AWS for Research

Stepping into the cloud

University of California, Merced

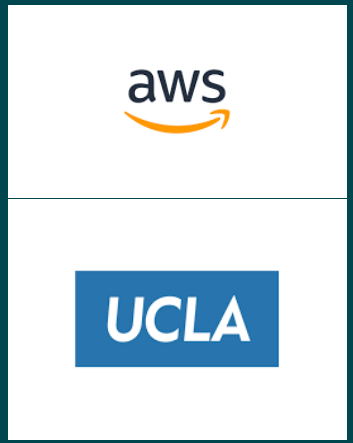
April 20, 2023

Scott Friedman, Ph.D.
AWS Higher Education Research

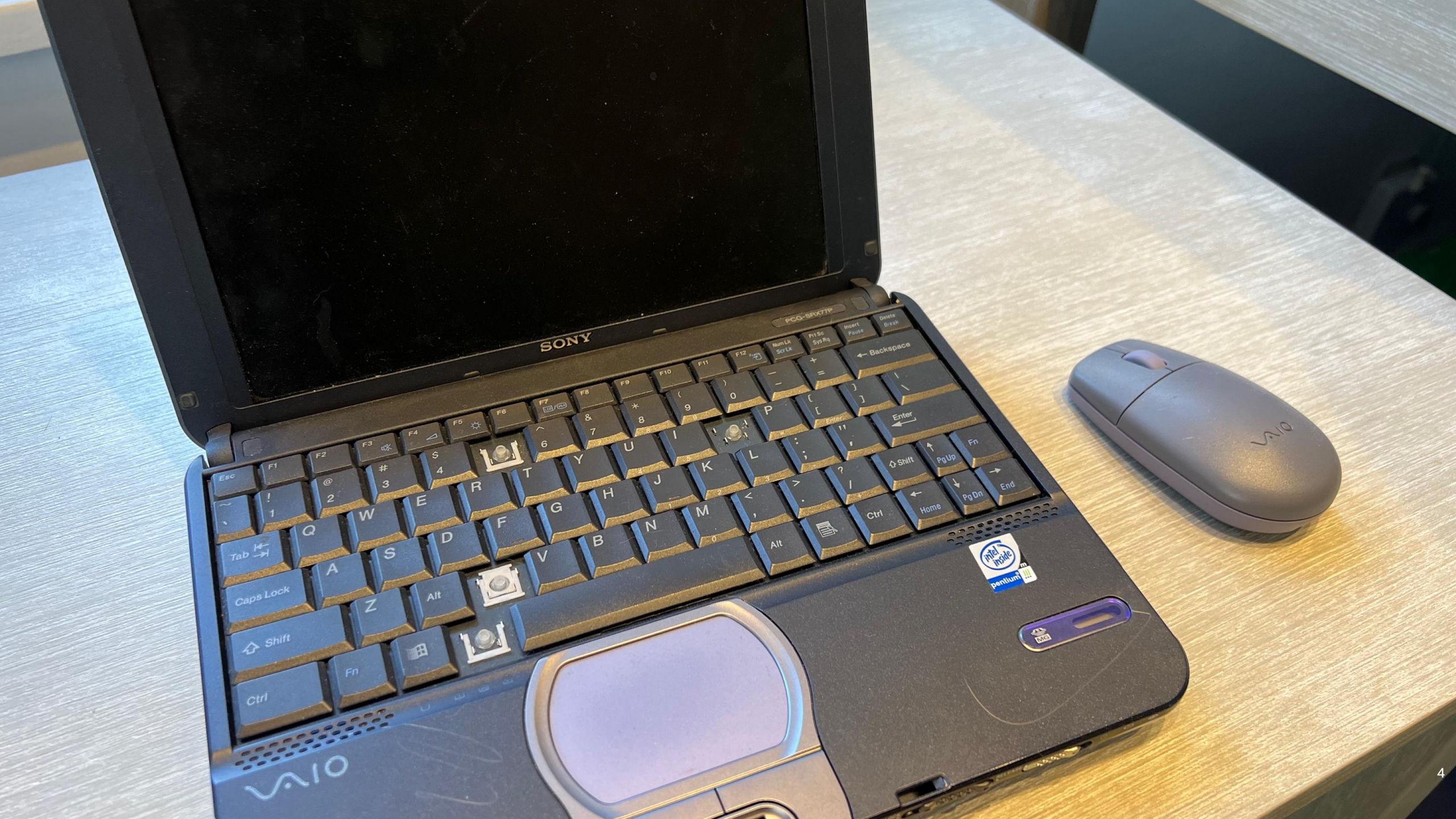
Who?

