Research Computing
HPC on AWS

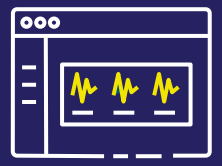
Boston Learning Days
June 18, 2025

SCOTT FRIEDMAN, PH.D. (HE/HIM)

Higher Education Research
scofri@amazon.com

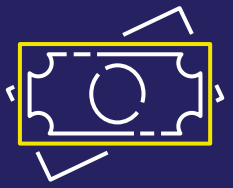
How does AWS enable research?

Enabling Research



Accelerate research
On demand access

Meet researcher capacity demands
Baseline capacity does not meet demand



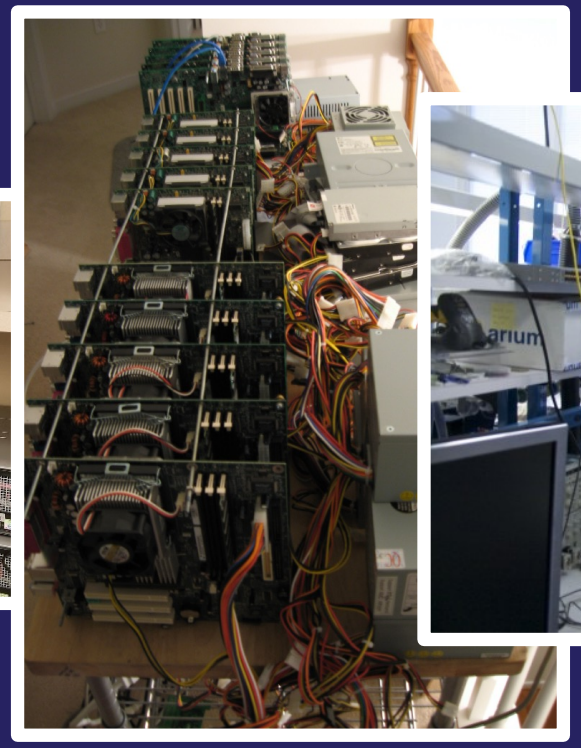
Flexible access to resources
Balance between cost and performance

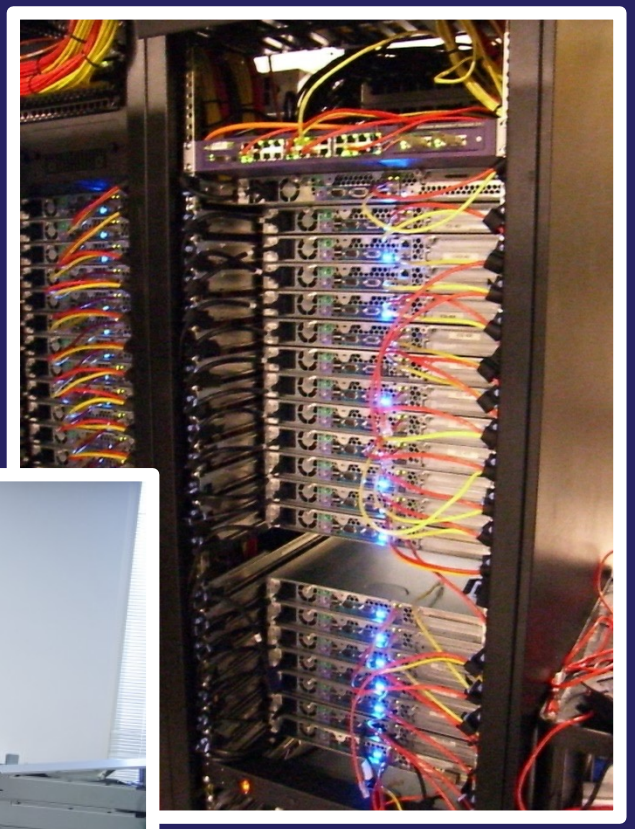
Access to services beyond HPC
AI/ML, managed compute, quantum

Tracks the pace of innovation
Access to the latest technologies

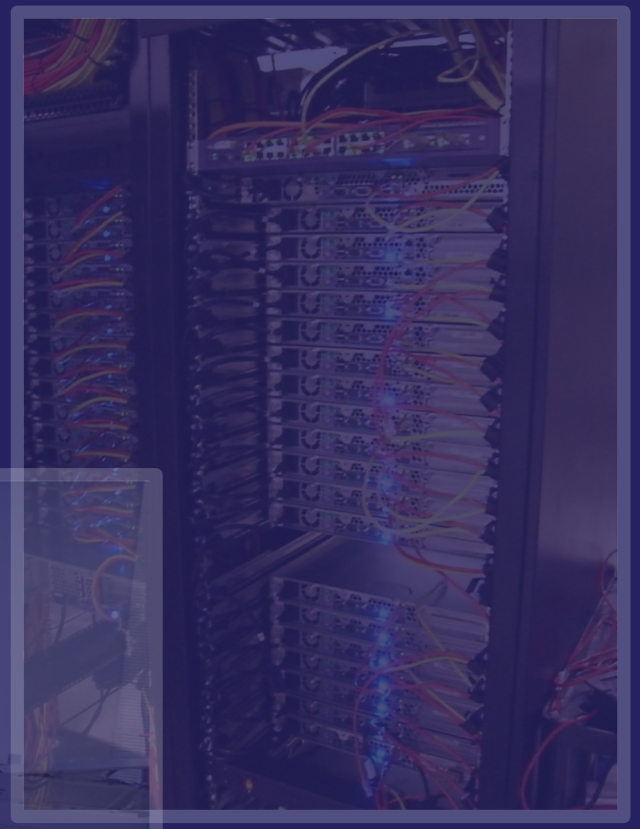
Research computing…

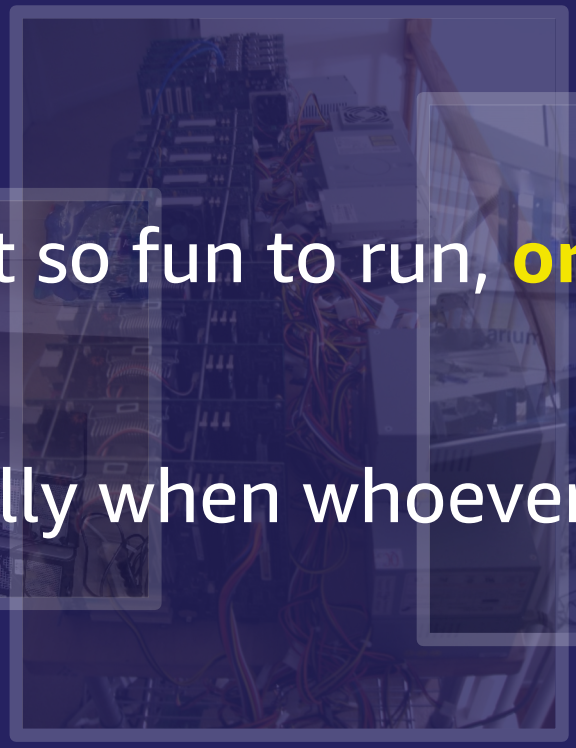
Familiar sights





Research computing...





Not so fun to run, **or keep running**

. . . especially when whoever set it up is **long gone**

Research computing…

Better

- No longer your problem
- Maybe free, condo, etc.

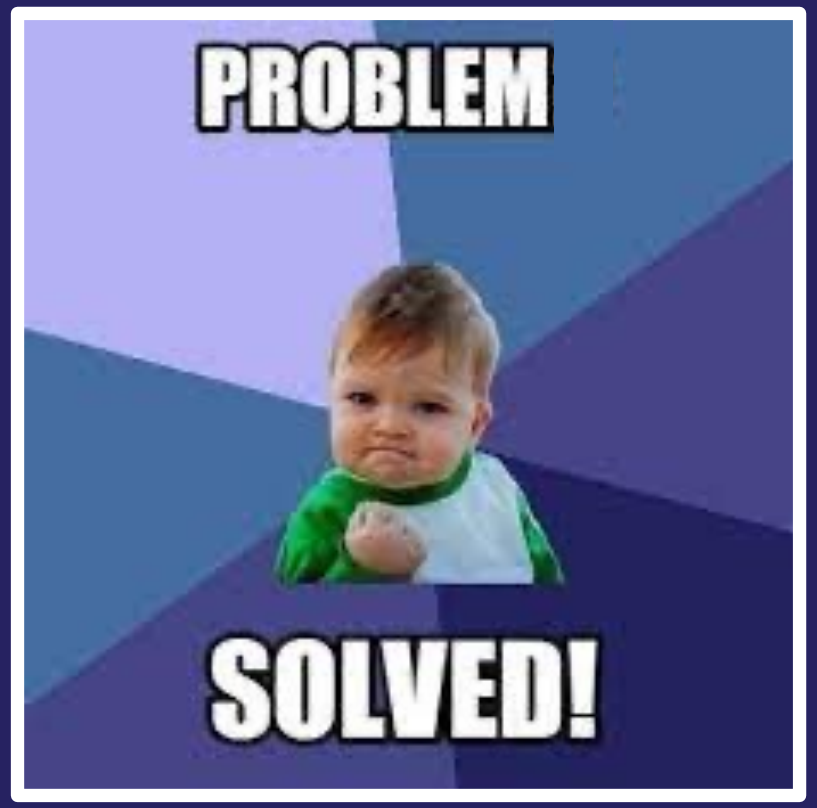
Give / Get

- Control for time
- Control for “support”
  - Institutional, other researchers
  - Power, cooling, staff, idle harvesting

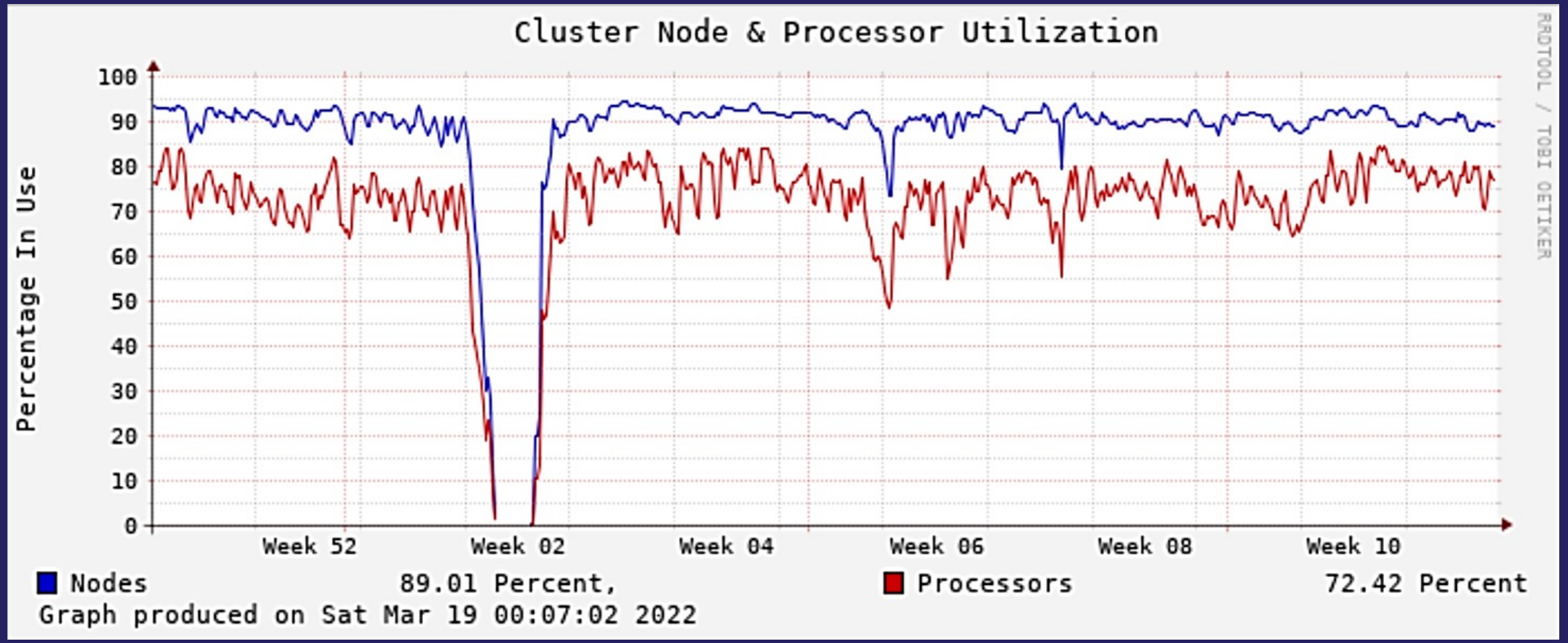
Research computing…

My message to you?

USE IT!



Research computing...



