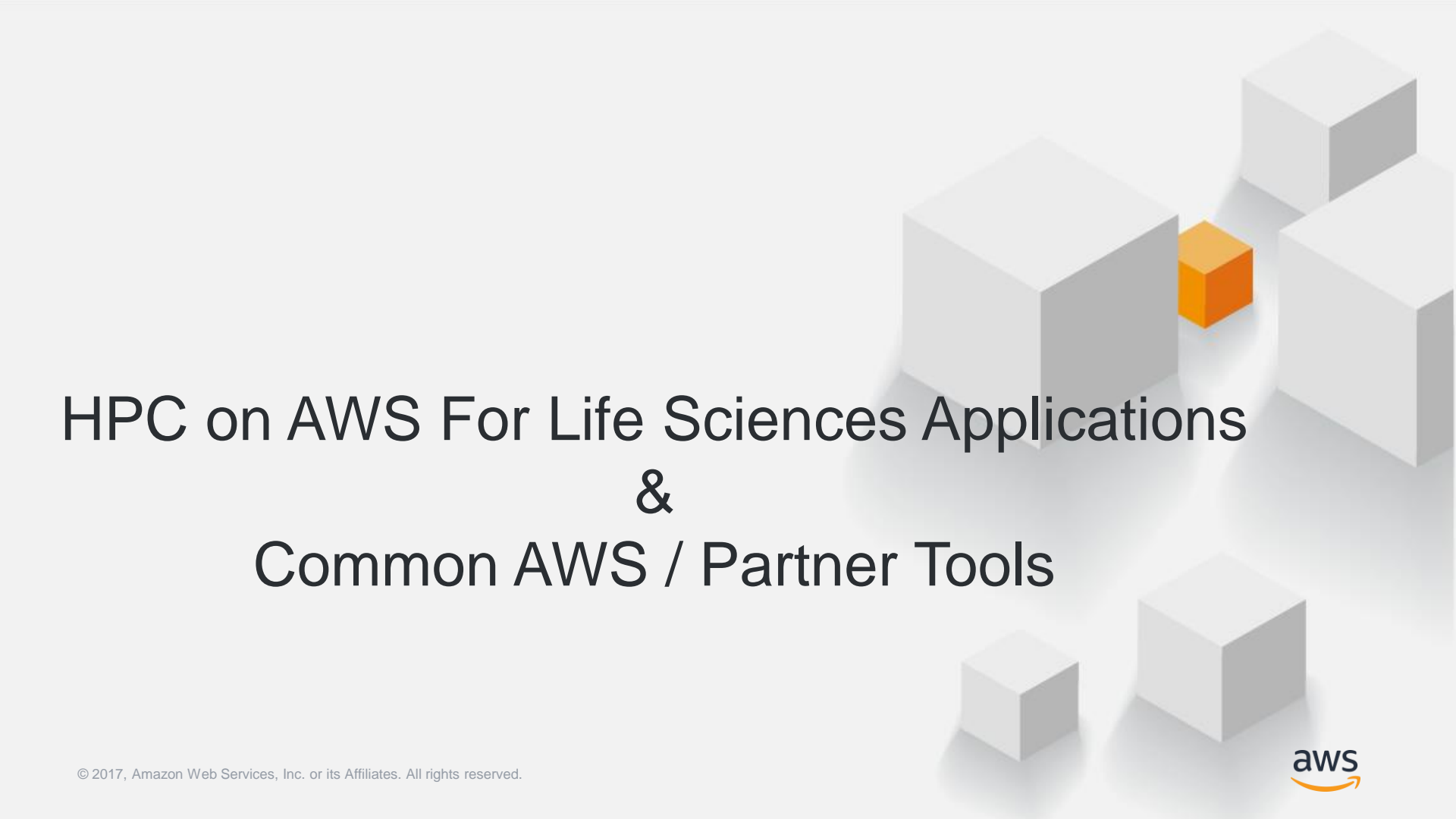
**Per Second Billing for EC2 Linux instances & EBS volumes**

# Deploying HPC on AWS

On AWS, secure and well-optimized HPC clusters can be automatically created, operated, and torn down in just minutes

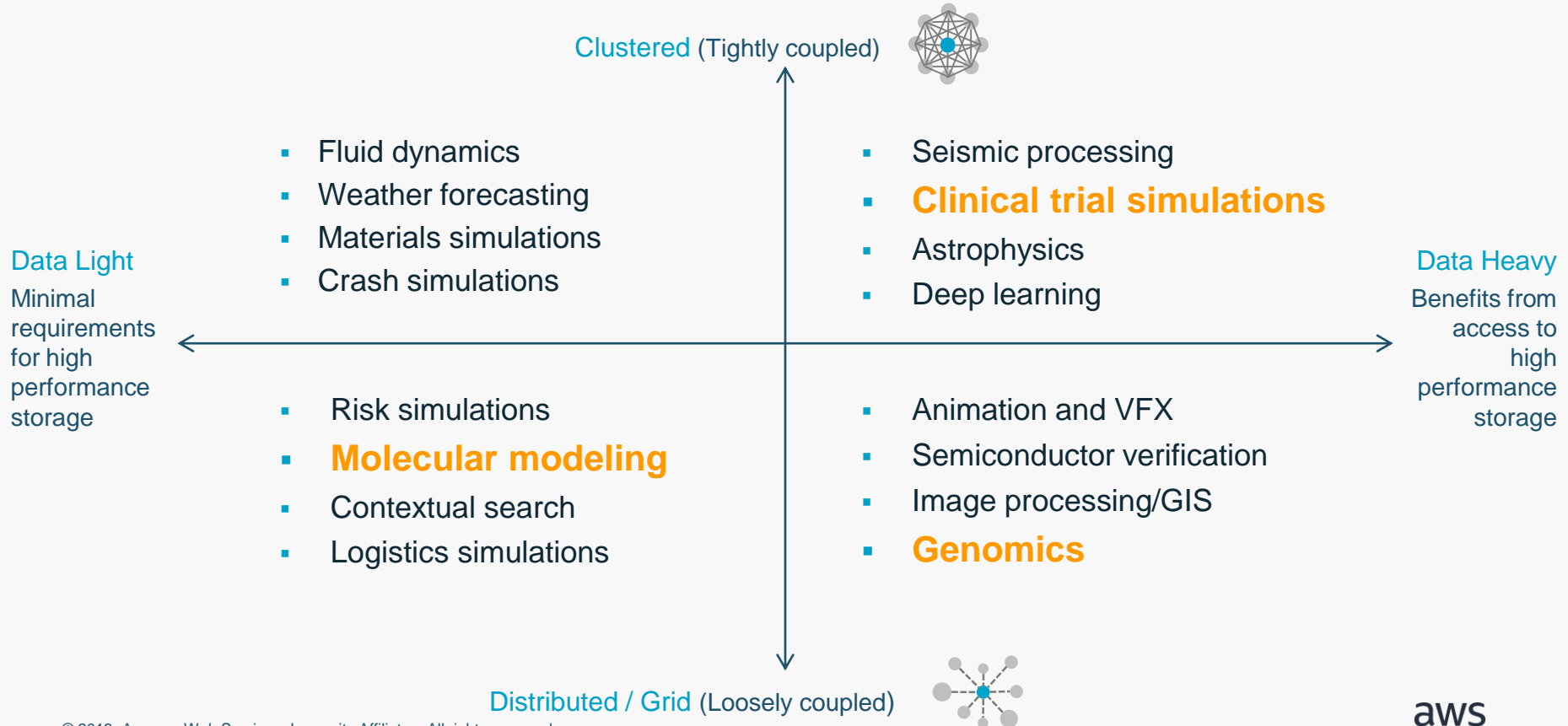
 Amazon S3 and Amazon Glacier





# HPC on AWS For Life Sciences Applications & Common AWS / Partner Tools

# Defining HPC – Example Use Cases



# AWS Benefits Directly Impact Life Sciences



## Accelerated Time to Insight

- Quickly derive actionable insights from research and development data inputs
- No large upfront investments in time, infrastructure, and money