Scott Friedman, Ph.D.
AWS Higher Education Research

CTO – Advanced Research Computing, UCLA



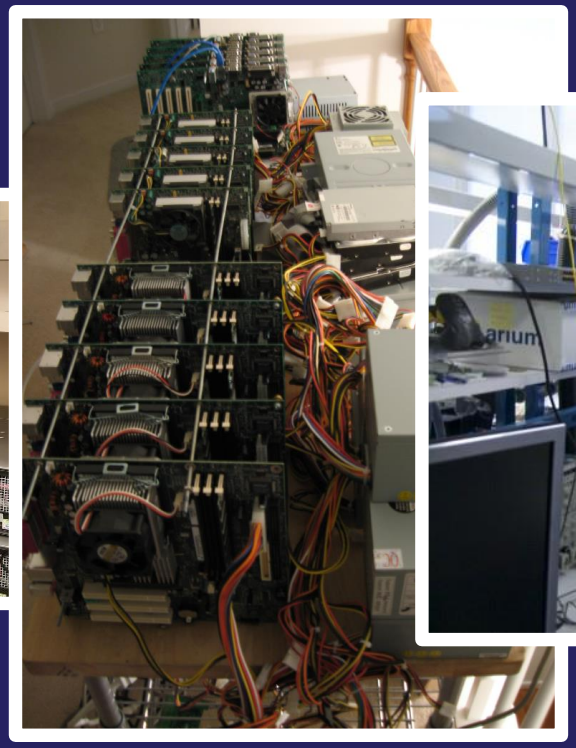
Why?