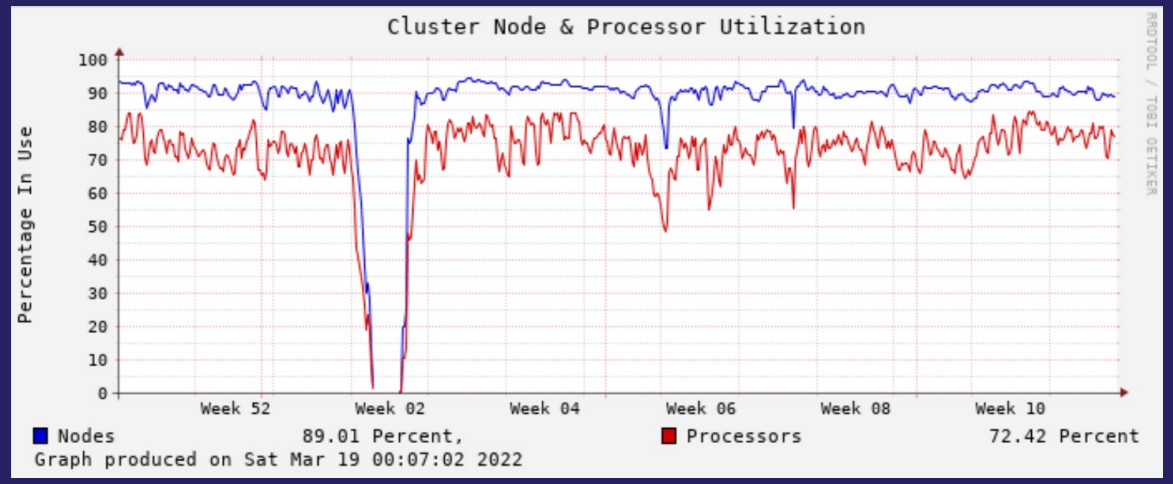
Research computing…

Familiar laboratory/campus system

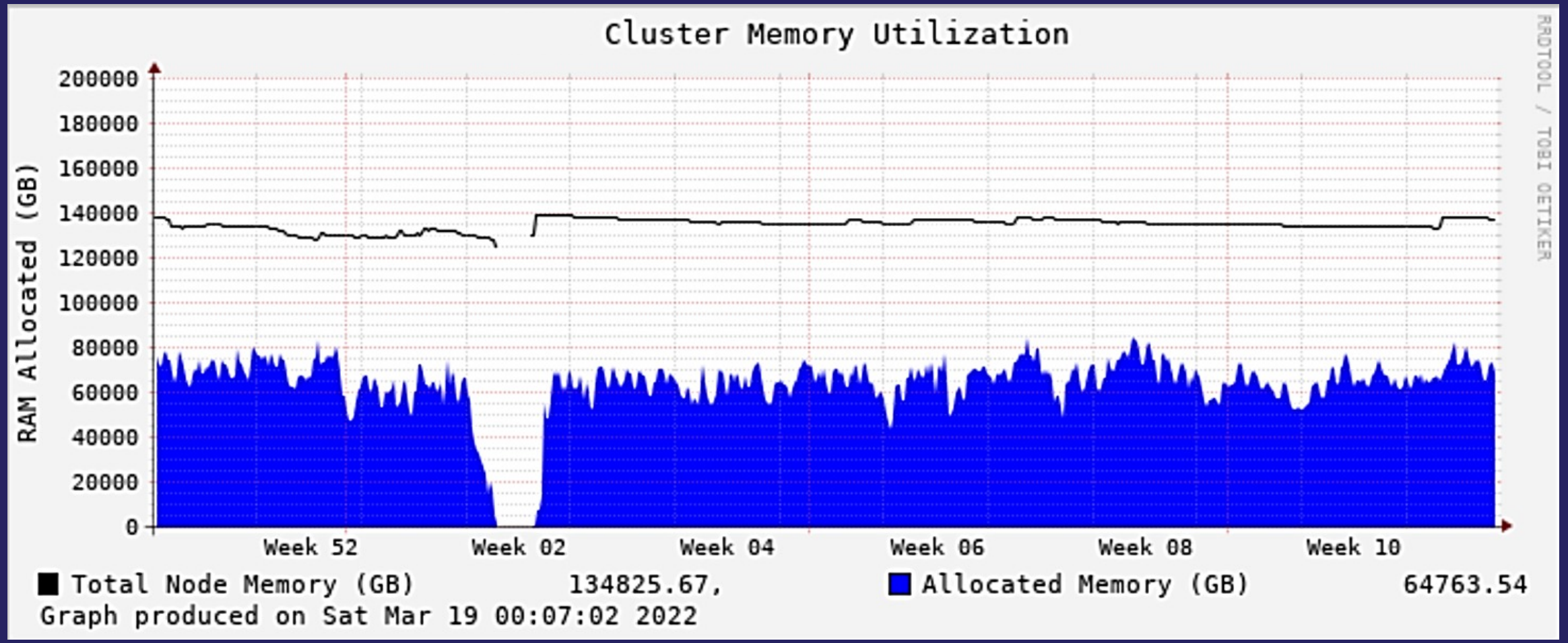
- High utilization capacity system (HPC + HTC)
- From a group/institutional perspective – has benefits
- Aggregate sustained usage

On average (even with maintenance)

- ~90% of nodes used in some way
- ~72% of cores in use



Research computing...



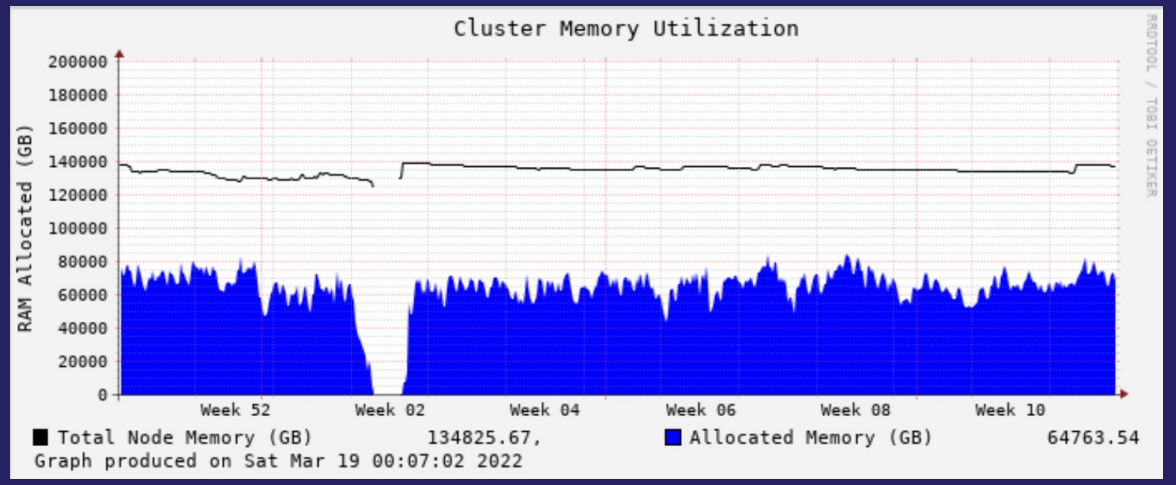
Research computing…

Familiar challenge

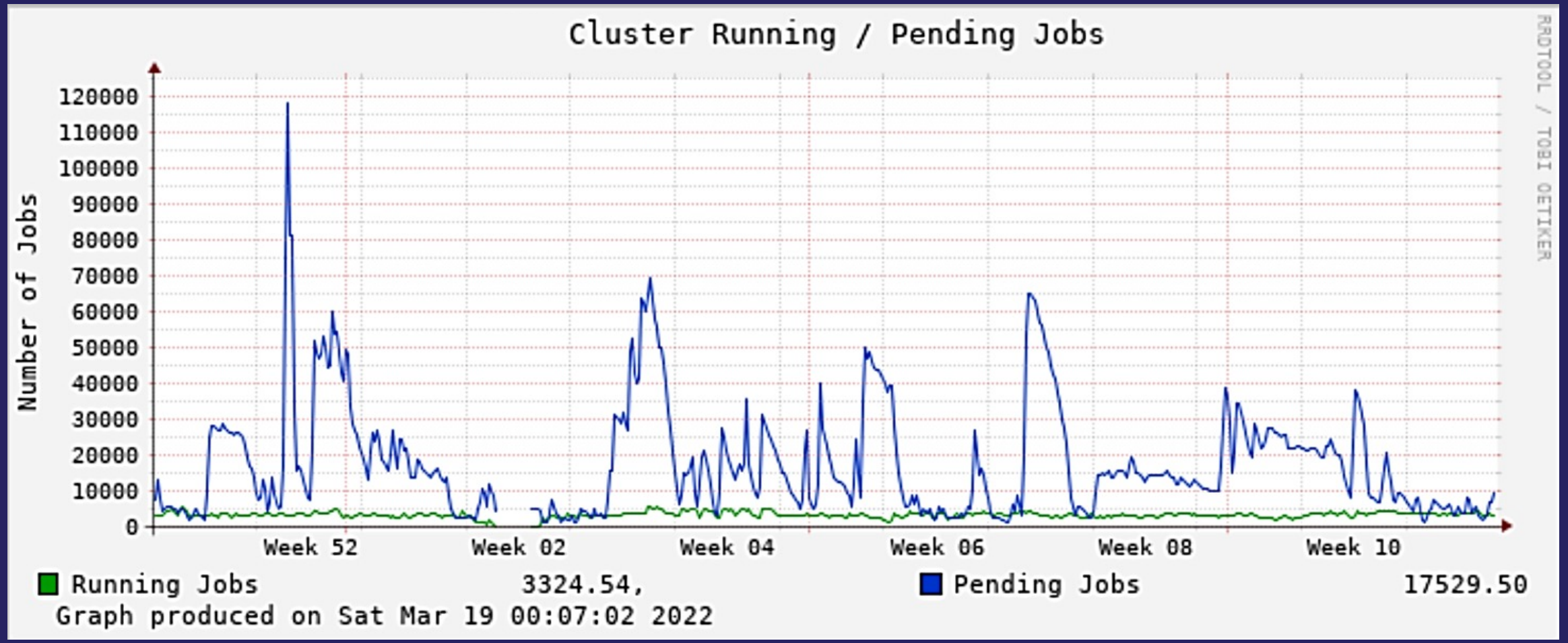
- Individual system resources end up over- or under-provisioned
  - CPU cores, memory, GPU, network
- Here memory is over-provisioned by 100%
  - 100% of the time (yes, extreme)

Do your best

- Purchase decisions made a priori
- Art more than science



Research computing…



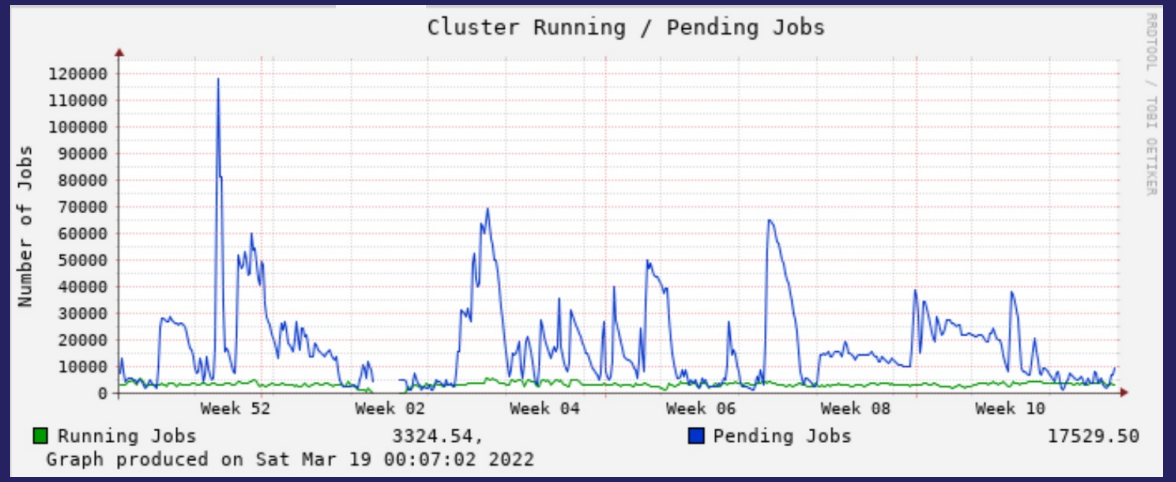
Research computing…

Familiar under-provisioned system

- ~5x on average vs. demand
- Resources are **finite** – space, power, enablement, equipment, and money
- Demand varies unpredictably over time

Support issue

- Why aren't **my** jobs running?
- I need **my** job to run now!
- Why are **others** using my nodes?



Research computing…

Familiar **individual** workload activity

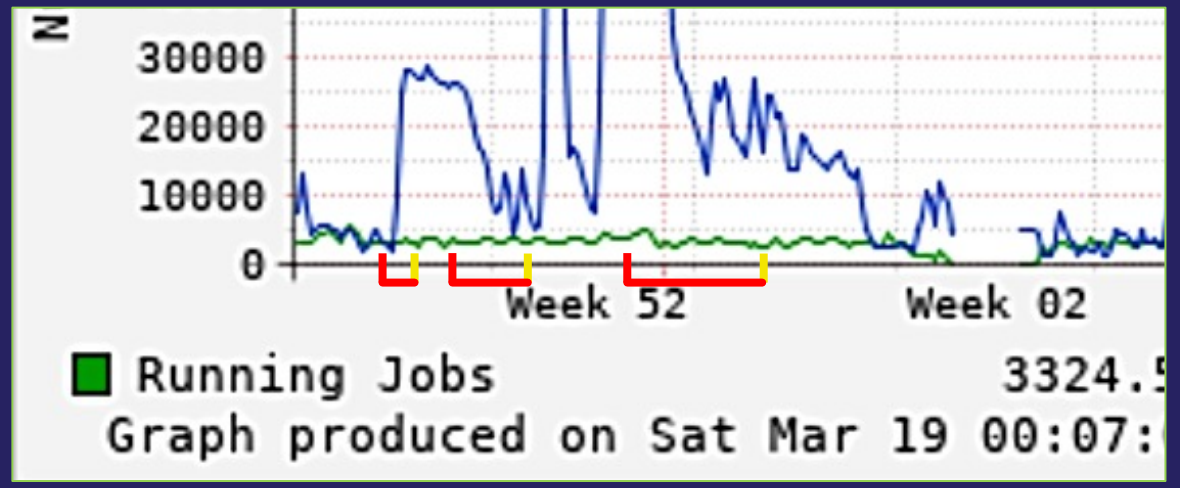
- Submitted in bursts
- Job's relationship to unknown aggregate demand
  - When will **my** job start?