## Scalability and Dynamic Resourcing

- Efficiently scale resources to meet shifting demands throughout the product lifecycle
- Encourages unrestricted scientific experimentation, product launches, and manufacturing runs



## Compliant and Secure Environment

- Streamlined and repeatable test environment
- Traceability helps organizations satisfy regulations and audits



## Global, Fault-Tolerant Infrastructure

- 55 AWS Availability Zones in 18 Regions worldwide means high availability
- Allows for seamless information-sharing between global stakeholders

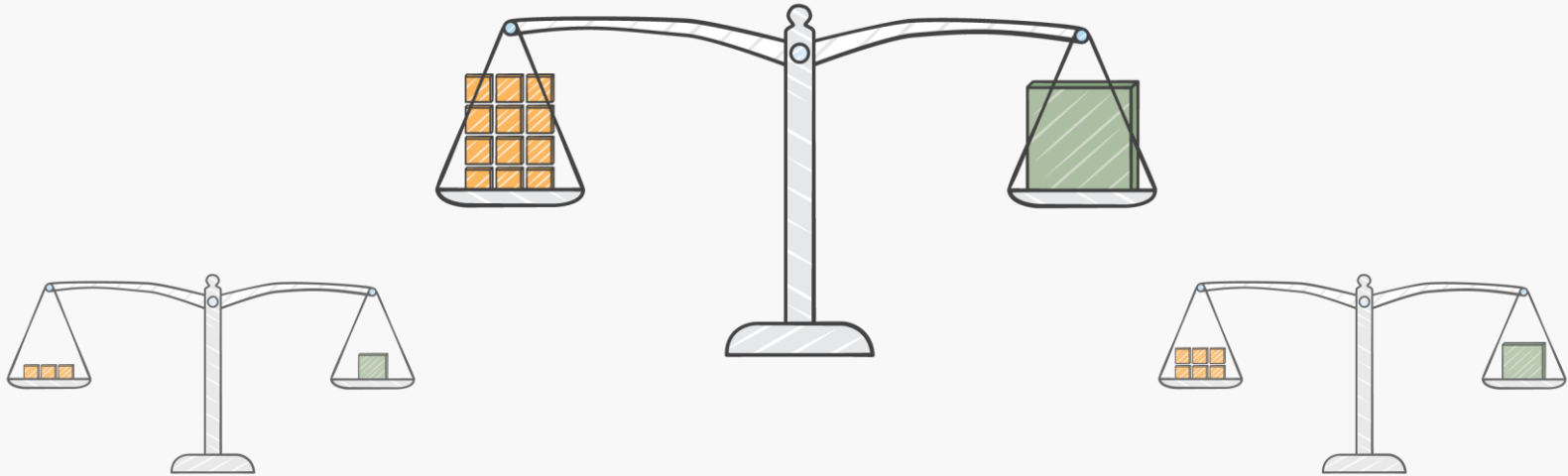


# Common AWS Tools for HPC Applications in Life Sciences

- Multiple Clusters
- CfnCluster
- AWS Batch
- Step Functions

# Deploy Multiple HPC Clusters

Running multiple clusters at the same time, and tuned for each workload



# HPC Automation with CfnCluster

## CfnCluster

CfnCluster is a tool used to build and manage High Performance Computing (HPC) clusters on AWS.

Once created, you can log into your cluster via the master node where you will have access to standard HPC tools such as schedulers, shared storage, and an MPI environment.



[Getting Started](#)



[CLI Reference](#)



[GitHub Project](#)



[Community Forum](#)

- CfnCluster simplifies deployment of HPC in the cloud, including integrating with popular HPC schedulers
- Built on AWS CloudFormation, easy to modify to meet specific application or project requirements