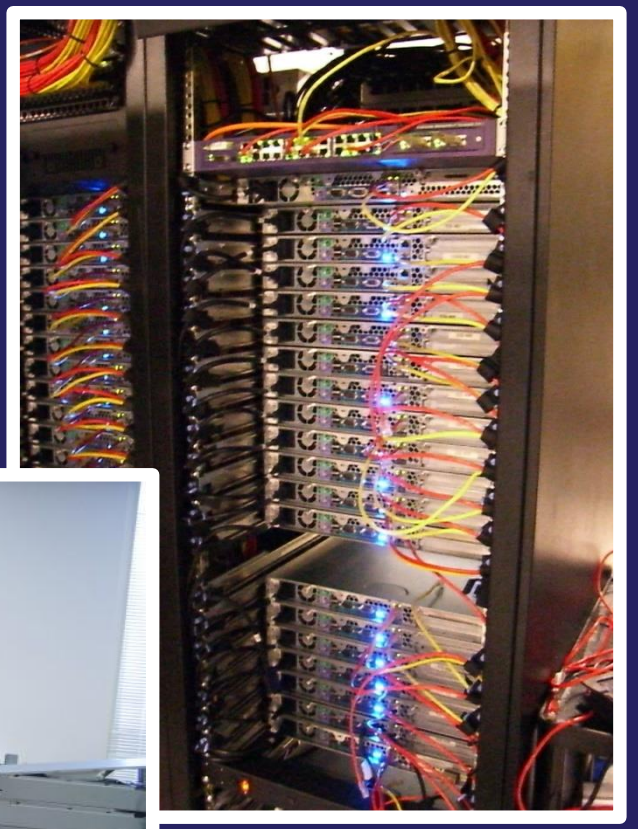

SONY
PCG-SRX77P
VAIO
intel inside
pentium III

Why AWS for research computing

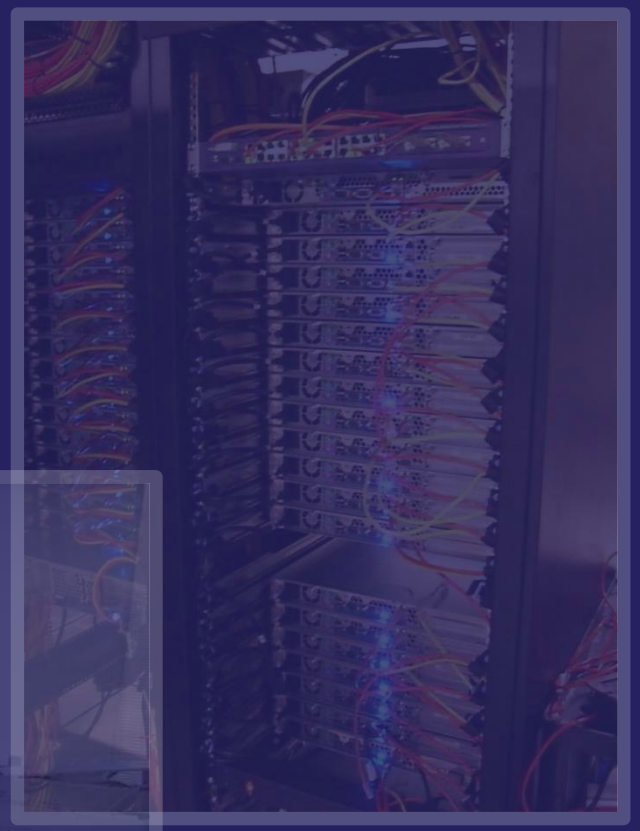
Research computing in higher education

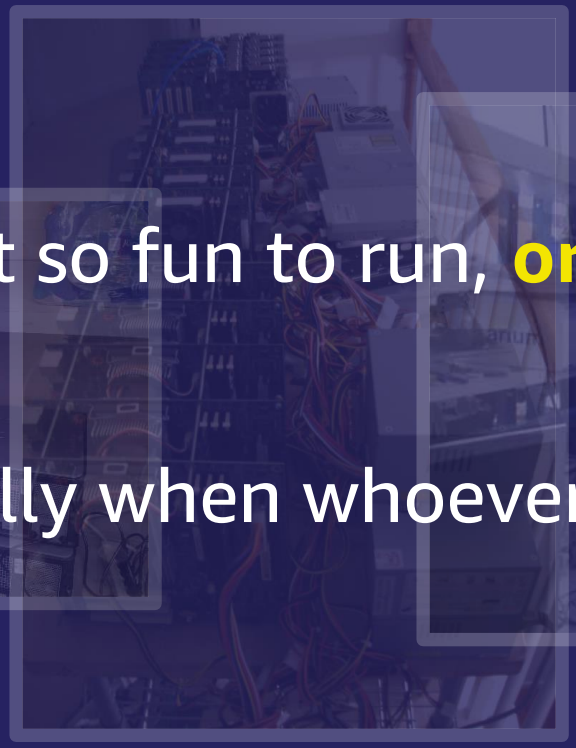
Familiar sights around campus





Research computing in higher education





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

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

Research computing in higher education

Better

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