L submit and wait

I schedule and run

Challenges

- Deadlines
  - Papers, conferences, graduating
- Long running jobs
  - Queue limits, maintenance



Research computing…

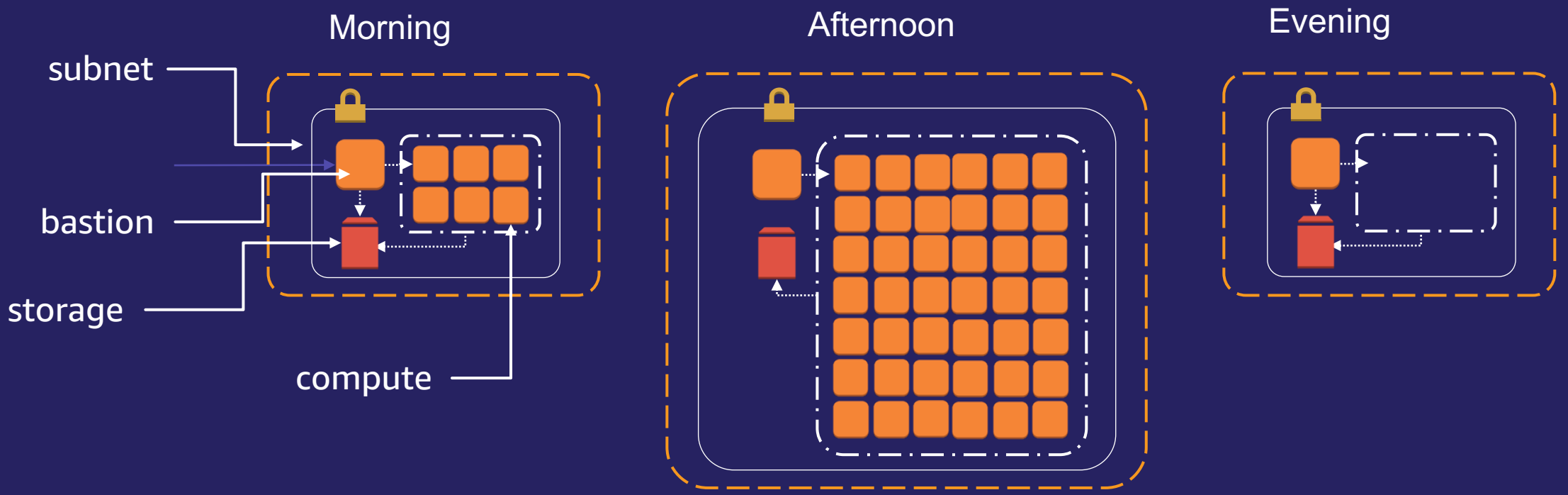
Lab/campus cluster – just fine, until . . .

- Interactive applications on batch system
- Any node type you want as long as it’s what was purchased
- Accelerators – what accelerators?
- Long-running applications, databases, portals
- Waiting…

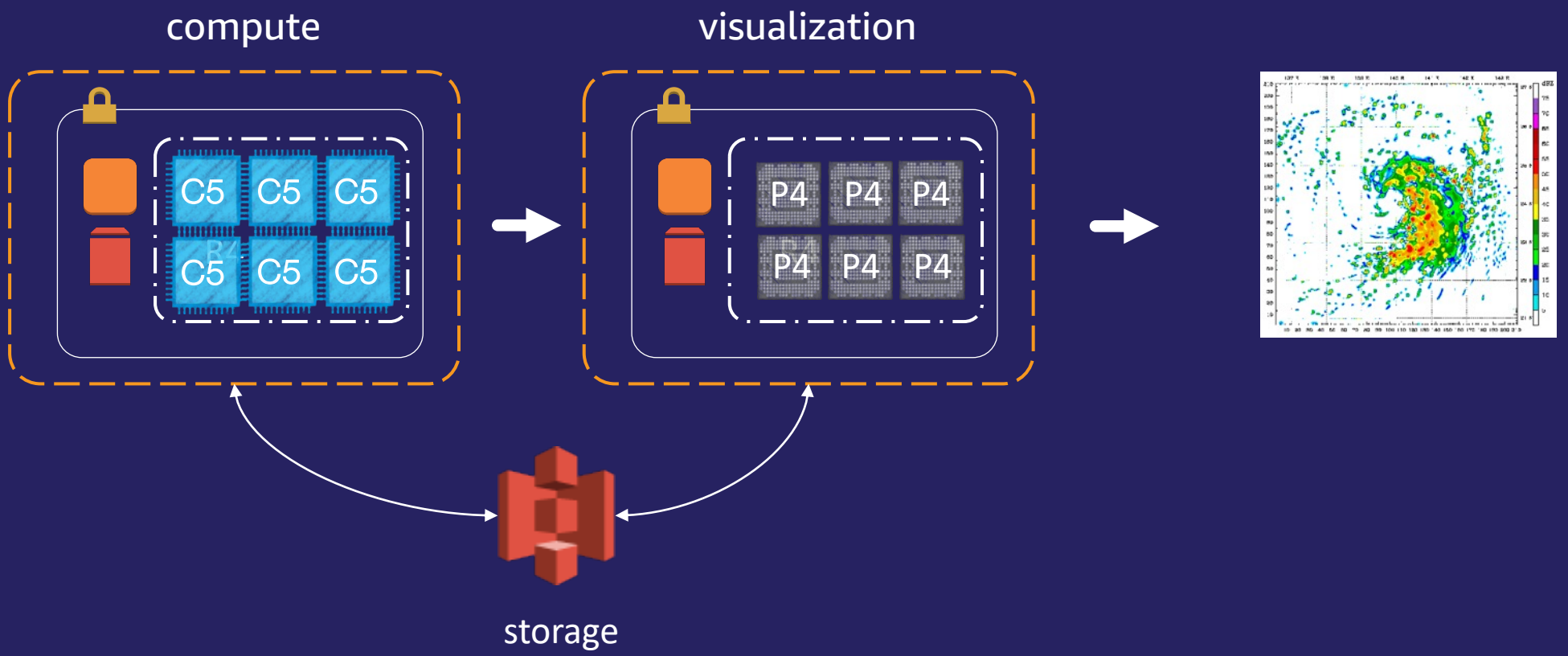
Researchers want to do research

How does AWS enable Research Computing?

Compute and Storage can be Adjusted Dynamically

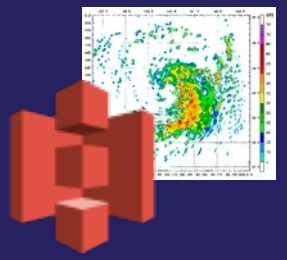


Compute and Storage can be Fit for Purpose



Compute and Storage can be Ephemeral

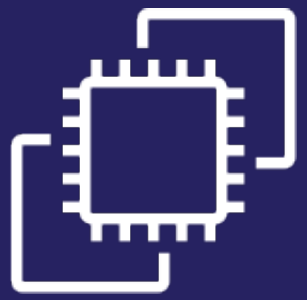
> poof <



storage

Compute and Storage can be Available on Your Schedule

Compute paradigms on AWS



Amazon EC2

Traditional Virtual, Bare Metal, and Accelerated computing

Amazon ECS, EKS, Fargate, and Batch

Container orchestration and execution

AWS Lambda

Serverless compute

Compute platforms on AWS

CATEGORIES

General purpose

Burstable

Compute intensive

Memory intensive

Storage (High I/O)

Dense storage

GPU compute

Graphics intensive

+

CAPABILITIES

Choice of processor
(AWS, Intel, AMD)

Fast processors
(up to 4.0 GHz)

High memory footprint
(up to 12 TiB)

Instance storage
(HDD and SSD)

Accelerated computing
(GPUs and FPGA)

Networking
(up to 400 Gbps)

Bare Metal

Size
(Nano to 32xlarge)

+

OPTIONS

Amazon EBS

Amazon Elastic Inference

=

MORE THAN

600+

INSTANCE TYPES

for virtually every workload and research need

AWS instance mapping to Nvidia GPUs

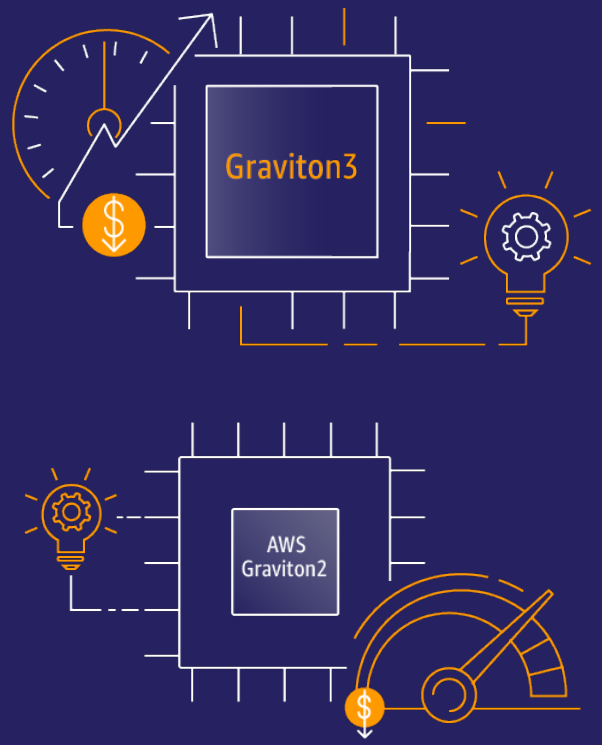
AWS instance/ NVIDIA GPU (professional grade):

Amazon EC2 P2	Amazon EC2 G3	N/A	Amazon EC2 P3	Amazon EC2 G4	Amazon EC2 P4de	Amazon EC2 P5
NVIDIA K80	NVIDIA M60	NVIDIA P100/40	NVIDIA V100 (16 GB)	NVIDIA T4	NVIDIA A100 (80GB)	NVIDIA H100
			Amazon EC2 P3dn	Amazon EC2 G5g*	Amazon EC2 P4d	Amazon EC2 P5e
			NVIDIA V100 (32 GB)	NVIDIA T4g	NVIDIA A100 (40GB)	NVIDIA H200
				Amazon EC2 G6/G6e	Amazon EC2 G5	
				NVIDIA L4/L40S	NVIDIA A10g	

NVIDIA GPU Architecture:

Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper

AWS Graviton processors



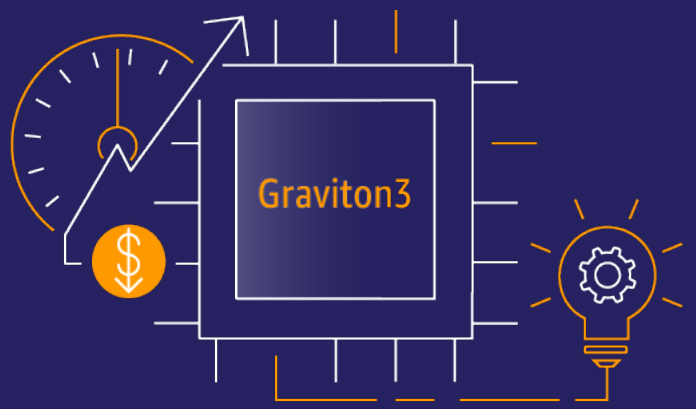
Custom AWS silicon with 64-bit Arm processor cores

Targeted optimizations for cloud-native workloads

Rapidly innovate, build, and iterate on behalf of customers

AWS Graviton3 EC2 C7g instances

- Supporting the best price performance for workloads in Amazon EC2



Up to 25% better performance compared to Graviton2

Up to 2x higher floating-point performance, up to 2x faster cryptographic workload performance, and up to 3x better machine learning performance compared to Graviton2

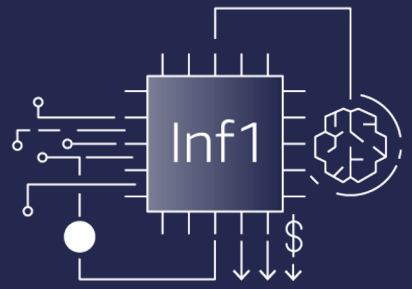
First generally available in the cloud to feature DDR5 memory

Up to 60% more energy efficient over comparable Amazon EC2 instances

C7g instances provide the best price performance for compute-intensive workloads in Amazon EC2