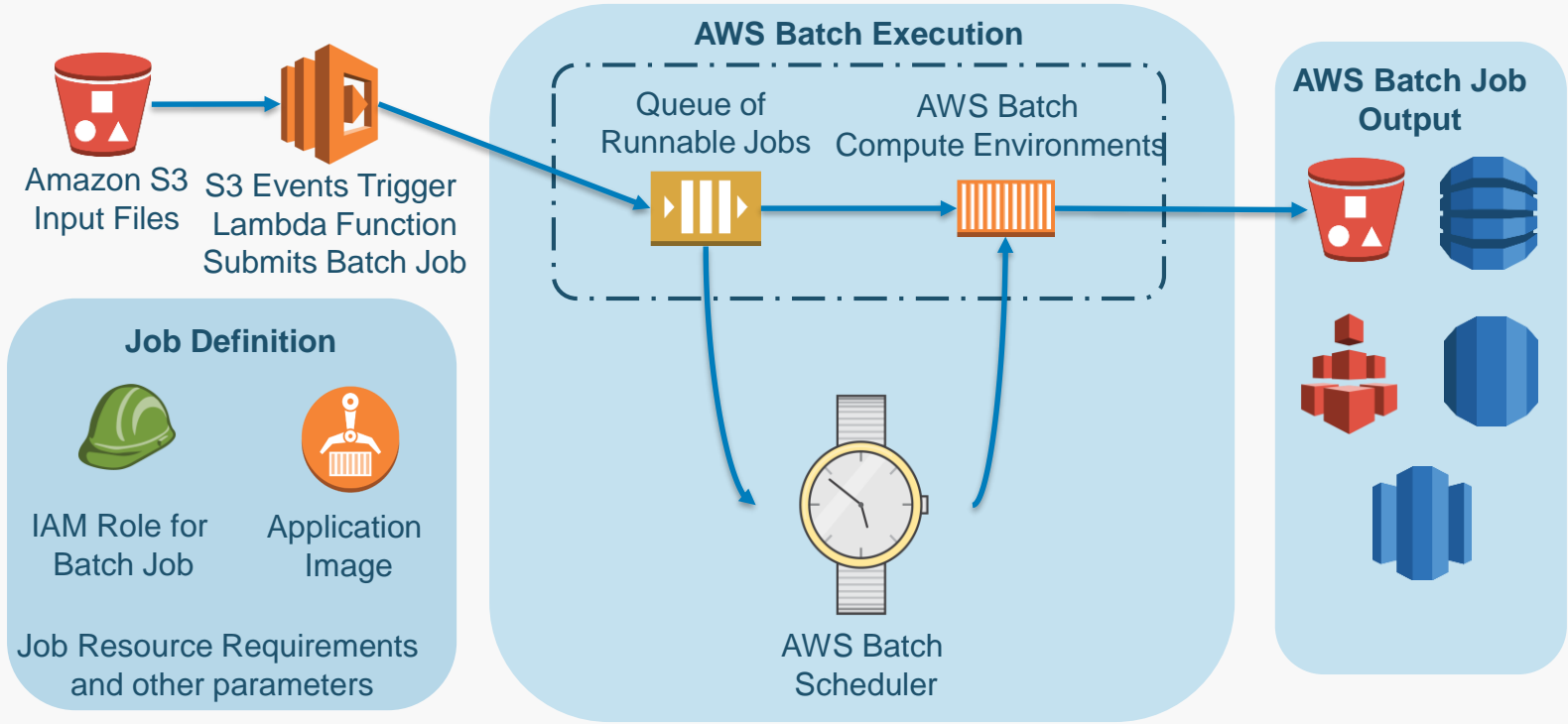
# AWS Batch

- AWS Batch dynamically provisions resources
- Plans, schedules, and executes
- No batch software to install



Focus on your **applications and results!**

# Example: AWS Batch Job Architecture



# AWS Step Functions...



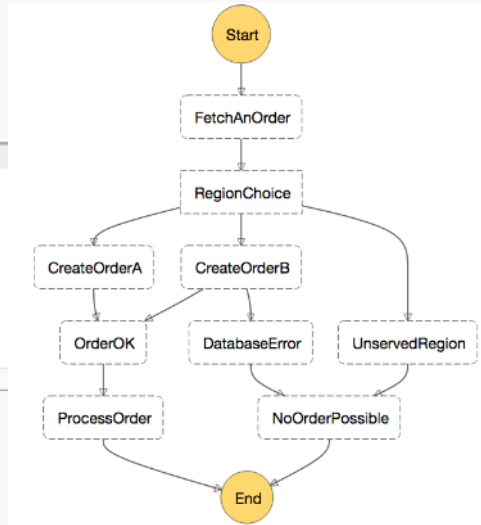
...makes it easy to coordinate the components of distributed applications using visual workflows

# Application Lifecycle in AWS Step Functions

## Define in JSON

```
Code
1 {
2   "Comment": "An AWL example using a choice state.",
3   "StartAt": "FirstState",
4   "States": {
5     "FirstState": {
6       "Type": "Task",
7       "Resource": "arn:aws:lambda:REGION:ACCOUNT_ID:function:FUNCTION_NAME",
8       "Next": "ChoiceState"
9     },
10    "ChoiceState": {
11      "Type": "Choice",
12      "Choices": [
13        {
14          "Variable": "$?.OrderType",
15          "Default": "OrderTypeA",
16          "Next": "CreateOrderA"
17        },
18        {
19          "Variable": "$?.OrderType",
20          "Default": "OrderTypeB",
21          "Next": "CreateOrderB"
22        }
23      ]
24    }
25  }
26 }
```

## Visualize in the console



## Monitor executions

Dashboard > Orders > New\_Order Execution Arn: aws:states:us-central-1:40241959445:execution:Orders\_New\_Order

New\_Order ✓

Graph Code

Execution Status: **Succeeded**

State Machine Arn: [arn:aws:states:us-central-1:40241959445:stateMachine:Orders](#)

Execution ID: [arn:aws:states:us-central-1:40241959445:execution:Orders/New\\_Order](#)

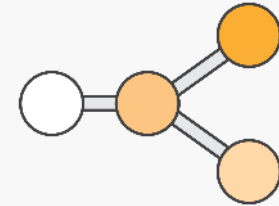
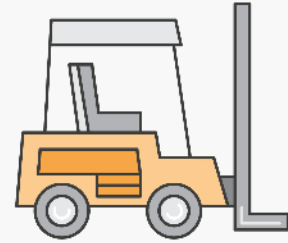
Started: Nov 20, 2016 9:55:28 AM

Completed: Nov 20, 2016 9:58:32 AM

ID	Type	Timestamp
1	ExecutionStarted	Nov 20, 2016 9:55:28 AM
2	TaskStateEntered	Nov 20, 2016 9:58:32 AM
3	LambdaFunctionScheduled	Nov 20, 2016 9:58:28 AM

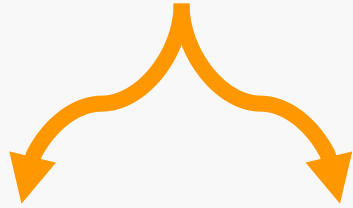
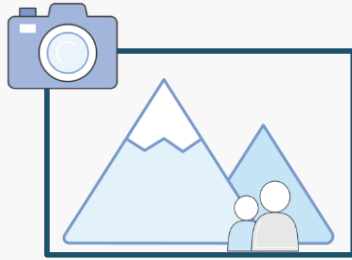
# Seven State Types

<b>Task</b>	<b>A single unit of work</b>
Choice	Adds branching logic
Parallel	Fork and join the data across tasks
Wait	Delay for a specified time
Fail	Stops an execution and marks it as a failure
Succeed	Stops an execution successfully
Pass	Passes its input to its output