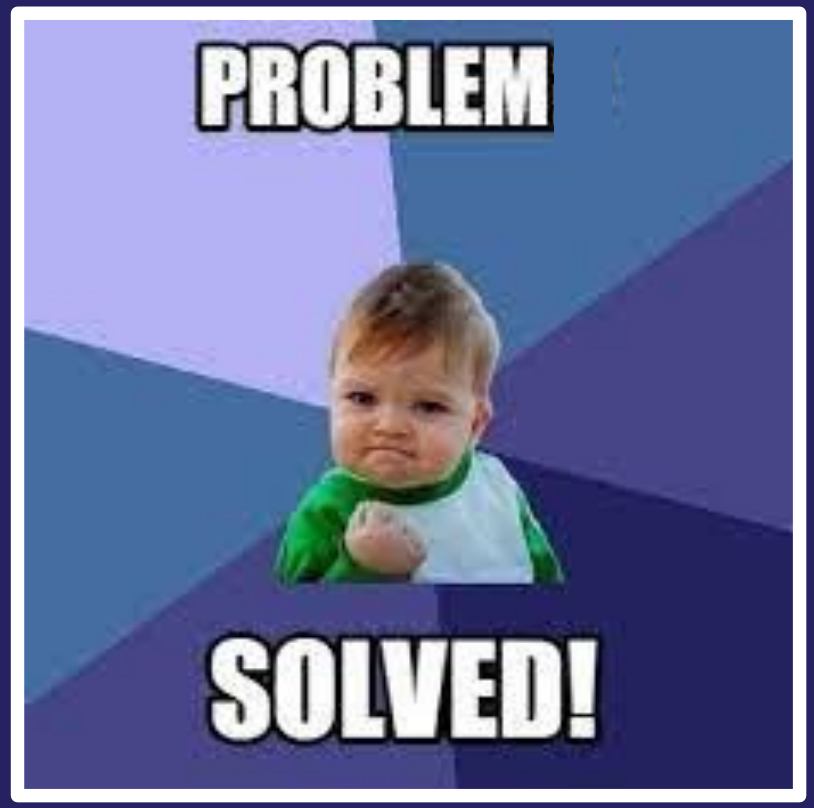
Give / Get

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

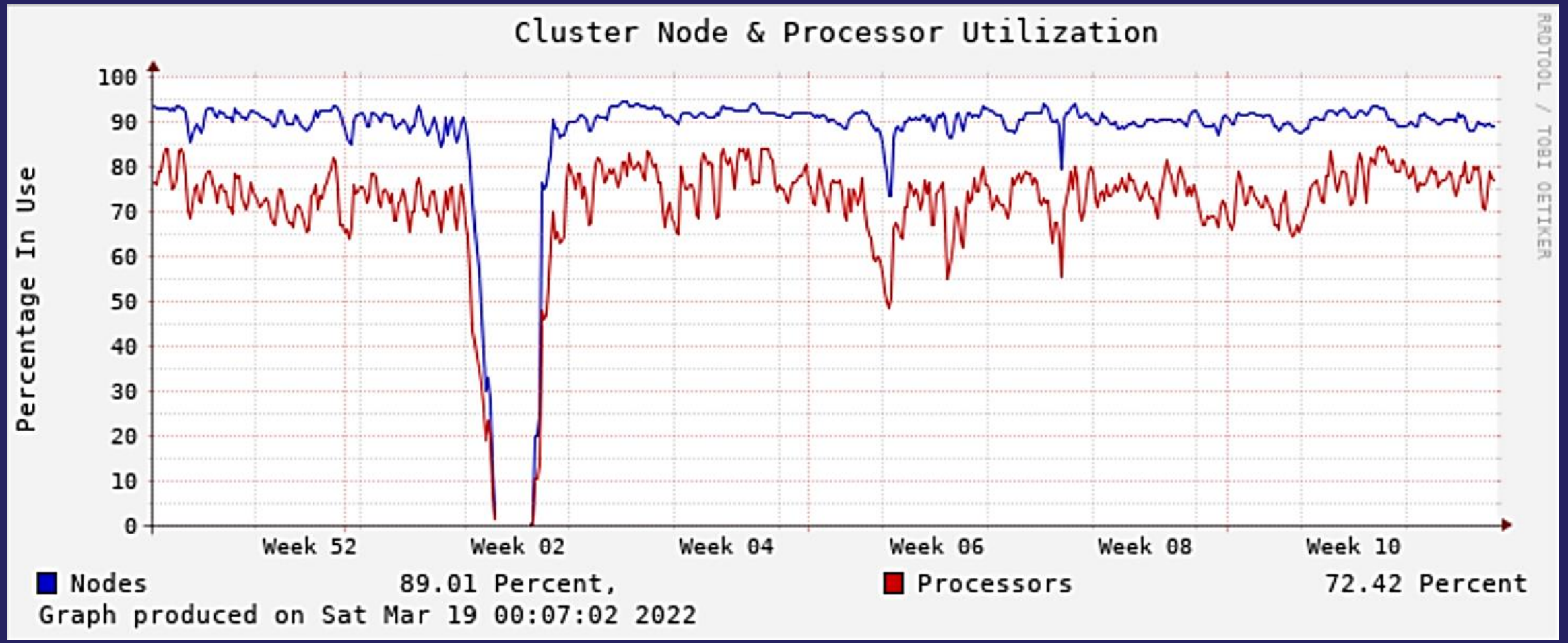
Research computing in higher education

My message to you?

USE IT!



Research computing in higher education



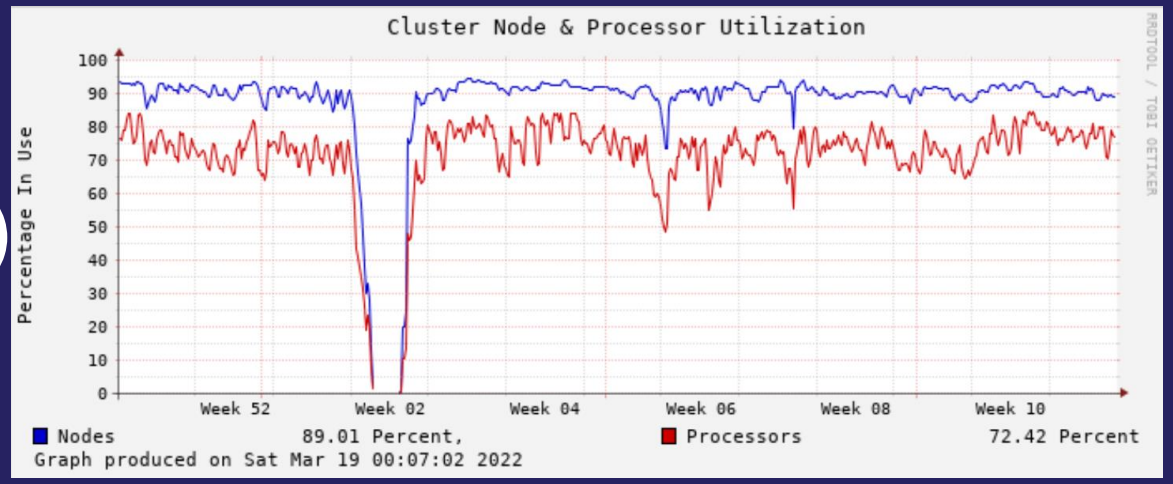
Research computing in higher education

Familiar lab/campus system

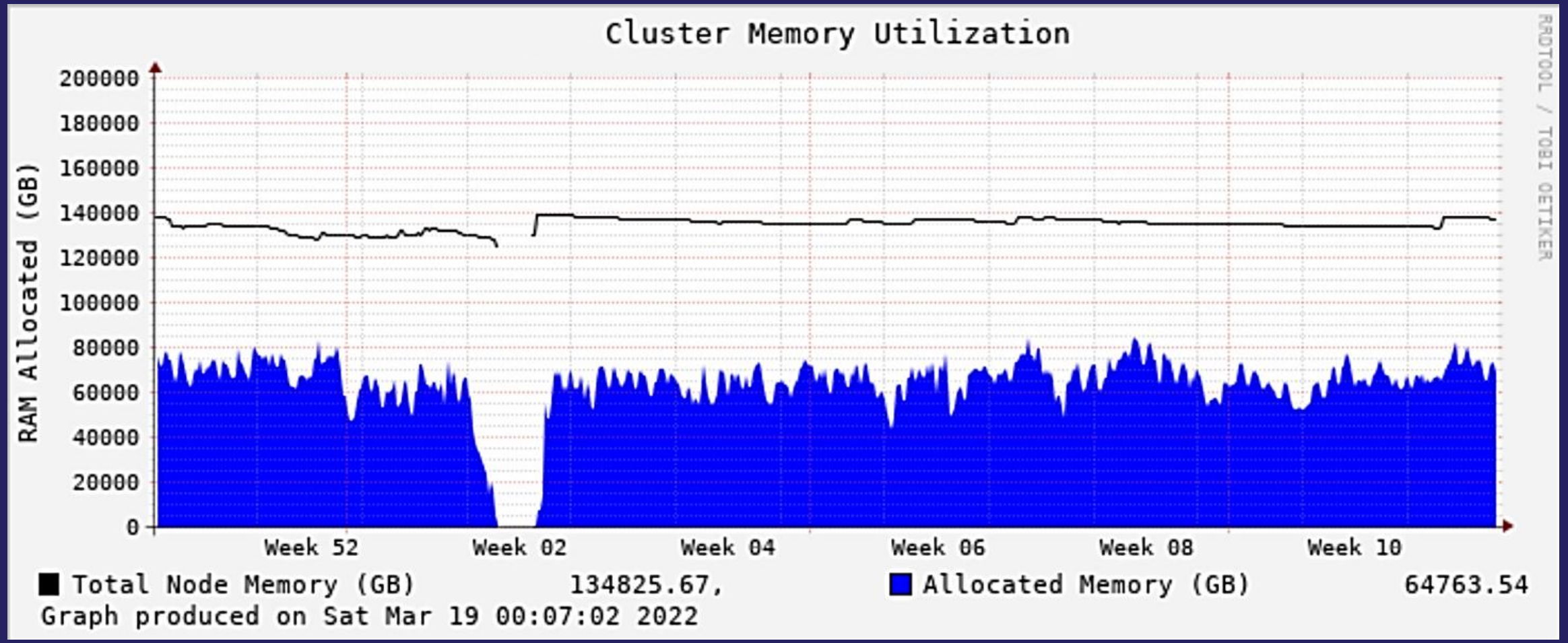
- Reasonably high utilization capacity system (HPC + HTC)
- From a group/institutional perspective – has benefits
- Aggregate sustained usage