Purpose-built accelerators for generative AI

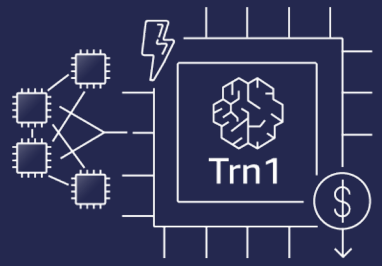
AWS Inferentia



Lowest cost per inference in the cloud for running deep learning (DL) models

Up to 70% lower cost per inference
than comparable Amazon EC2 instances

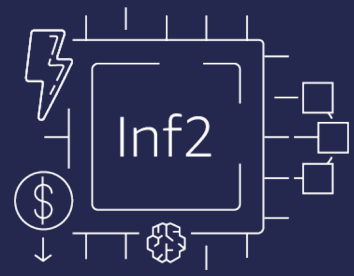
AWS Trainium



The most cost-efficient, high-performance training of LLMs and diffusion models

Up to 50% savings on training costs
over comparable Amazon EC2 instances

AWS Inferentia2



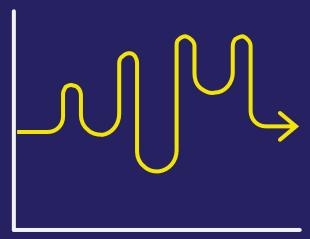
High performance at the lowest cost per inference for LLMs and diffusion models

Up to 40% better price performance
than comparable Amazon EC2 instances

Flexible options to optimize cost

On-Demand

Pay-for-what you use with no long-term commitments



Stateful Spiky workloads

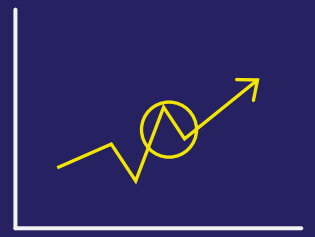
Savings Plans

Significant savings for 1 or 3 year hourly usage commitments

Committed & steady-state usage

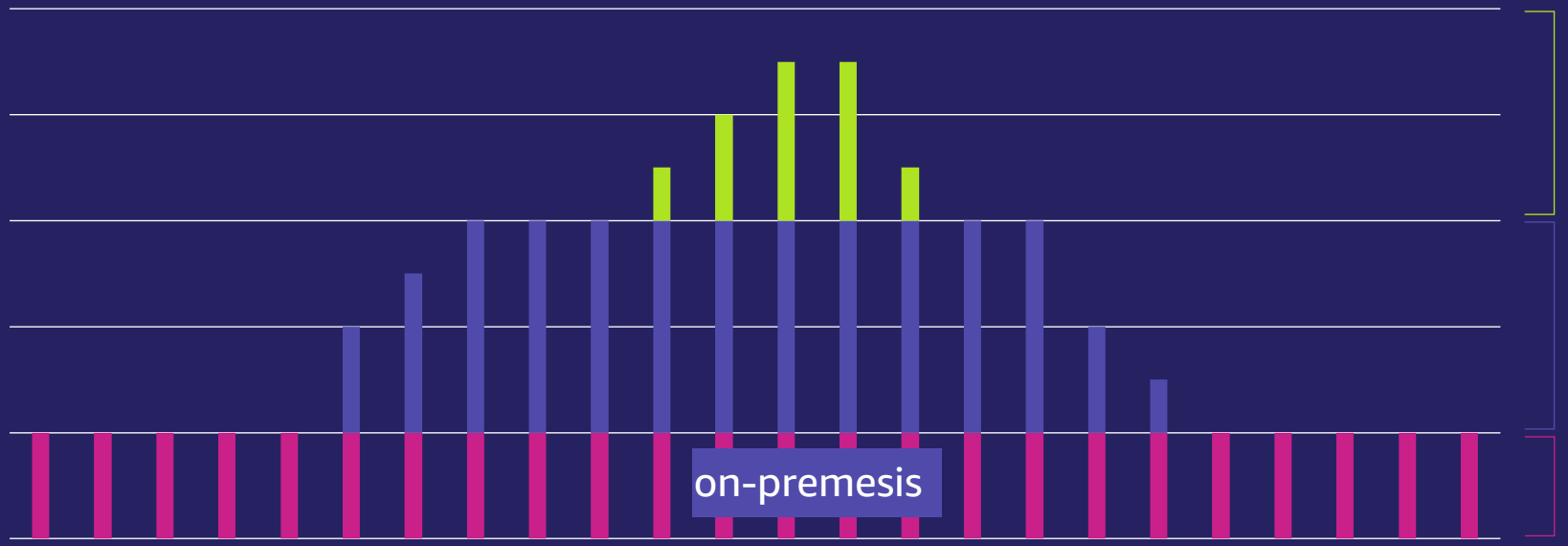
Spot

Spare capacity at up to 90% off On-Demand prices



Fault-tolerant, flexible, stateless workloads

How these options interrelate



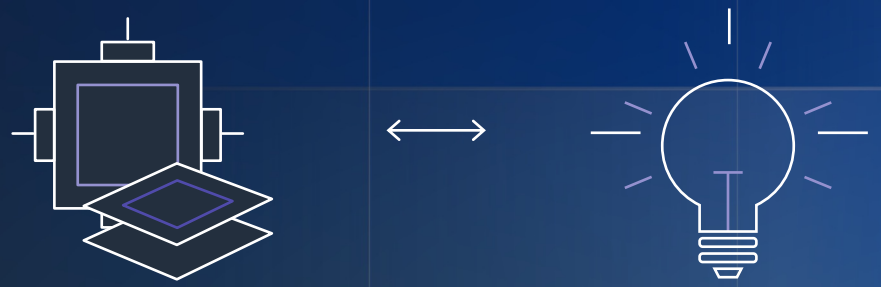
SCALE USING **SPOT** FOR FLEXIBLE, FAULT-TOLERANT WORKLOADS

SCALE USING **ON-DEMAND** FOR NEW OR STATEFUL SPIKY WORKLOADS

USE **SAVINGS PLANS** FOR KNOWN/STEADY-STATE WORKLOADS

AWS SERVICES MAKE THIS EASY AND EFFICIENT

- Amazon EC2 Auto Scaling
- EC2 Fleet
- Amazon Elastic Container Service (Amazon ECS)
- Amazon Elastic Kubernetes Service (Amazon EKS)
- AWS Thinkbox
- Amazon EMR
- AWS CloudFormation
- AWS Batch



HPC on AWS

Flexible configuration and virtually unlimited scalability to grow and shrink your infrastructure as your HPC workloads dictate, not the other way around

AWS cloud native HPC

AWS ParallelCluster

Simplifies deployment of HPC in the cloud, including integrating with popular HPC schedulers

Integrated with AWS Batch, Amazon FSx for Lustre and Elastic Fabric Adapter

AWS Batch

AWS Batch dynamically provisions resources, plans, schedules, and executes

No additional components to install

AWS bursting from on-premises

Use your existing on-premises scheduler and software stack to burst jobs into AWS providing availability, scale, and capacity for your research.

What is bursting?

HED HPC BURSTING

- **Level set:**
- Bursting is an ability to schedule jobs from an on-premises cluster and run jobs in AWS.
- Resources provisioned in the cloud
  - Operate like any other on-premises cluster node or storage.
  - Are provisioned as needed and deprovisioned when idle by the on-premises scheduler.
- Access is enabled and controlled through queues/partitions defined by the on-premises scheduler.
- User experience is (should be) identical whether on-premises or cloud. i.e., submit a job to a partition.

What are the cloud alternatives?

HED HPC BURSTING

- **The obvious alternative is Parallelcluster and Parallelcluster Service (PCS)**
- **There are others:**
  - RONIN, etc. (installable software)
  - Rescale, etc. (PaaS)
  - Roll your own
- **Why not these?**
  - Any of these may be appropriate depending on customer needs.
  - Bursting is and and, not an or.

Why bursting?

- Bursting is a way to enable a large number of research computing users on AWS.
- Researchers are already using and know how to use on-premises systems.
  - Their software, data, and workflows are already there.
- On-premises systems are not going anywhere!
  - The financial incentives are too strong and not changing anytime soon.

- Bursting allows research IT to leverage knowledge and skills they already have.
  - Acquisition of cloud skills needed for bursting are moderate for IT who have limited time.
  - Cluster configurations can be leveraged for bursting, no need to recreate from scratch.

Why bursting?

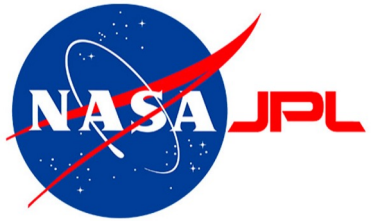
HED HPC BURSTING

- Bursting *augments* on-premises systems.
  - **It's an and not an or.**
- Scales to zero: not using = not paying.
- It's an opt-in capability: no user is forced to burst.
  - Good use cases: GPU, high-memory, etc.
- Ability to implement cost controls through the scheduler itself.
- Burst to specific AWS resources and accounts.

Easily run your Python code on AWS

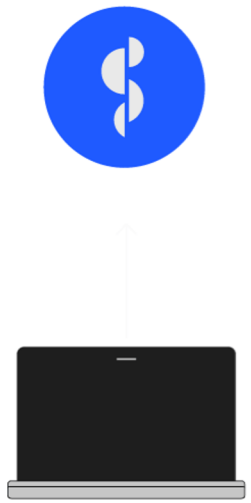
Leverage the power of the cloud without the fuss.

Spend your time doing research, not learning infrastructure



Run your Python code in the cloud

- Run securely in your AWS account
- Run from anywhere you use Python
- Run on any hardware



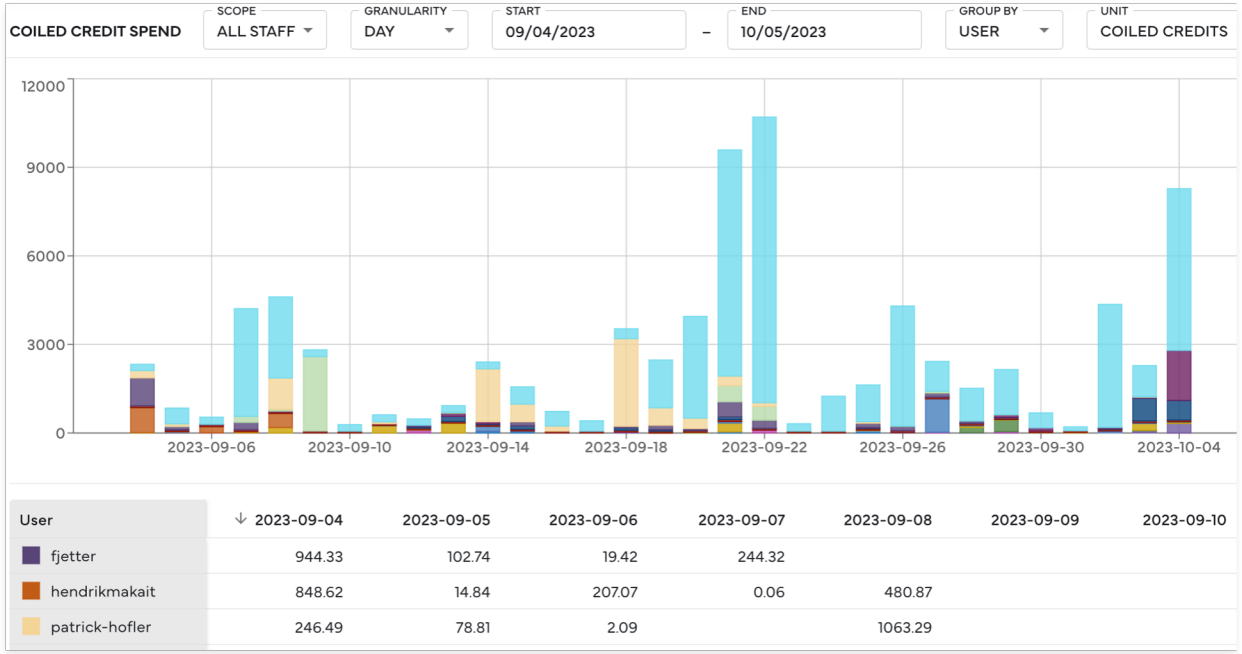
Minimal code changes

- Run your Python functions in the cloud
- Scale your favorite libraries with Dask

```python
import coiled
import pandas as pd

df = pd.read_csv("data.csv")  # Read local data

@coiled.function()            # This function runs remotely
def process(df):
    df = df[df.name == "Alice"]
    return df

result = process(df)          # Process on remote VM
print(result)
```