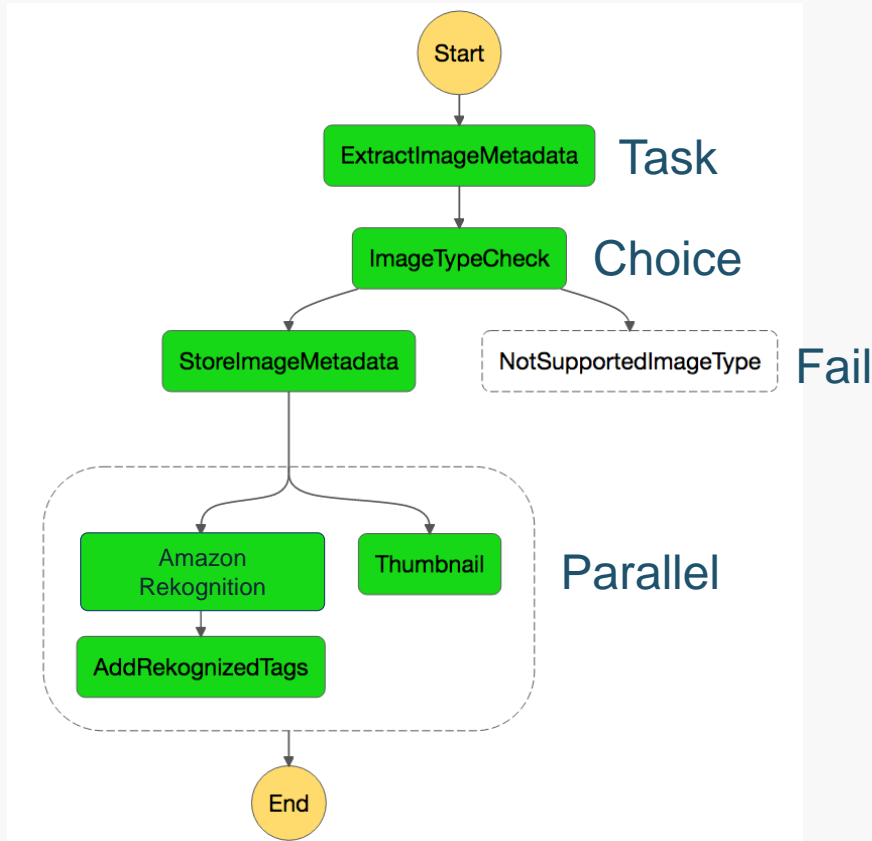




# Build Visual Workflows Using State Types



Mountains  
People  
Snow



# Partner Tools : Alces Flight



# HCLS Specific Applications and Tools

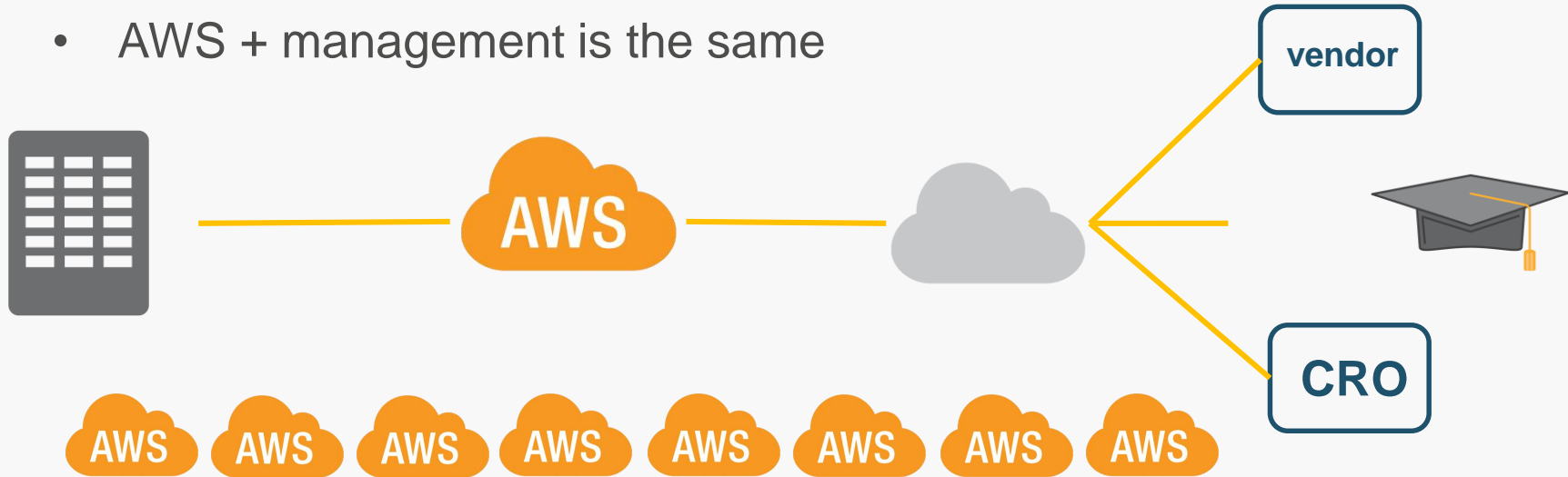
## Customer Success Stories

# Collaborative Research Environments

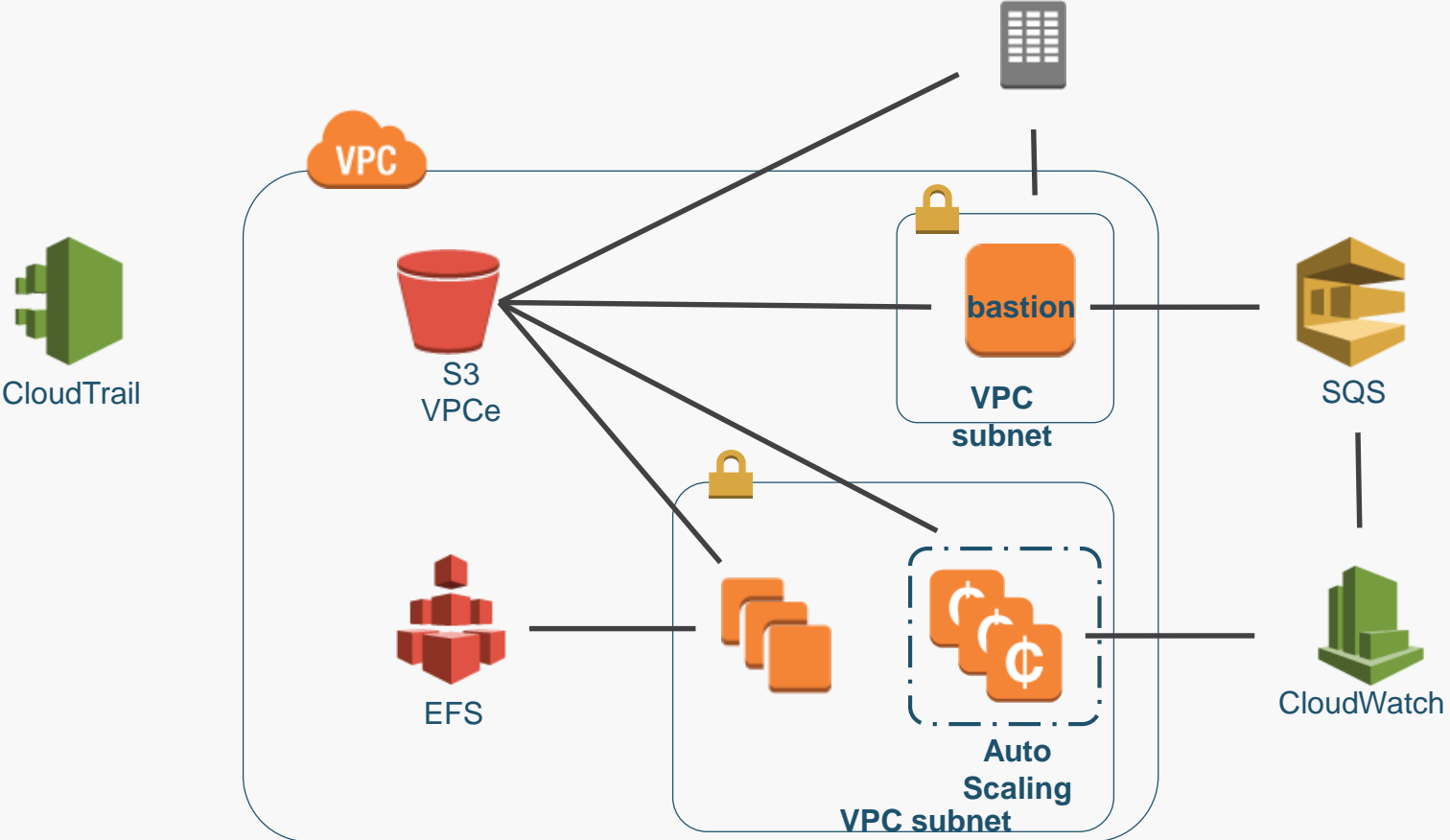
# Collaboration Structure

## Many Collaborations

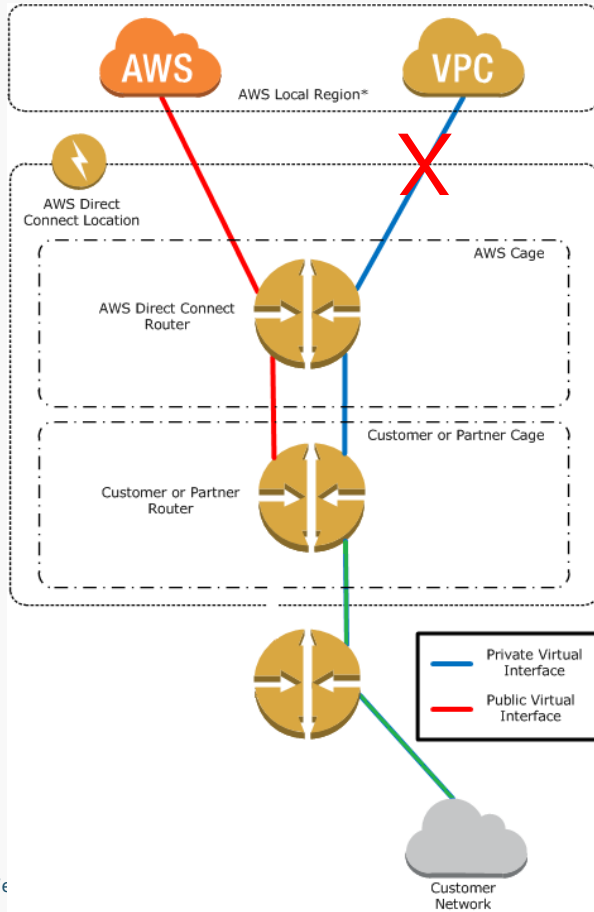
- Two collaborations models
- Multi-AWS account
- AWS + management is the same



# Example HPC Collaboration Architecture



# Connectivity



- Multi-account model + VPC
- Connectivity options
- Big decisions factors

# HPC Clusters in Healthcare & Life Sciences

## HPC on AWS for Cancer Drug Research

### The Challenge

- Slower time to results due to wait times and longer times to run jobs on fixed configurations available
- Hard to collaborate with external entities due to security and compliance issues
- Inability to scale beyond the fixed number of cores that were available on premises

### The Solution

- The company runs many HPC workloads on hundreds of Amazon EC2 instances and uses Amazon S3 and Amazon Glacier to store hundreds of terabytes of genomic data
- Using Amazon VPC, AWS Access and Identity Management, AWS Direct Connect to collaborate securely

### The Result

- HPC job time reduced to hours instead of weeks
- More parallel work being achieved leading to increased productivity



“By spinning up a few hundred nodes on AWS and getting results in less than a day, our scientific researchers have a lot more freedom to ask questions that weren’t even possible before. The speed is important, but equally important is the additional intellectual curiosity this enables for researchers”

Lance Smith  
Associate Director of IT,  
Celgene



# Celgene Uses AWS to Speed Cancer Research



Using AWS, a single scientist can launch hundreds of compute nodes. That's a capability we just didn't have before.

**Lance Smith**  
Associate Director of IT



Celgene is a biotechnology firm that creates drugs that fight cancer and other diseases and disorders.

- Wanted to improve its high-performance computing (HPC) capabilities
- Needed to enable collaboration between its own researchers and academic research labs
- Reduces HPC computational jobs from weeks to less than one day
- Enables secure collaboration between internal and external researchers
- Gives each scientist the ability to launch hundreds of compute nodes

# Baylor: Seamless Global Collaboration

“ “ The AWS Cloud enables swift collaboration even with hundreds of terabytes of data

**Dr. Narayanan Veeraraghavan**  
Lead Programmer Scientist  
Genome Sequencing Center, Baylor  
College of Medicine



- CHARGE Project required global collaboration of 200 scientists at 5 institutions
- 20 genomic sequences generating a PB a month