On average (even with maintenance)

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



Research computing in higher education



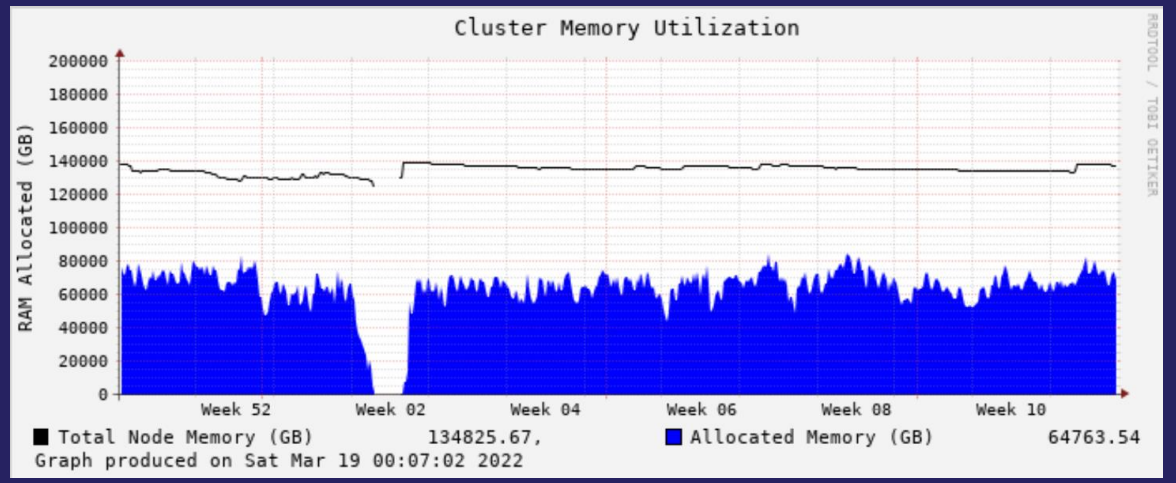
Research computing in higher education

Familiar challenge

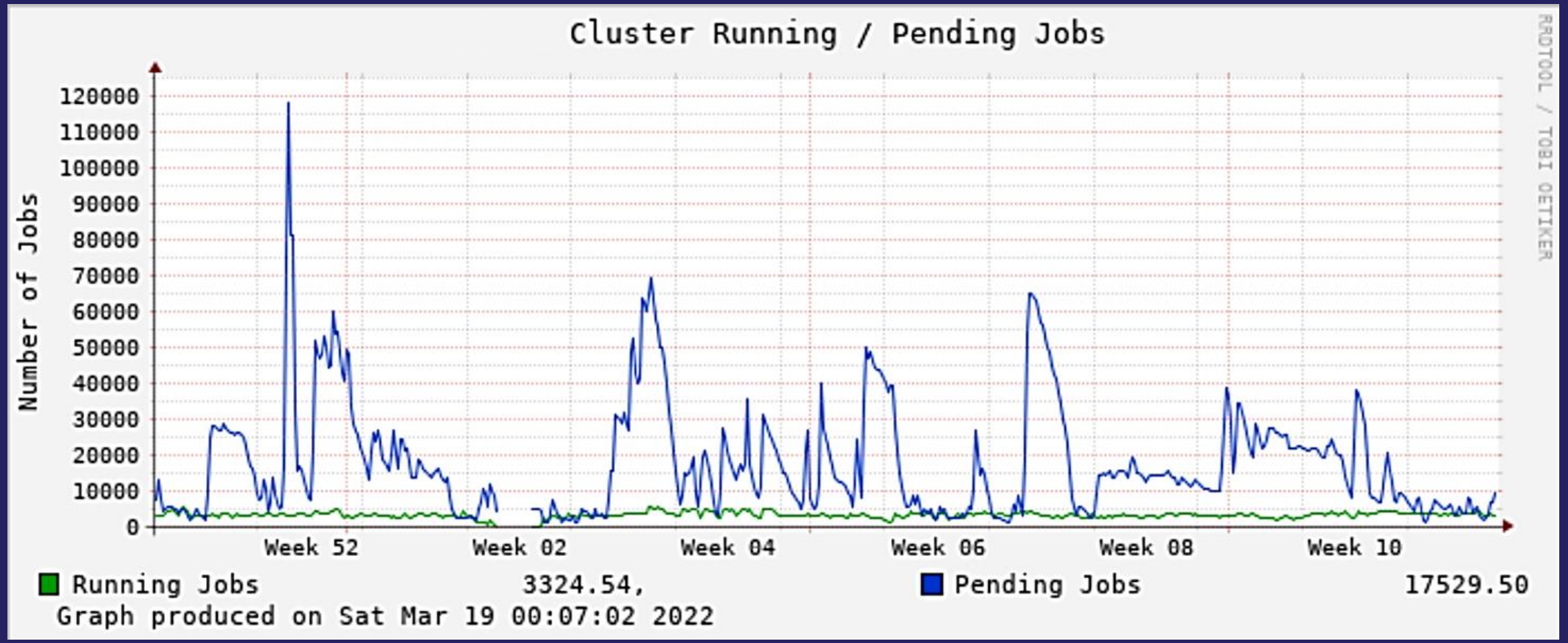
- Individual system resources end up over- or under-provisioned
  - CPU cores, memory, GPU, network
- Here memory is over-provisioned by 100%
  - 100% of the time (yes, extreme)