Easily collaborate with your team

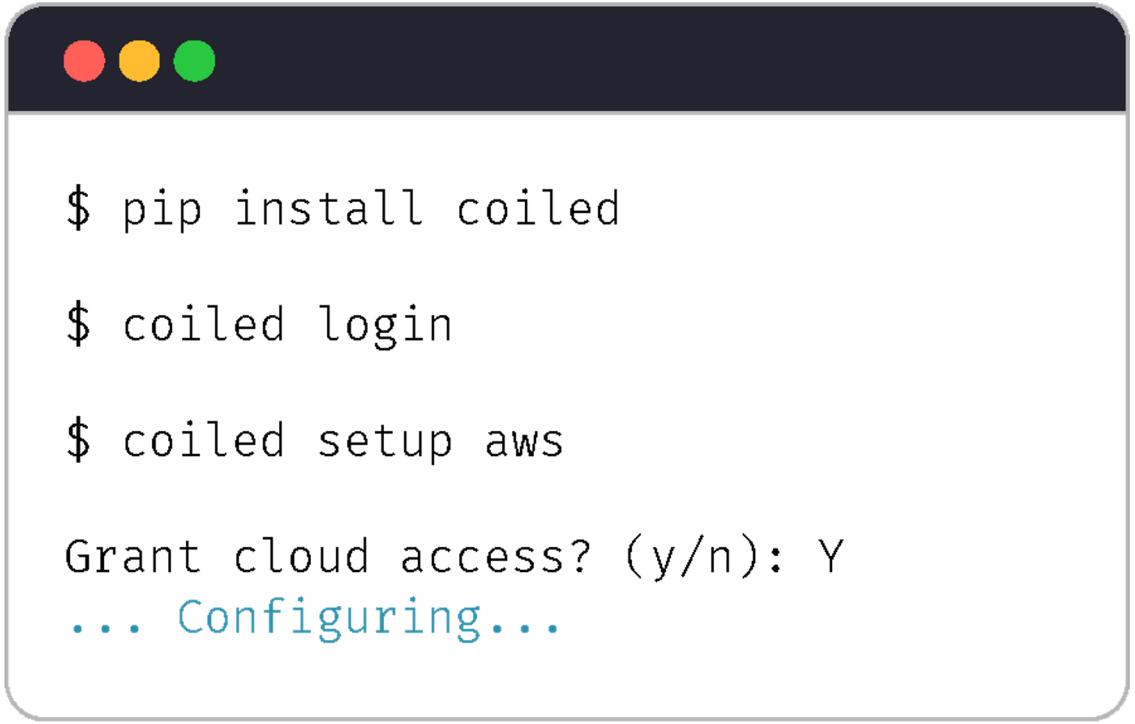
- Visibility & monitoring of team usage
- Idle shutdown
- Set spend limits
- Save money with spot instances, ARM



User	2023-09-04	2023-09-05	2023-09-06	2023-09-07	2023-09-08	2023-09-09	2023-09-10
fjetter	944.33	102.74	19.42	244.32			
hendrikmakait	848.62	14.84	207.07	0.06	480.87		
patrick-hofler	246.49	78.81	2.09		1063.29		

Easy to get started

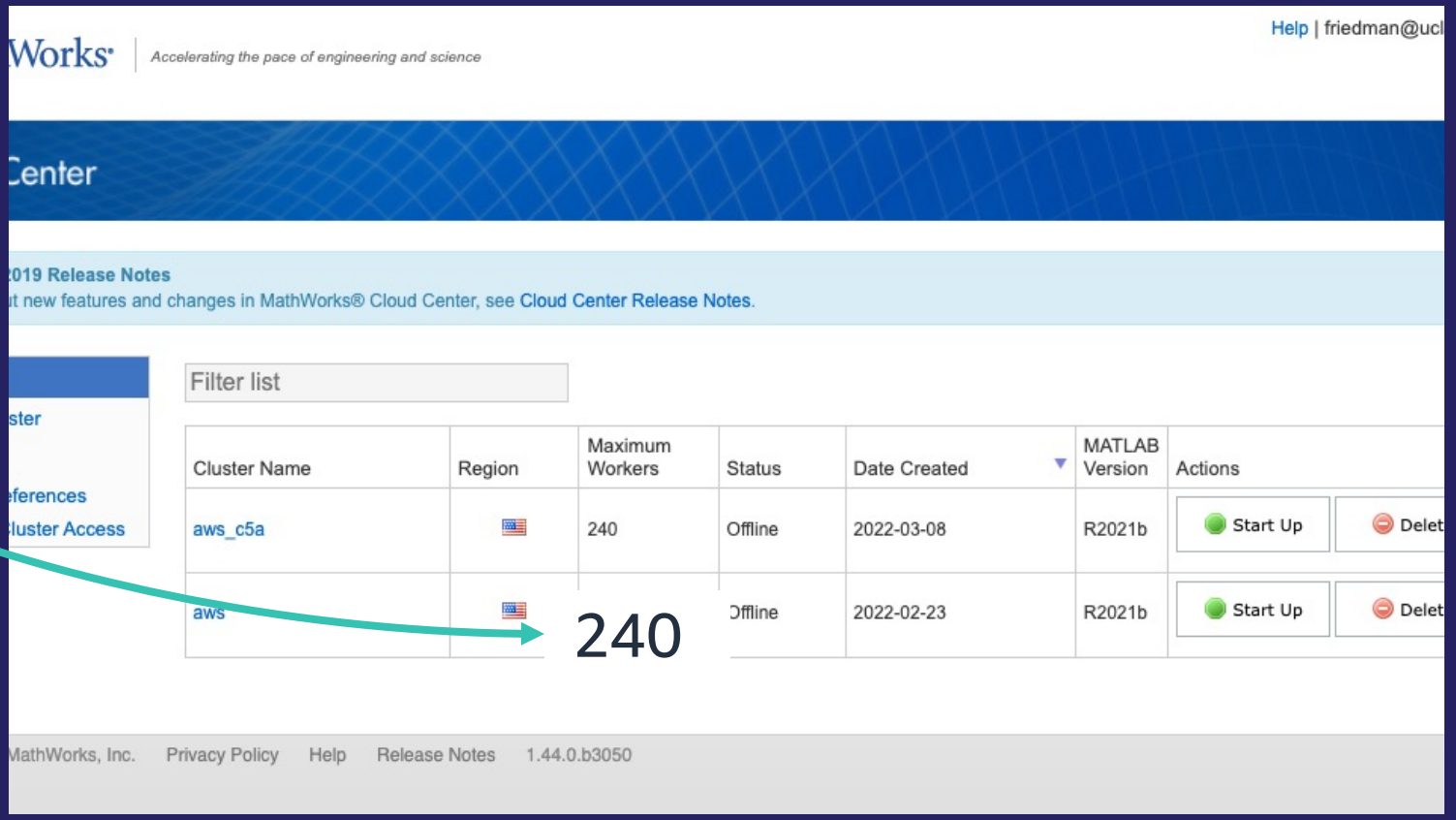
Start work on the generous free tier with 10,000 CPU-hours per month, more than enough for most individuals.



MATLAB + Cloud Center

Create cloud parallel pool* from your laptop

What?



https://cloudcenter.mathworks.com/

* Parallel Computing Toolbox Required

Research Data and Storage

The universe of research data and its challenges

Growing Exponentially

From new sources

Increasingly diverse

Used by many researchers

Analyzed by many applications

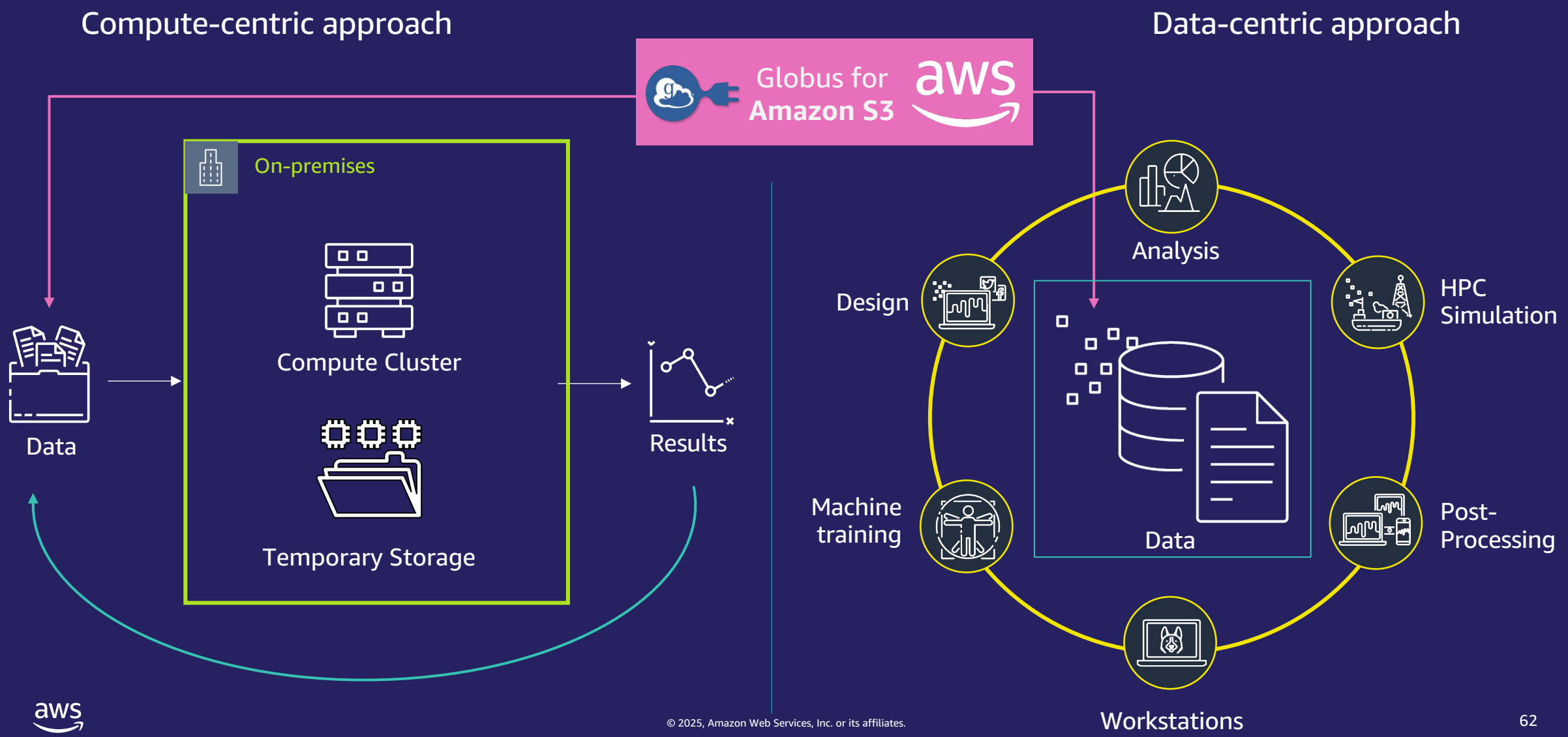
Administrative overhead

Lack of scalability

Lack of Agility

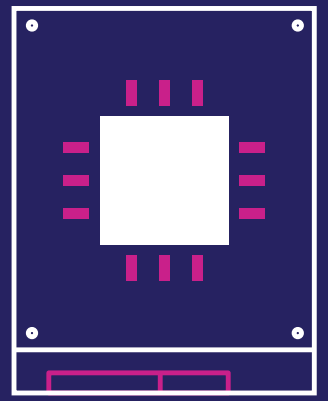
Data silos...

AWS enables a data oriented approach to research

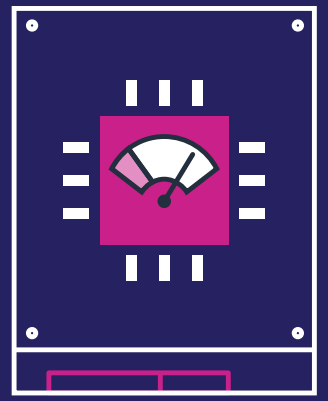


Elastic Block Storage (EBS)

SSD



gp2 – gp3
General Purpose SSD
$0.08-$0.10/GB/mo*

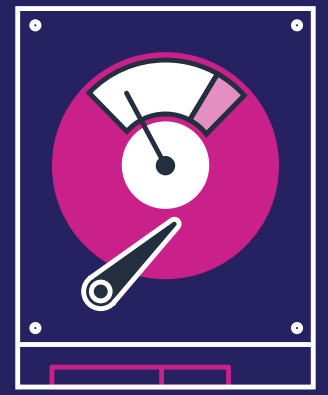


io1 – io2
Provisioned IOPS SSD
$0.125/GB/mo*

* Additional provisioned throughput or IOPS charges may apply

HDD

st1
Throughput Optimized HDD
$0.045/GB/mo



sc1
Cold HDD
$0.015/GB/mo

S3 Object Storage

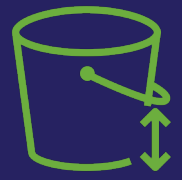
Amazon S3

S3 Standard

- Active, frequently accessed data
- Milliseconds access
- ≥ 3 AZ
- $0.0210/GB

S3 Intelligent-Tiering

- Data with changing access patterns
- Milliseconds access
- ≥ 3 AZ
- $0.0210 to $0.0125/GB ($0.004 to $0.00099/GB Archive)
- No retrieval fees
- Monitoring fee per Obj.
- Min storage duration
- Min object size



S3 Standard-IA

- Infrequently accessed data
- Milliseconds access
- ≥ 3 AZ
- $0.0125/GB
- Retrieval fee per GB
- Min storage duration
- Min object size

S3 One Zone-IA

- Re-creatable, less accessed data
- Milliseconds access
- 1 AZ
- $0.0100/GB
- Retrieval fee per GB
- Min storage duration
- Min object size

S3 Glacier Instant Retrieval

- Archive data instant retrieval
- Milliseconds access
- ≥ 3 AZ
- $0.0040/GB
- Retrieval fee per GB
- Min storage duration
- Min object size

S3 Glacier Flexible Retrieval

- Archive data
- Select minutes or hours
- ≥ 3 AZ
- $0.0036/GB – ($4.10/TB)
- Retrieval fee per GB
- Min storage duration
- Min object size

S3 Glacier Deep Archive

- Archive data
- Select 12 or 48 hours
- ≥ 3 AZ
- $0.00099/GB - ($1.01/TB)
- Retrieval fee per GB
- Min storage duration
- Min object size

Frequent ← *Access Frequency* → *Infrequent*

aws

Elastic File System (EFS) – Unlimited network storage

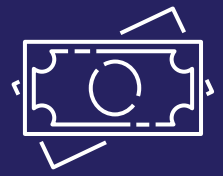
This is NFS!