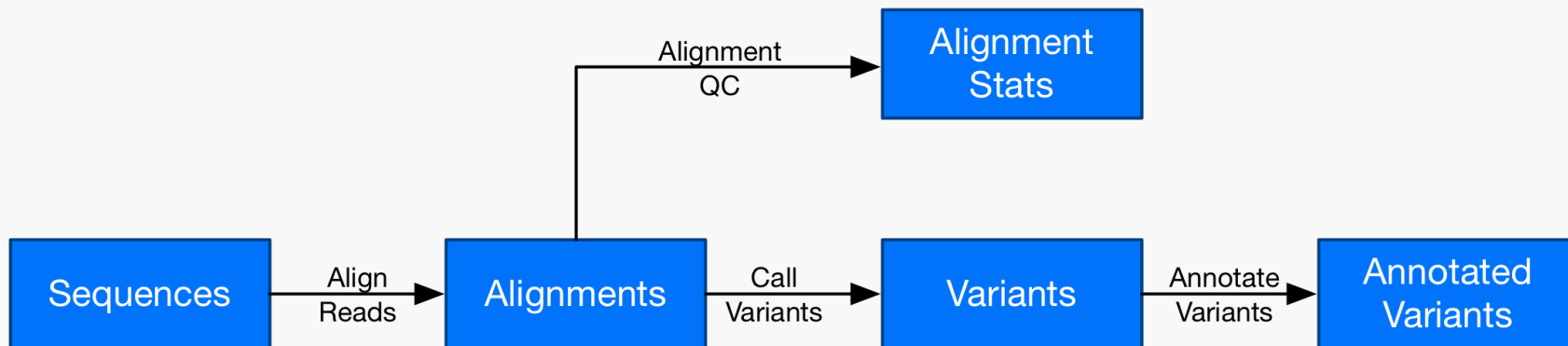


**When extended to the AWS Cloud, first analysis completed 5x faster vs. on premises**

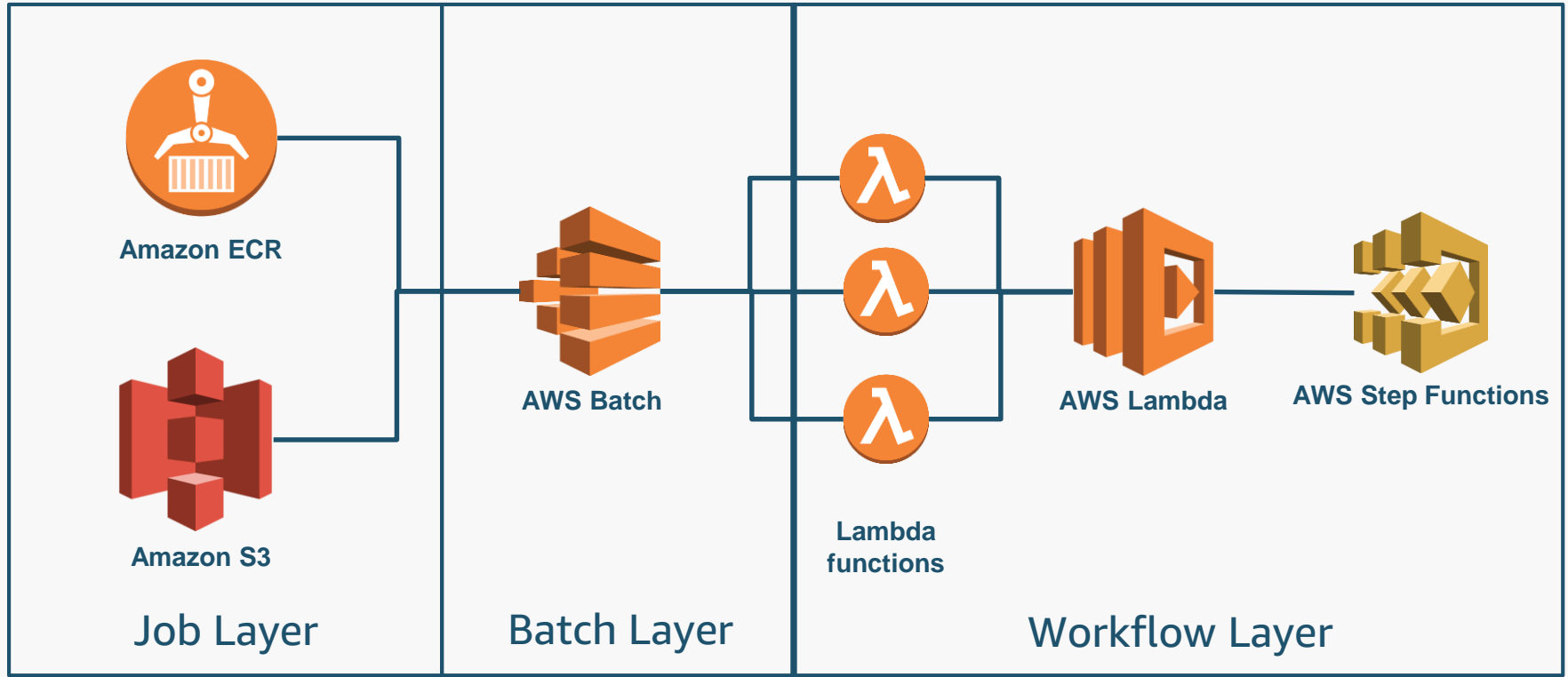
# Genomics

# Genomics Data Processing

## Typical workflow in genomics analysis



# A Reference Architecture for Genomics



# U.C. Berkeley Researchers Uses AWS to Look At More Accurate Data, Faster



We're able to solve real world problems because AWS gives us access to real world resources

**Michael Franklin**  
Director, AMP Lab



- The AMP Lab is a multidisciplinary research effort at the University of California, Berkeley aimed at building scalable machine learning and data analysis technology
- Needed to be able to scale compute resources quickly to analyze algorithms that are used in genomics work
- Researchers are able to easily scale Amazon EC2 instances simultaneously to process genome data faster and more cost effectively

# Genomics Processing on FPGA

## Children's Hospital of Philadelphia and Edico Genome Achieve Fastest-Ever Analysis of 1,000 Genomes



Orlando, Fla., Oct 19, 2018 – The Children's Hospital of Philadelphia (CHOP) and Edico Genome today set a new scientific world standard in rapidly processing whole human genomes into data files usable for researchers aiming to bring precision medicine into mainstream clinical practice. Utilizing Edico Genome's DRAGEN™ Genome Pipeline, deployed on 1,000 Amazon EC2 F1 instances on the Amazon Web Services (AWS) Cloud, **1,000 pediatric genomes were processed in two hours and 25 minutes.**



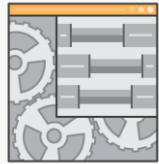
# FPGA Acceleration Using F1

## Amazon FPGA Image (AFI)

An F1 instance can have any number of AFIs

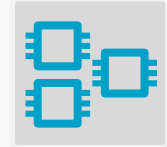
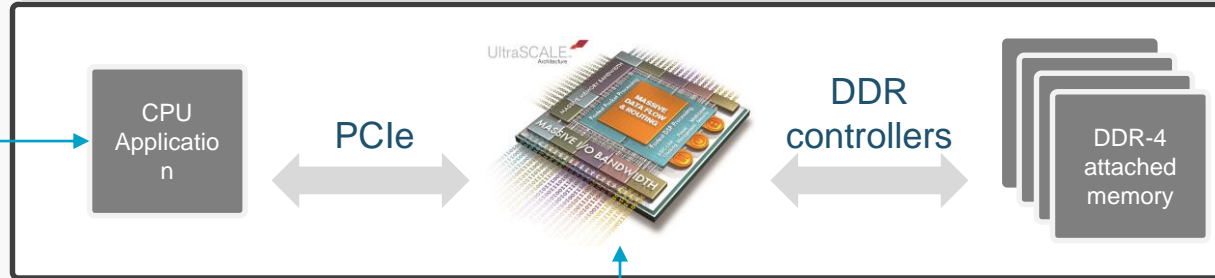
An AFI can be loaded into the FPGA in seconds

## Amazon Machine Image (AMI)



EC2  
F1

Launch instance and load AFI





# Clinical Trial Simulations

# Bristol-Myers Squibb Runs Clinical Trial Simulations Faster with AWS

At Bristol Myer-Squibb, Compute-intensive clinical trial simulations that previously took 60 hours could now be finished in only 1.2 hours on the AWS Cloud. They were able to reduce the number of subjects from 60 to 40 and the number of blood samples per subject from 12 to 5



**Bristol-Myers Squibb**

- Bristol-Myers Squibb needed a cost-effective and secure cloud solution to host research data for scientists that would be used in conjunction with on-premises systems.
- BMS can take advantage of on-demand capacity to run clinical trial simulations 98 percent faster than in its previous environment.

Bristol-Myers Squibb (BMS) is a global biopharmaceutical company with a mission to discover, develop, and deliver innovative medicines that help patients prevail over serious diseases.

# Medidata Uses AWS to Provide Physicians with Potentially Life-Saving Access to Patient Information



The rate of innovation in AWS is brilliant – it allows us to keep pushing forward-thinking solutions to our customers.