Do your best

- Purchase decisions made a priori
- Art more than science



Research computing in higher education



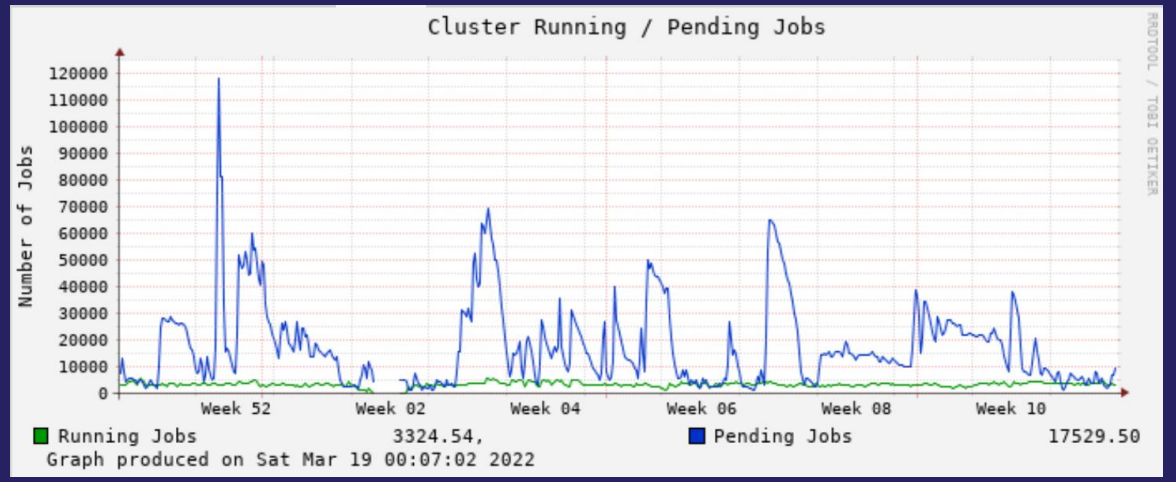
Research computing in higher education

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 in higher education

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 in higher education

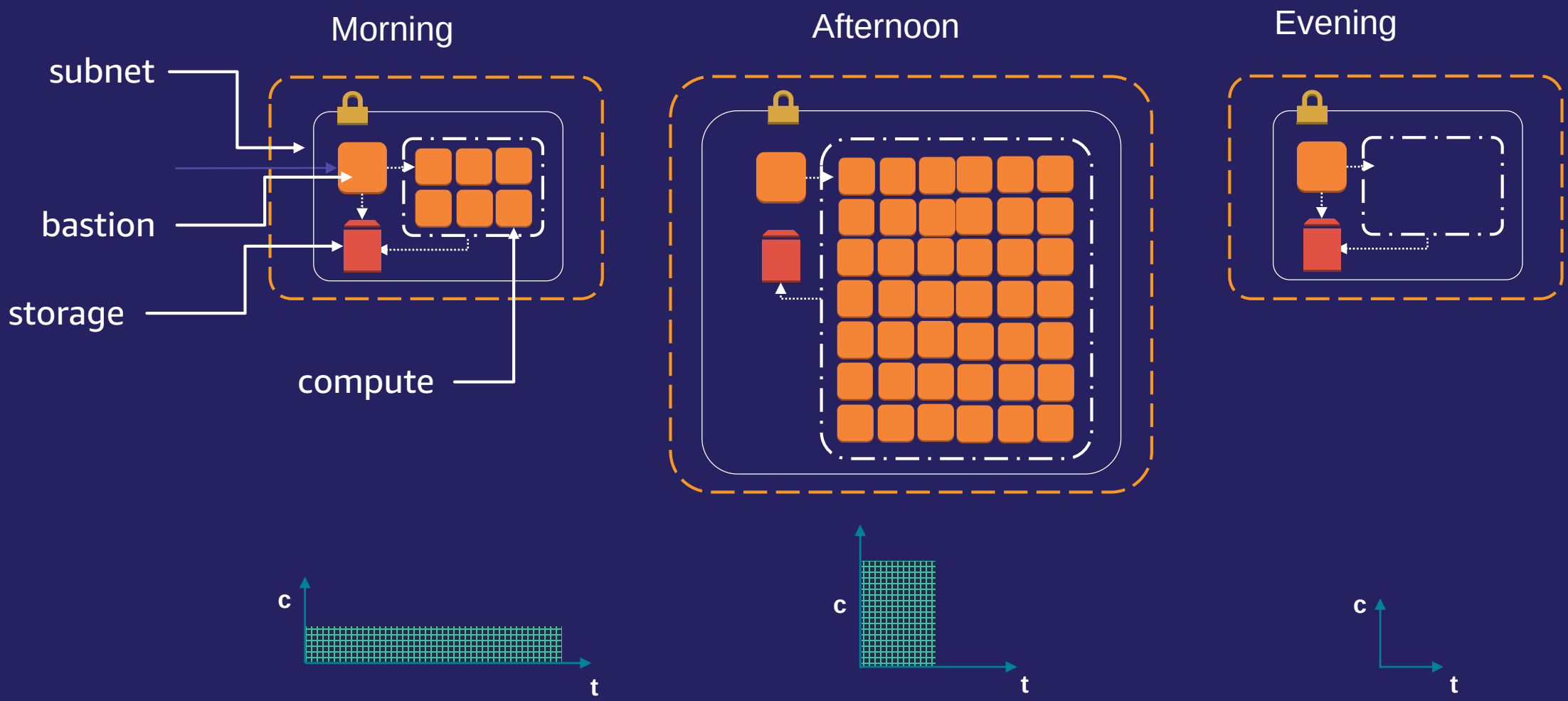
Lab/campus cluster – just fine, until . . .