Amazon FSx for Lustre

FULLY MANAGED SHARED STORAGE BUILT ON THE WORLD'S MOST POPULAR HIGH-PERFORMANCE FILE SYSTEM

Sub-ms latencies, **hundreds of GB/s of throughput,** millions of IOPS

Concurrent access for thousands of instances and **100,000s of cores**



Cost-optimized file systems with HDD and SSD storage options

Flexible deployment options for short- and longer-term workloads

Learn more: Amazon FSx for Lustre, https://aws.amazon.com/fsx/lustre/

aws

Amazon File Cache

HIGH-SPEED CACHE FOR DATASETS STORED ANYWHERE—ACCELERATE AND SIMPLIFY HYBRID WORKLOADS

- **Fast—access cached data** at sub-millisecond latencies and hundreds of GB/s of throughput
- **Agile—burst compute-intensive** workloads from on premises to compute resources on AWS
- **Simple—unify datasets** across S3 buckets and NFS file systems into a single namespace

---

Media and entertainment

VFX
rendering/transcoding

HPC

Financial services, health and life sciences, microprocessor design, manufacturing, weather forecasting, and energy

AI/ML

ML model training
and large-scale analytics

Globus

A data management platform that provides secure, reliable, and efficient data transfer and sharing throughout the research lifecycle

- Web-based graphical interface is used by over 150,000 researchers around the world
- Free to use
- Subscriptions enable enhanced functionality such as data sharing and storage connectors
- Globus for Amazon S3 Connector makes transferring data into S3 simple, reliable, and efficient
- Fire and forget transfers to AWS storage

Global Data Egress Waiver

Why

Researchers need predictable budgets

Who

Available to degree-granting /research institutions

Must use a research network (e.g. Internet2) or AWS Direct Connect

What

Waives charges for data downloads

Waives up to 15% of the customer total account bill

Data uploads are always free

How

Ask your AWS Account Manager

Or via Research Portals (e.g. Arcus, GEANT…)

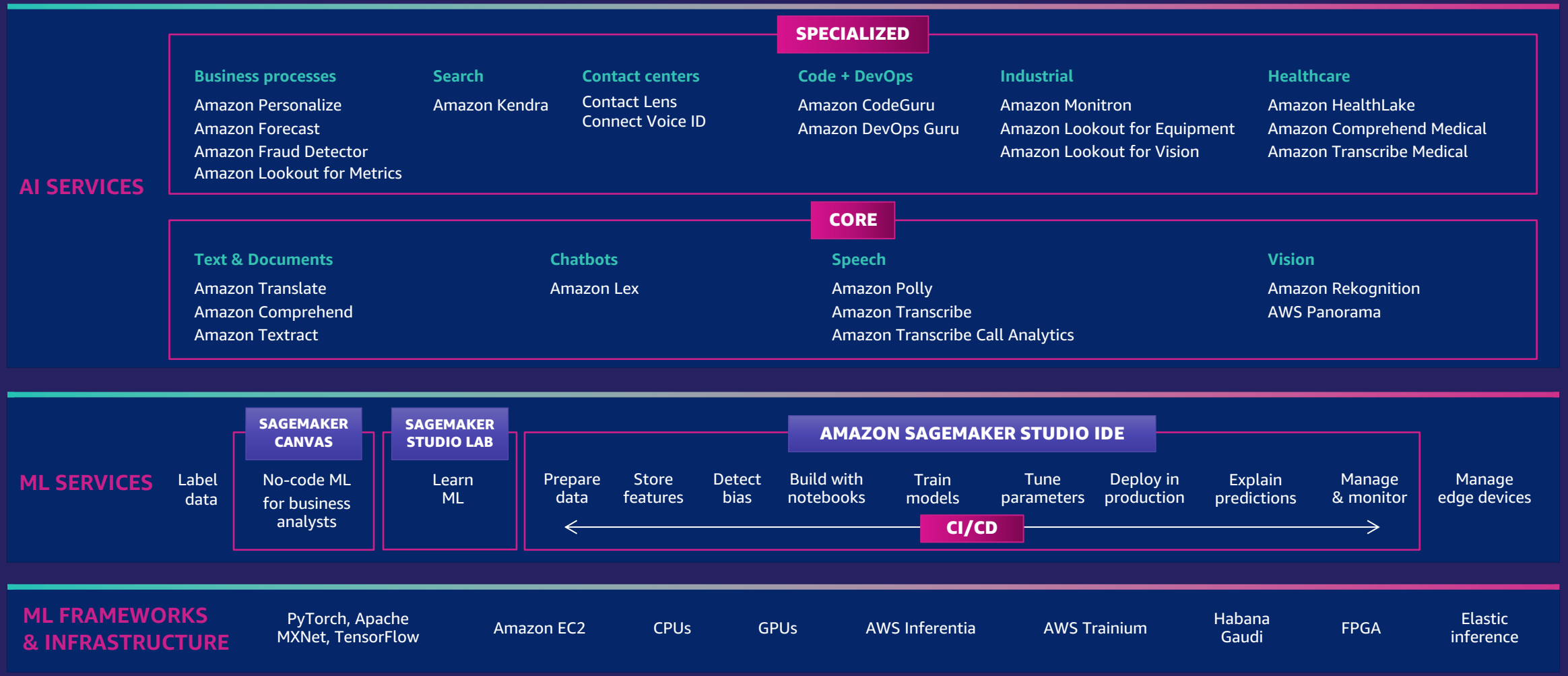
Or via certain Resellers (e.g. DLT/I2 NET+, VMWare on AWS…)

aws

Machine Learning

The AWS ML stack

Broadest and most complete set of machine learning capabilities

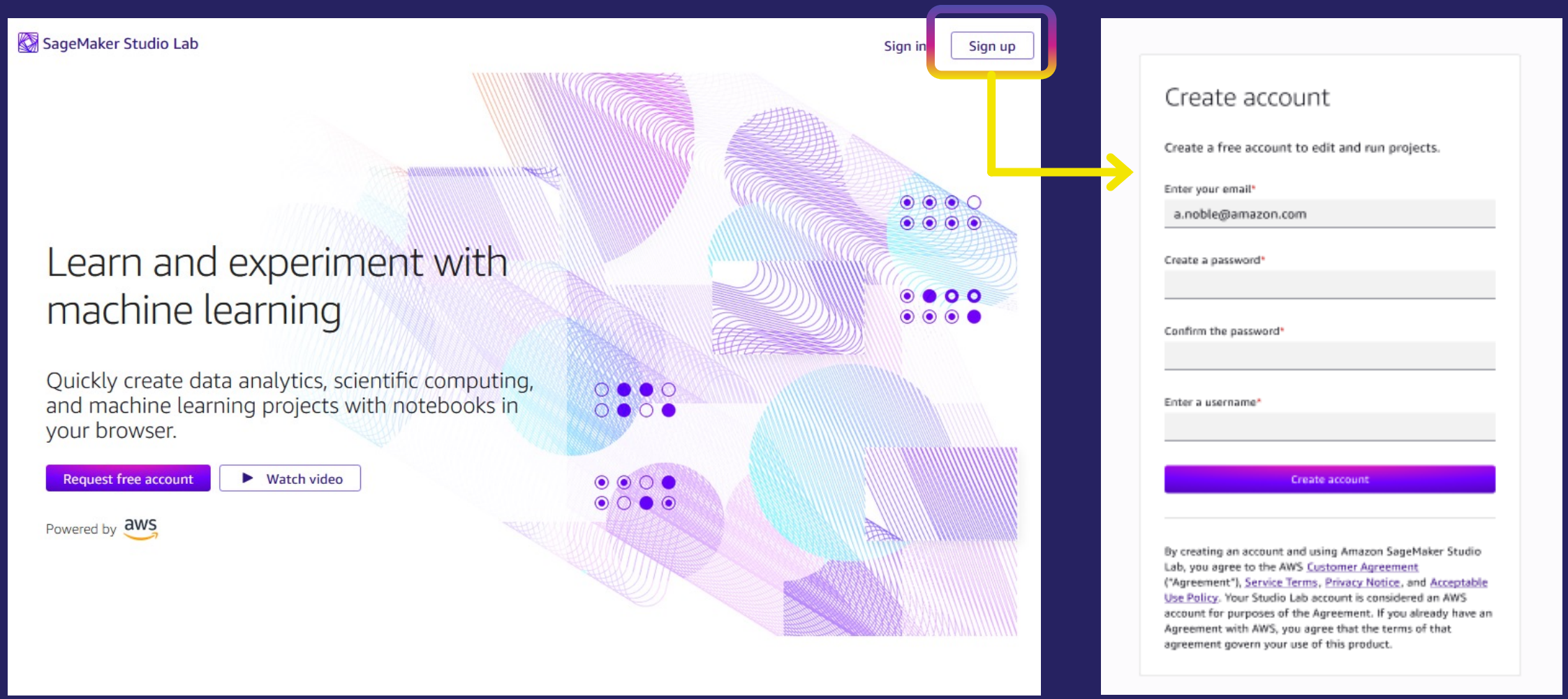


aws machine learning

https://aws.amazon.com/machine-learning/

https://studiolab.sagemaker.aws/

Request and register for an account; no credit card or AWS account required



Tailored to aspiring data scientists

→Jupyter notebook environment	Based on JupyterLab
→Easy to get started	**No-cost**, no cloud infrastructure setup
→Satisfactory compute	CPU (T3.XL) and GPU (G4D.XL)
→Time to code	Save ML project, pick up where left
→Version control management	Integrated with Git
→Supportive community	Integrated with GitHub
→Full support of shell commands	Terminal access

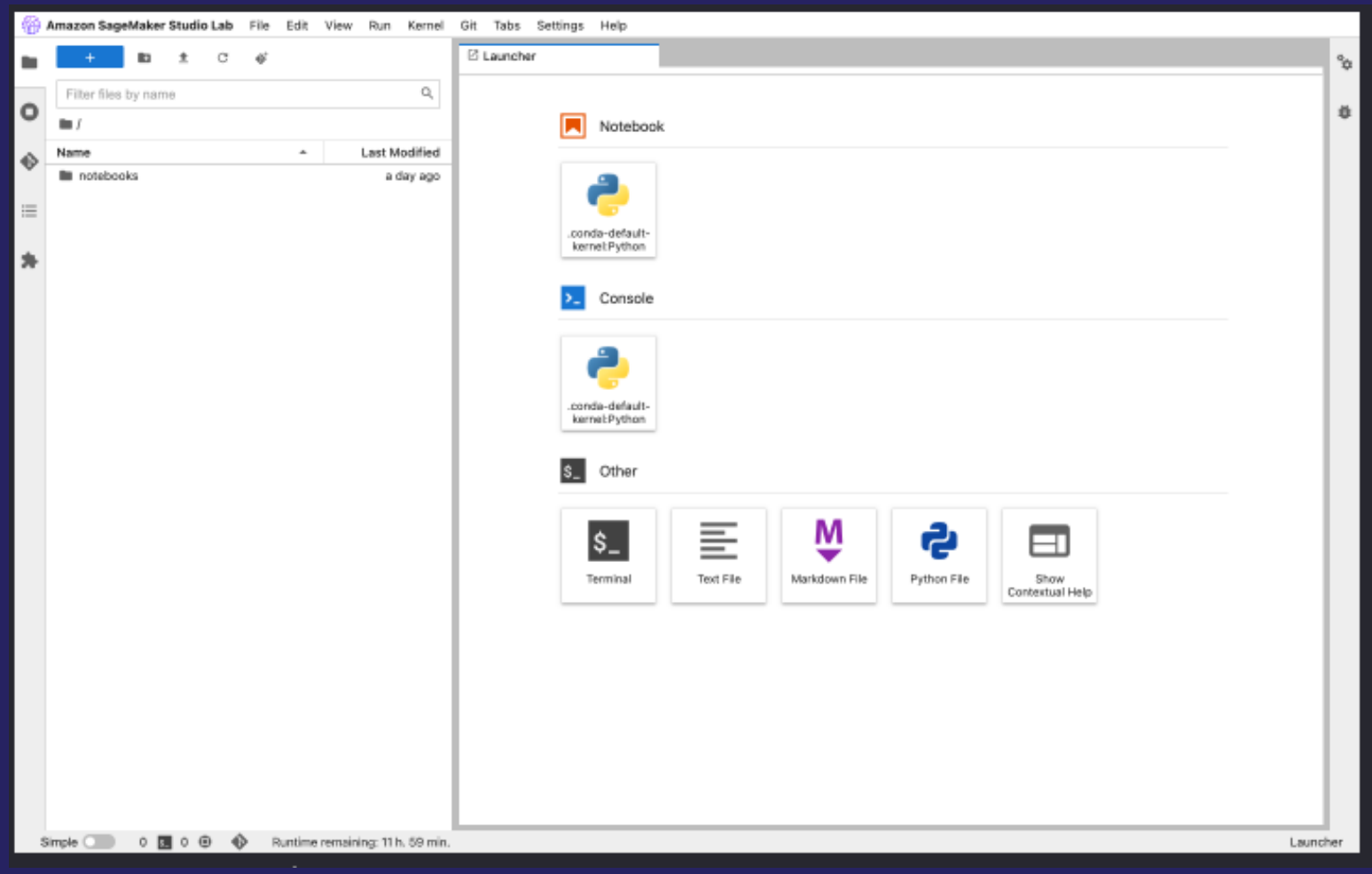
Notebook development environment

Familiar JupyterLab experience

- Terminal access
- Git/GitHub

Your ML environment on AWS

- Compute dedicated to you
- 12 hours CPU/4 hours GPU
- *%pip install* your libraries
- 15 GB storage for your project
- Unlimited user sessions
- Pick up where you left off