**Michelle Marlborough**  
VP of Product Strategy

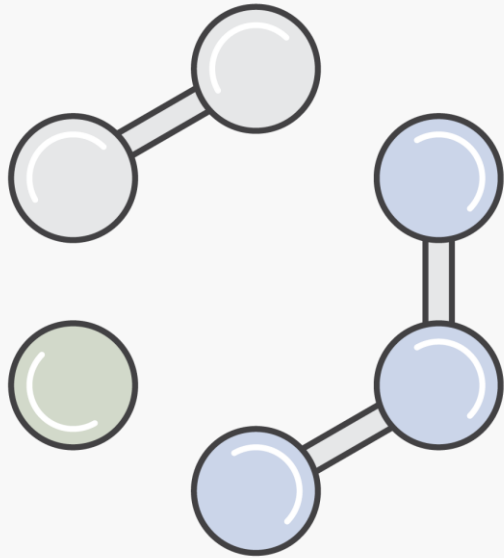


- The high availability, scalability, and storage capability offered by AWS help Medidata provide critical patient information to physicians and create new and innovative medical solutions

Medidata provides software that helps physicians deliver better, faster, and safer clinical trials

# Molecular Modeling & Computational Chemistry

# Computational Chemistry



- Calculate the properties of molecules and solids
- Property space varies over the HPC calculation space from massively parallel to tightly coupled
- Many industry standard partner solutions available on AWS

# Schrödinger uses AWS to Run Simulations Faster



This performance increase, coupled with the ability to quickly scale in response to new compound ideas, gives our customers the ability to bring lifesaving drugs to market more quickly

Robert Abel,  
Senior Vice President of Science

**SCHRÖDINGER**



- Amazon EC2 P3 instances with high performance GPUs allowed Schrodinger to perform four times as many simulations in a day as they could with P2 instances.

Schrödinger's mission is to improve human health and quality of life by developing advanced computational methods that transform the way scientists design therapeutics and materials

# AWS Helps Pfizer Focus on Large Scale Data Analysis

“ AWS enables Pfizer’s WRD to explore specific difficult or deep scientific questions in timely, scalable manner and helps Pfizer make better decisions more quickly.

**Dr. Michael Miller**  
Head of HPC for R&D

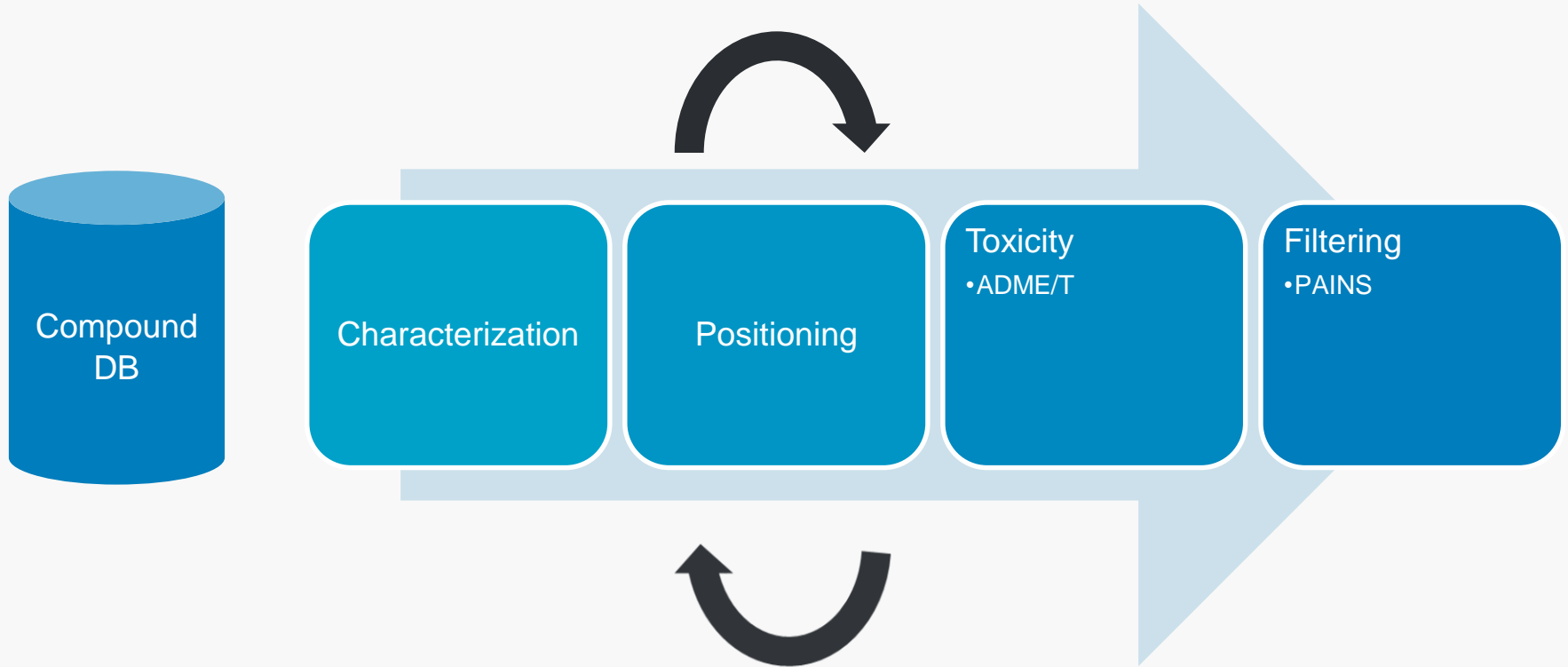


- Pfizer’s HPC software and systems for worldwide research and development (WRD) support large scale data analysis, research projects, clinical analytics and modeling
- The company chose AWS because it offered an additional level of security and integrated easily into the already present infrastructure
- Pfizer saved money with AWS by not having to invest in additional hardware and software

# High Throughput Screening



# Virtual High-Throughput Screening



# Novartis: Acceleration of Pre-Clinical R&D

“ “  
We completed the equivalent of thirty-nine years of computational chemistry in just under 9 hours for a cost of around \$4200.