- Interactive applications on batch system
- Any node type you want as long as it's what is available
- Accelerators – what accelerators?
- Long-running applications, databases, portals
- Extensive waiting

Researchers want to do research

How does AWS enable Research Computing?

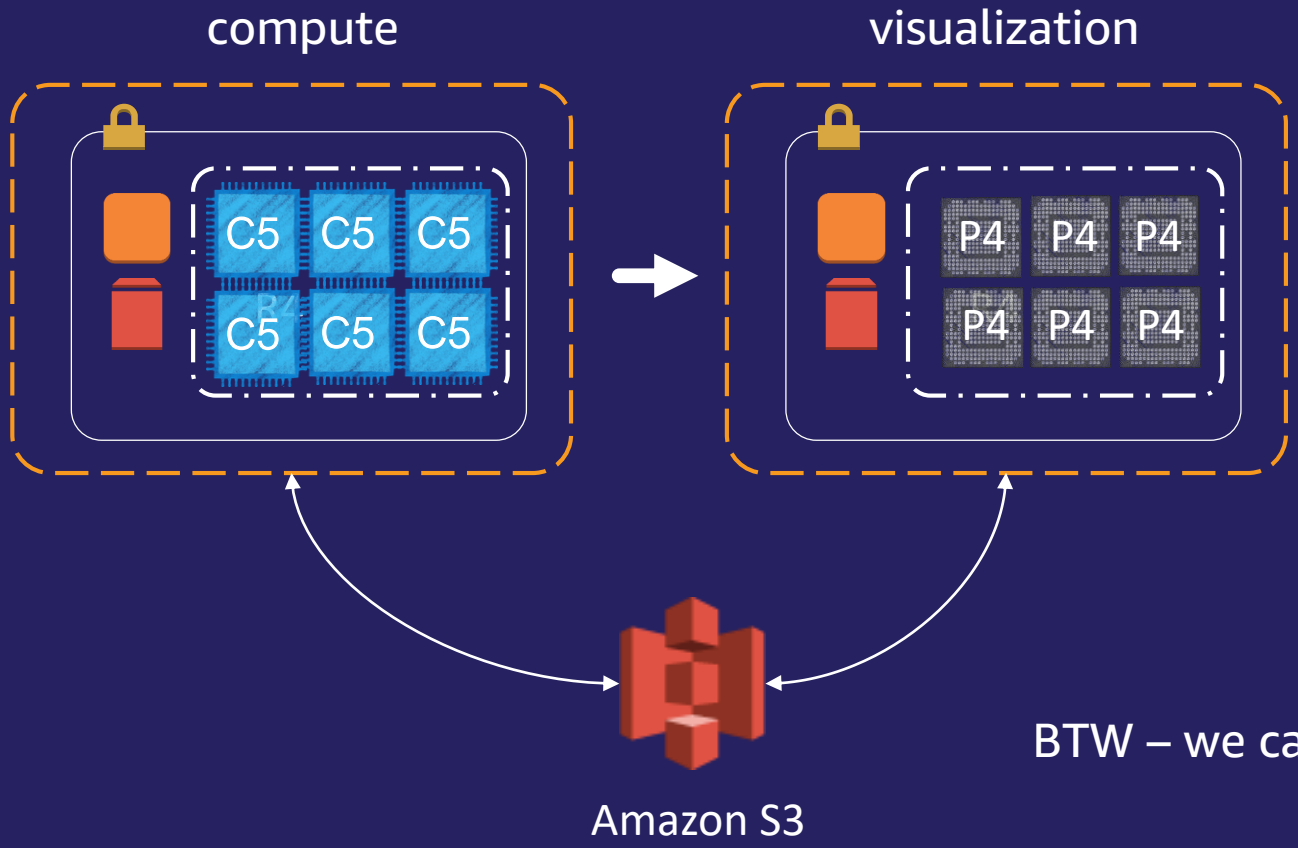
Compute and Storage can be Adjusted Dynamically

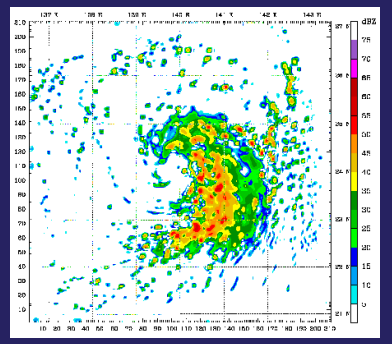


Goldilocks FTW!