Built to make ML more accessible

Amazon SageMaker overview

Amazon SageMaker

PREPARE

SageMaker Ground Truth
Label training data for machine learning

SageMaker Data Wrangler
Aggregate and prepare data for machine learning

SageMaker Processing
Built-in Python, BYO R/Spark

SageMaker Feature Store
Store, update, retrieve, and share features

SageMaker Clarify
Detect bias and understand model predictions

BUILD

SageMaker Studio Notebooks
Jupyter notebooks with elastic compute and sharing

Built-in and Bring your-own Algorithms
Dozens of optimized algorithms or bring your own

Local Mode
Test and prototype on your local machine

SageMaker Autopilot
Automatically create machine learning models with full visibility

SageMaker JumpStart
Pre-built solutions for common use cases

TRAIN & TUNE

Managed Training
Distributed infrastructure management

SageMaker Experiments
Capture, organize, and compare every step

Automatic Model Tuning
Hyperparameter optimization

Distributed Training Libraries
Training for large datasets and models

SageMaker Debugger
Debug and profile training runs

Managed Spot Training
Reduce training cost by 90%

DEPLOY & MANAGE

Managed Deployment
Fully managed, ultra low latency, high throughput

Kubernetes & Kubeflow Integration
Simplify Kubernetes-based machine learning

Multi-Model Endpoints
Reduce cost by hosting multiple models per instance

SageMaker Model Monitor
Maintain accuracy of deployed models

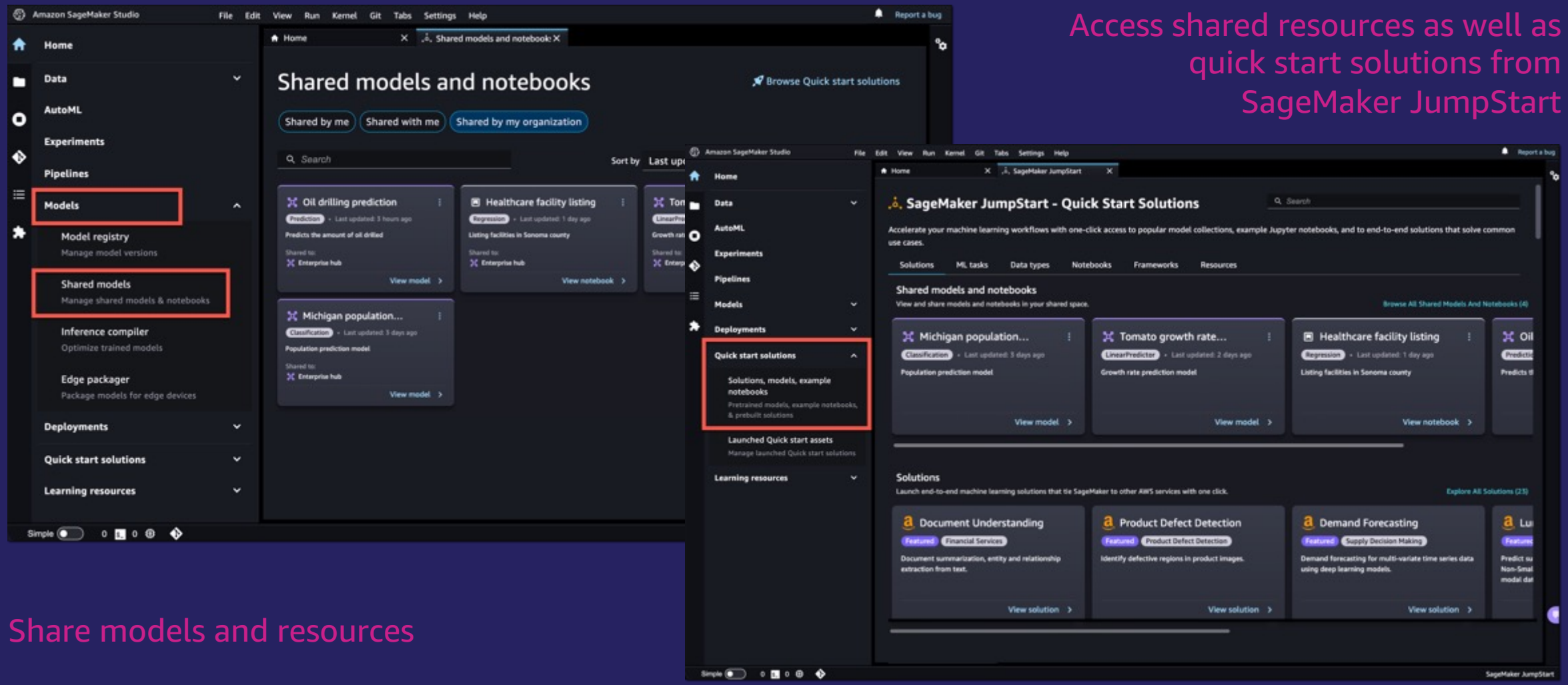
SageMaker Edge Manager
Manage and monitor models on edge devices

SageMaker Pipelines
Workflow orchestration and automation

SageMaker Studio
Integrated development environment (IDE) for ML

Amazon SageMaker JumpStart

Begin with proven solutions and collaborate to innovate



Access shared resources as well as quick start solutions from SageMaker JumpStart

Share models and resources

aws machine learning

 |

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models

Choice of industry-leading FMs available via a single API

Customize your models using your organization’s data

Enterprise-grade security and privacy

Amazon Bedrock supports leading foundation models

AI21labs

Jurassic-2

Contextual answers, summarization, paraphrasing

ANTHROP\C

Claude 3, Claude 2.1 & Claude Instant

Summarization, complex reasoning, writing, coding

cohere

Command & Embed

Text generation, search, classification

Meta

Llama 2

Dialogue use cases and language tasks

Mistral AI

Mistral 7B, Mixtral 8x7B

Text summarization, Q&A, Text classification, Text completion, code generation

stability.ai

Stable Diffusion XL 1.0

High-quality images and art

amazon

Amazon Titan

Summarization, image and text generation and search, Q&A

Privately customize models with your data

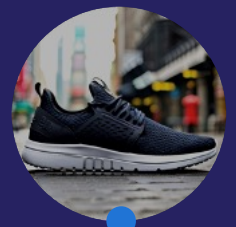
FINE-TUNING AND CONTINUED PRE-TRAINING

Deliver tailored, differentiated tail user experiences with customized FMs

Fine-tune Llama 2, Command, and Titan FMs for specific tasks with labeled data

Use continued pre-training to adapt Titan Text FMs to your domain with unlabeled data

None of your inputs to or outputs from Amazon Bedrock will be used to train the original base models



Labeled data
AMAZON S3

MODEL

Amazon—**Titan Text**

Cohere—**Command**

Meta—**Llama2**

AMAZON BEDROCK

Meta—**Llama2**

Copy

FINE-TUNED MODEL

GENERATED CONTENT

Social media

Display ads

Web copy

Knowledge bases for Amazon Bedrock

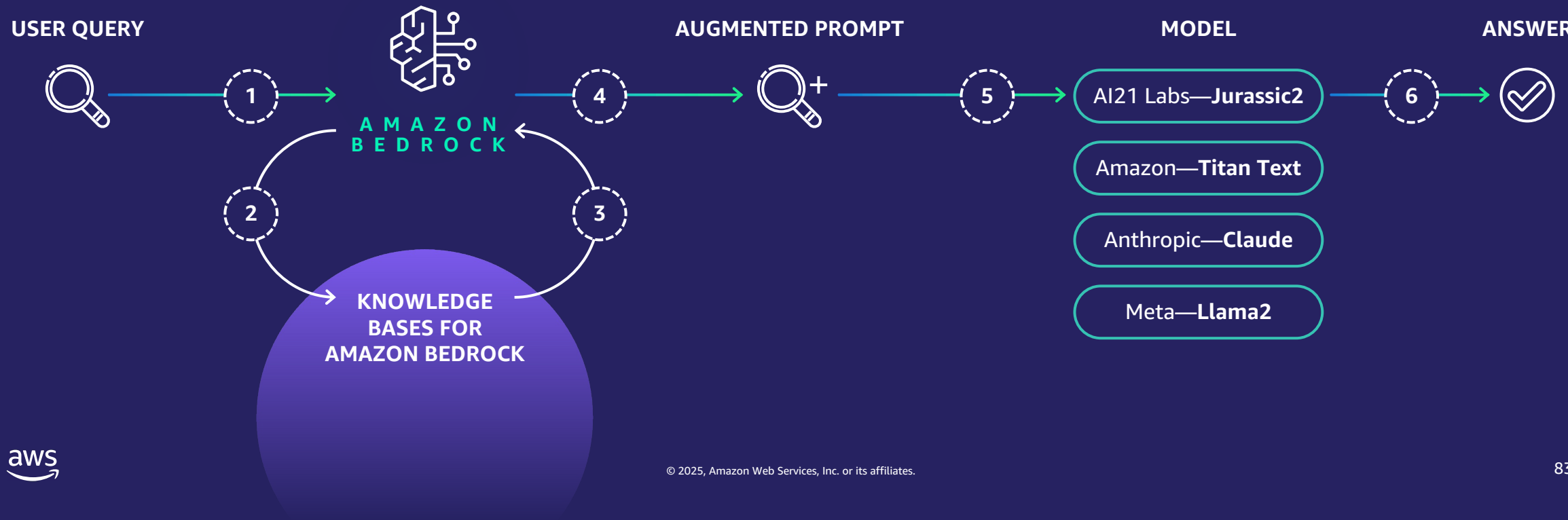
NATIVE SUPPORT FOR RETRIEVAL AUGMENTED GENERATION (RAG)

Securely connect FMs to data sources for RAG to deliver more relevant responses

Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multi-turn conversations

Automatic citations with retrievals to improve transparency



Data Security and Privacy

Amazon Bedrock

Helps keep your data secure and private

None of the customer's data is used to train the underlying models

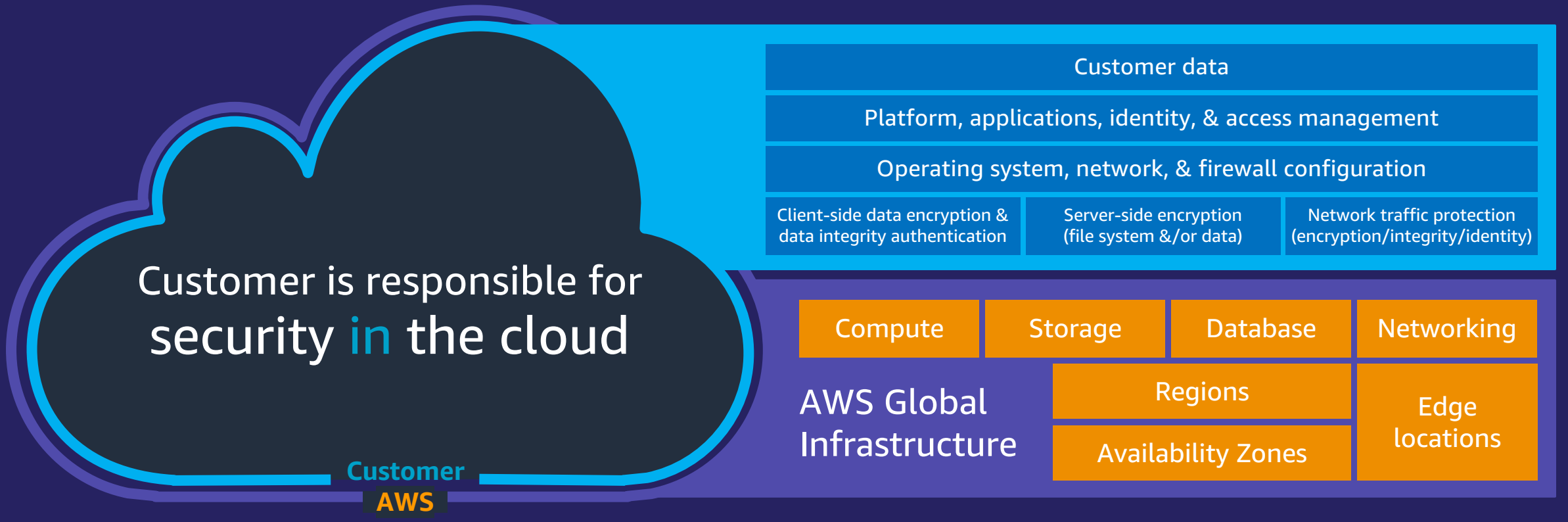
All data is encrypted in transit and at rest; data used for customization is securely transferred through customer's VPC