Steve Litster

Global Head of Scientific Computing, Novartis



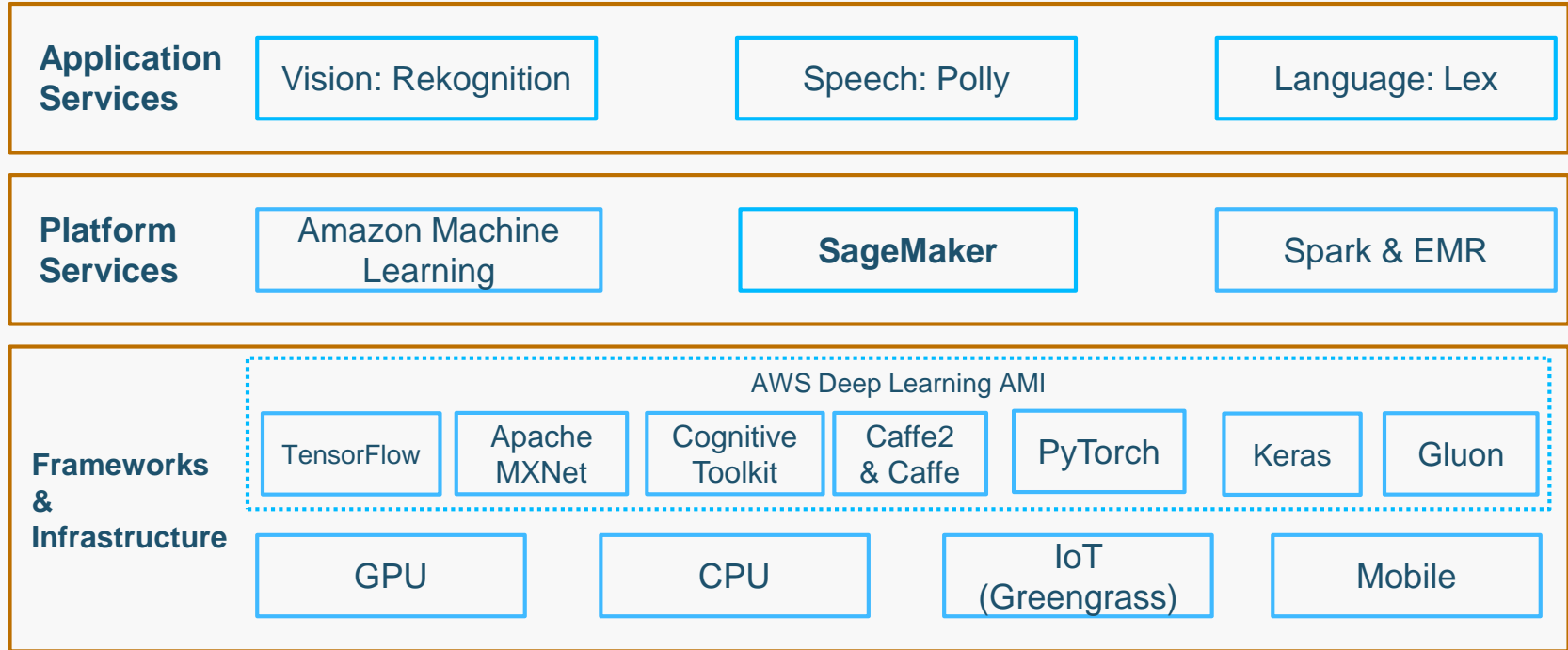
- Existing infrastructure to screen 10 million compounds in a computational model not available
- New infrastructure would have cost approximately \$40 million to build



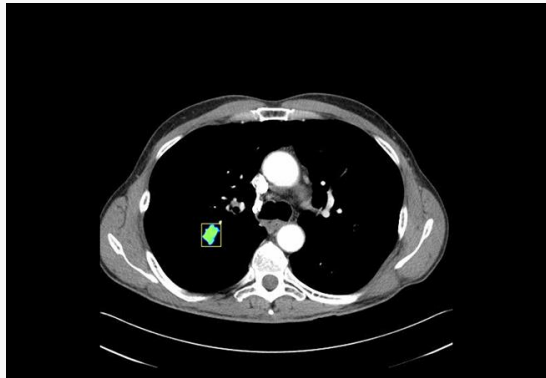
**Novartis used AWS for HPC computational chemistry**

# Deep Learning

# Wide and Deep ML/DL Stack

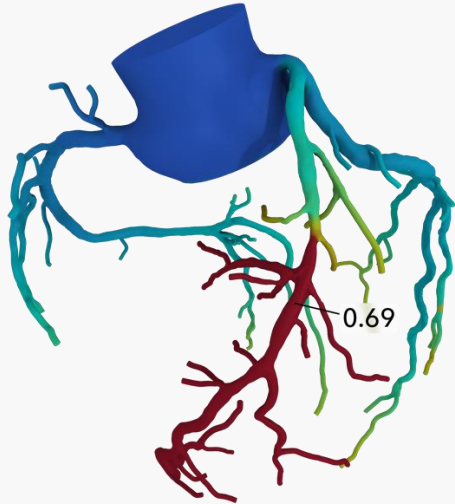


# Early Cancer Detection with Deep Learning on AWS



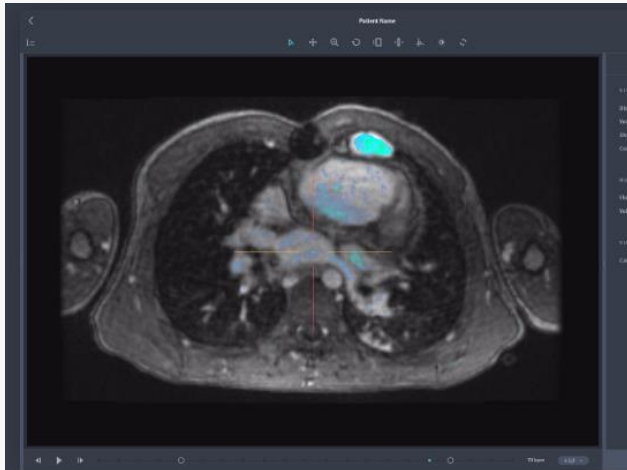
- Matrix Analytics uses deep learning in its LungDirect solution to track disease progression for patients diagnosed with pulmonary nodules in their lungs.
- Deep learning algorithms assess the malignancy risk of pulmonary nodules based on factors such as nodule size, shape, density, volume, as well as patient demographics.
- Use AWS Deep Learning AMI and the TensorFlow machine learning framework to train computer vision algorithms for CT scans.

# Using Deep Learning to Detect Heart Disease



- HeartFlow creates personalized medical technology using deep learning to help diagnose heart disease.
- Accelerated by NVIDIA GPUs, HeartFlow's solution analyzes CT scans to create a 3D model of a patient's heart and coronary arteries.
- In addition to creating an accurate 3D model, the system simulates the flow of blood in each vessel.
- Uses the Caffe deep learning framework on P2 instances; exploring TensorFlow on G3.

# Advancing Cardiac Visualization with Deep Learning



Arterys uses deep learning for innovation in medical imaging to deliver data-driven clinical patient care.

- Analyzes thousands of cardiac MRI images from global hospitals to understand patients and treatments.
- Enabled seamless visualization of medical images and solved limited computation power available to doctors today by moving from CPU to GPU computing.
- Reduced time for medical imaging analysis from 30 minutes to seconds using deep learning and Amazon G2 and S3.

# Lessons Learned & Best Practices



# Lessons learned/best practices

- Dive deep into workload and plan accordingly
- Look carefully into I/O-storage architecture
- Optimize: look at innovative accelerated options for FPGAs and GPUs
- Engage partner solutions where possible

Where to get more information ?

# More Information

## Web Pages:

AWS HPC Landing Page: [www.aws.amazon.com/hpc](http://www.aws.amazon.com/hpc)

AWS Genomics Page: <https://aws.amazon.com/health/genomics/>

AWS Biotech & Pharma Page: <https://aws.amazon.com/health/biotech-pharma/>

## Blogs:

[Building High Throughput Genomics Batch Workflows on AWS](#)

## White Papers

[Architecting for Genomic Data Security and Compliance in AWS](#)

[How Cloud HPC is Reducing Time to Insights in Pre-clinical Research](#)

[AWS Genomics Guide](#)