Data remains in the region where API processed

Support for GDPR, SOC, ISO, CSA compliance and HIPAA eligibility

Security and compliance

Shared security model



AWS is responsible for
security **of** the cloud

Shared operational responsibility model explained

More opinionated

Less opinionated

	AWS manages	Customer manages
AWS Lambda Serverless functions	• Data source integrations • Application-level runtime and updates • Physical hardware, software, networking, and facilities • Provisioning	• Application code • Security and network configuration
AWS Fargate Serverless containers	• Container orchestration, provisioning • Cluster scaling • Physical hardware, host OS/kernel, networking, and facilities	• Application code • Service scaling • Data source integrations • Security config network config, management tasks, application runtime updates
ECS/EKS Container-management as a service	• Container orchestration control plane • Physical hardware software, networking, and facilities	• Application code • Data source integrations • Worker hosts and service scaling • Security config and updates, network config, firewall, management tasks
EC2 Infrastructure-as-a-Service	• Physical hardware software, networking, and facilities	• Application code • Data source integrations • Instance scaling • Security config and updates, network config, management tasks • Provisioning, managing scaling and patching of servers

Research presents a unique challenge

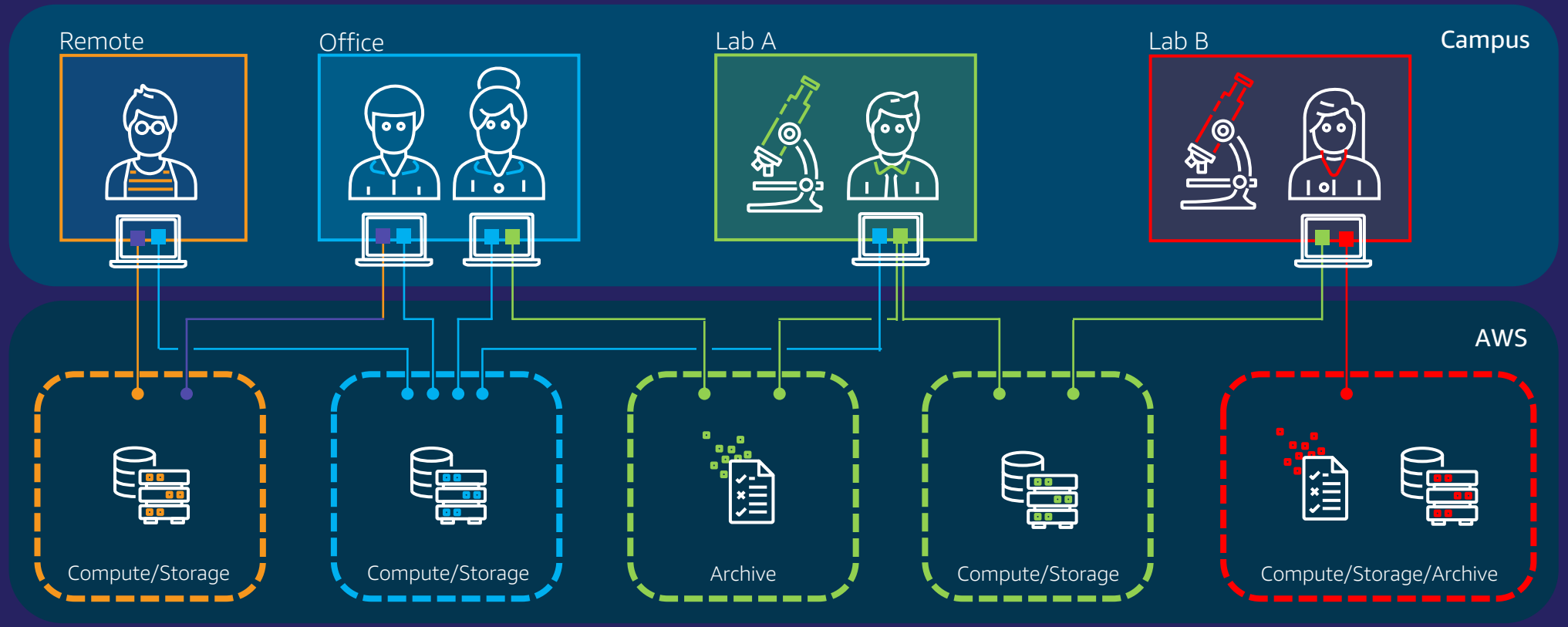
On-premises solutions for research have limits.

Factors:

- Based on point-in-time technology and are either overly generic or tailored to an initial research project's needs.
- Adapting to evolving research and compliance requirements is complex and expensive.
- Researchers want to share and collaborate. Data silos make this challenging.

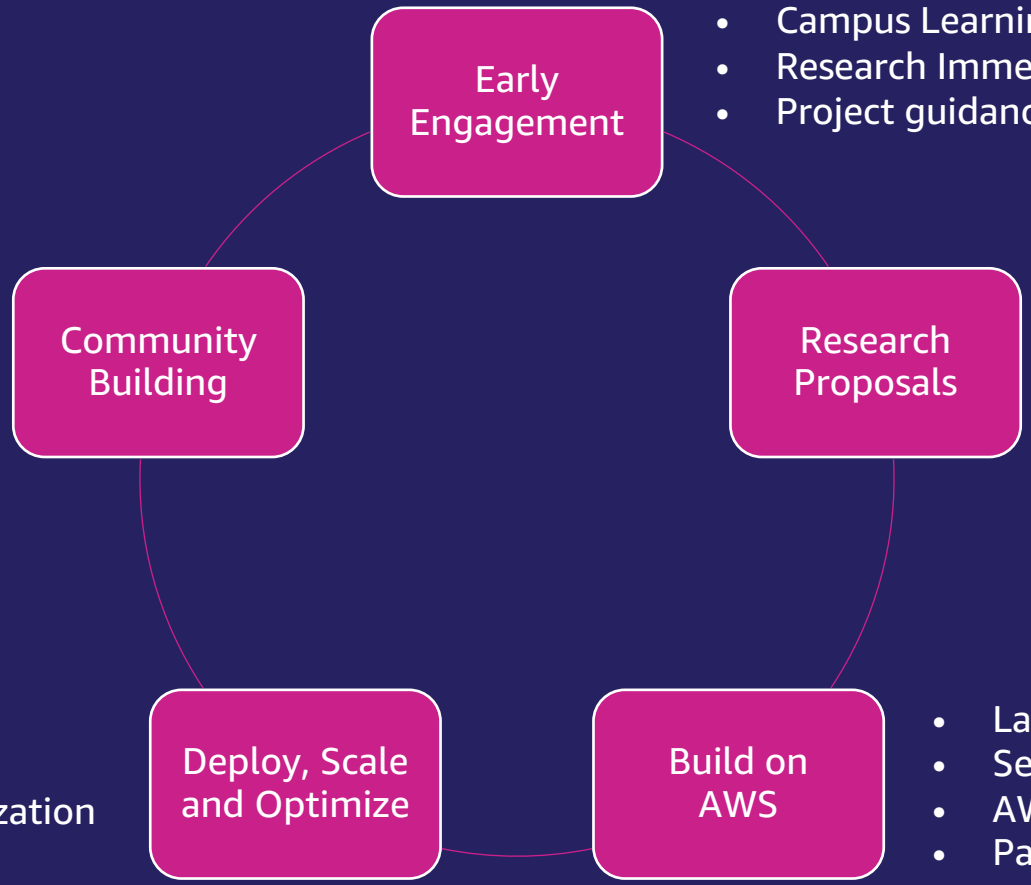
AWS provides a solution

Enables deployment of repeatable research environments that enable collaboration, efficiency, and security.



Working with researchers

How we work with researchers



- Campus Learning Days
- Research Immersion Day
- Project guidance

- Cloud architecting
- Proof of Concept assistance
- Letters of Collaboration

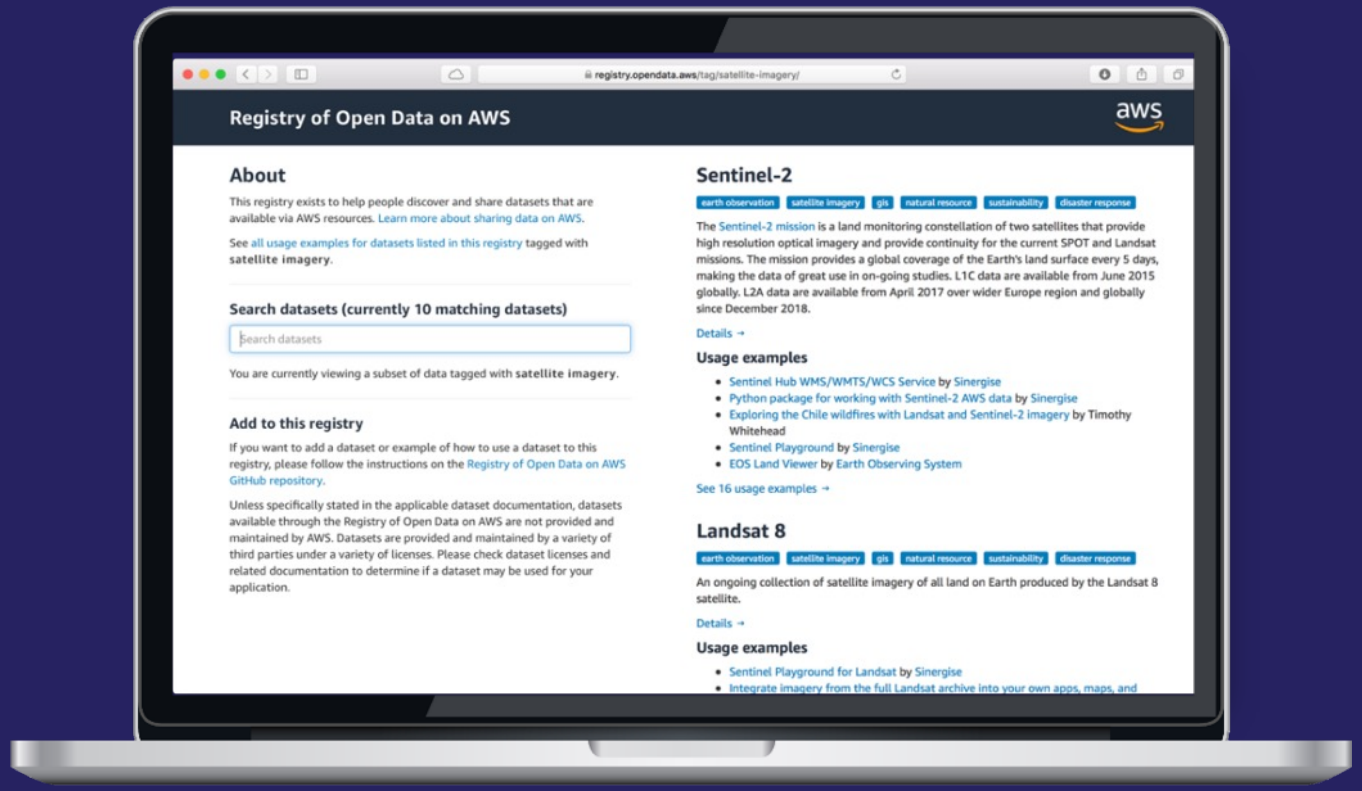
- Landing Zones, Account Vending, SSO
- Secure Infrastructure: CMMC, HIPAA, etc.
- AWS Training
- Partner solutions

- Deployment guidance
- Performance/Cost optimization

- Workshops/Research Days
- Research Summits Talks
- Articles/Blogs/Webinars
- Customer Stories

Open data on AWS

AWS hosts a variety of public datasets to lower the cost and improve the speed of research.



Examples include:

- 1000 Genomes Project
- The Cancer Genome Atlas
- International Cancer Genome Consortium
- Landsat 8
- Common Crawl
- SpaceNet
- OpenStreetMaps
- …and more

https://registry.opendata.aws/

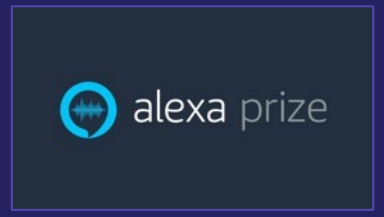
Amazon’s Investment in Research

Academics at Amazon

We hire world-class academics as Amazon Scholars and Amazon Visiting Academics to work on large-scale technical challenges, while they continue to teach and conduct research at their universities.

Collaborations

Whether you’re a faculty member, student, developer, thought leader or a policy maker, Amazon offers a number of ways for you to engage with the company’s science community.



aws

Programs & Collaborations

Letters of Collaboration / Grant PoC Support

Amazon Science: Funding, Amazon Scholars, Internships

Cloud Credits for Research

Global Data Egress Waiver

Amazon Machine Learning Solutions Lab

Amazon Quantum Solutions Lab

AWS Data Lab

NSF Cloudbank

NIH STRIDES

Thank you!

Scott Friedman, Ph.D. (he/him)
scofri@amazon.com