Scale resources to be “just right”

Compute and Storage can be Fit for Purpose





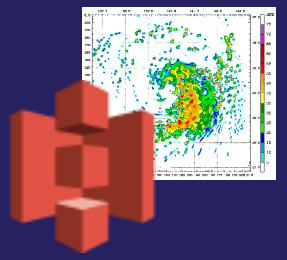
BTW – we can dial-in CPU and memory as well

Be the Snowflake

Deploy exactly what you need, when you need it

Compute and Storage can be Ephemeral

> poof <



Amazon S3

Not Using? Not Paying!

Only pay for what you use – only while you use it

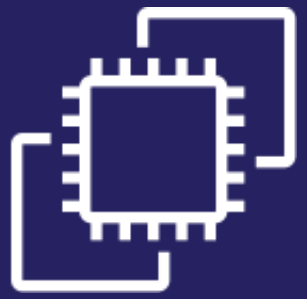
Compute and Storage can be Available on Your Schedule

Skip to the head of the line

You don’t even need to know anyone

AWS compute for research

Compute paradigms



Amazon EC2

Traditional Virtual, Bare Metal, and Accelerated computing

Amazon ECS, EKS, Fargate, and Batch

Container orchestration and execution

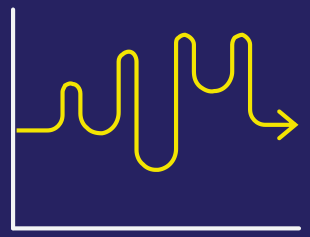
AWS Lambda

Serverless compute

Multiple pricing 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

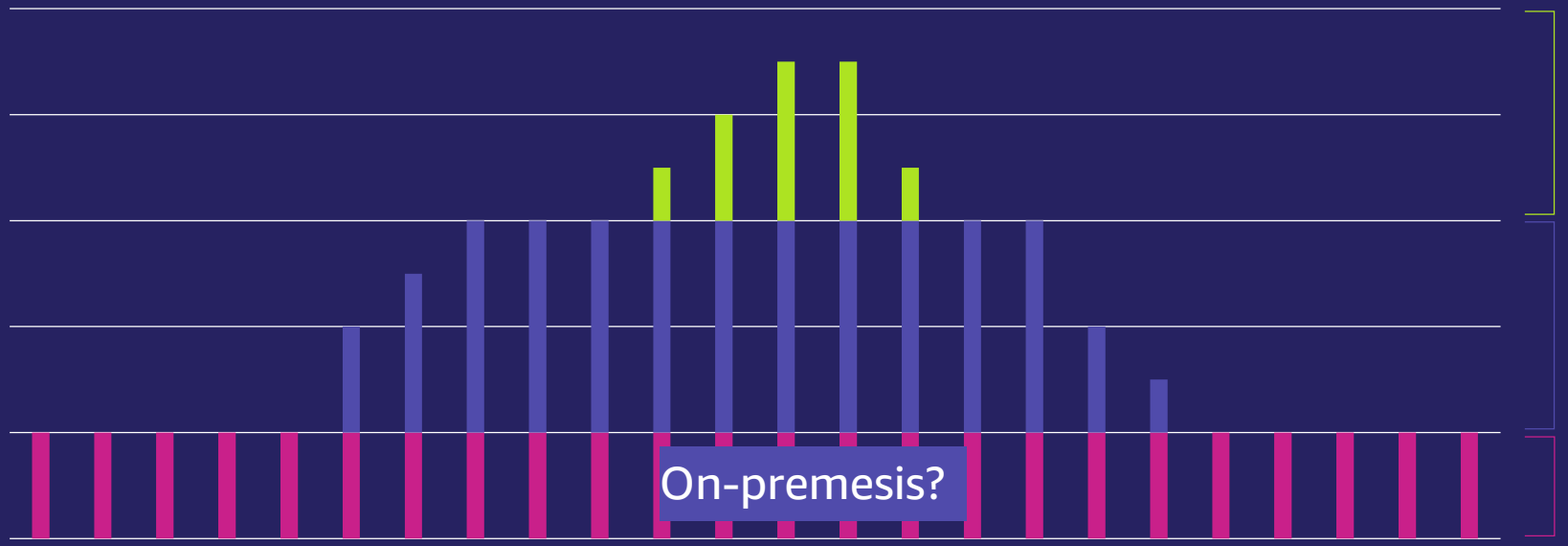
Spot

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

Fault-tolerant, flexible, stateless workloads

The best practice is to combine all purchase options

How these options interrelate



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

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

USE **RIs** AND **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

Compute platform choice

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

300+

INSTANCE TYPES

for virtually every workload and research need

HPC Research Computing great, but not for everyone

HPC

Everything else

How to avoid complicated

Lightsail for Research

Key Features

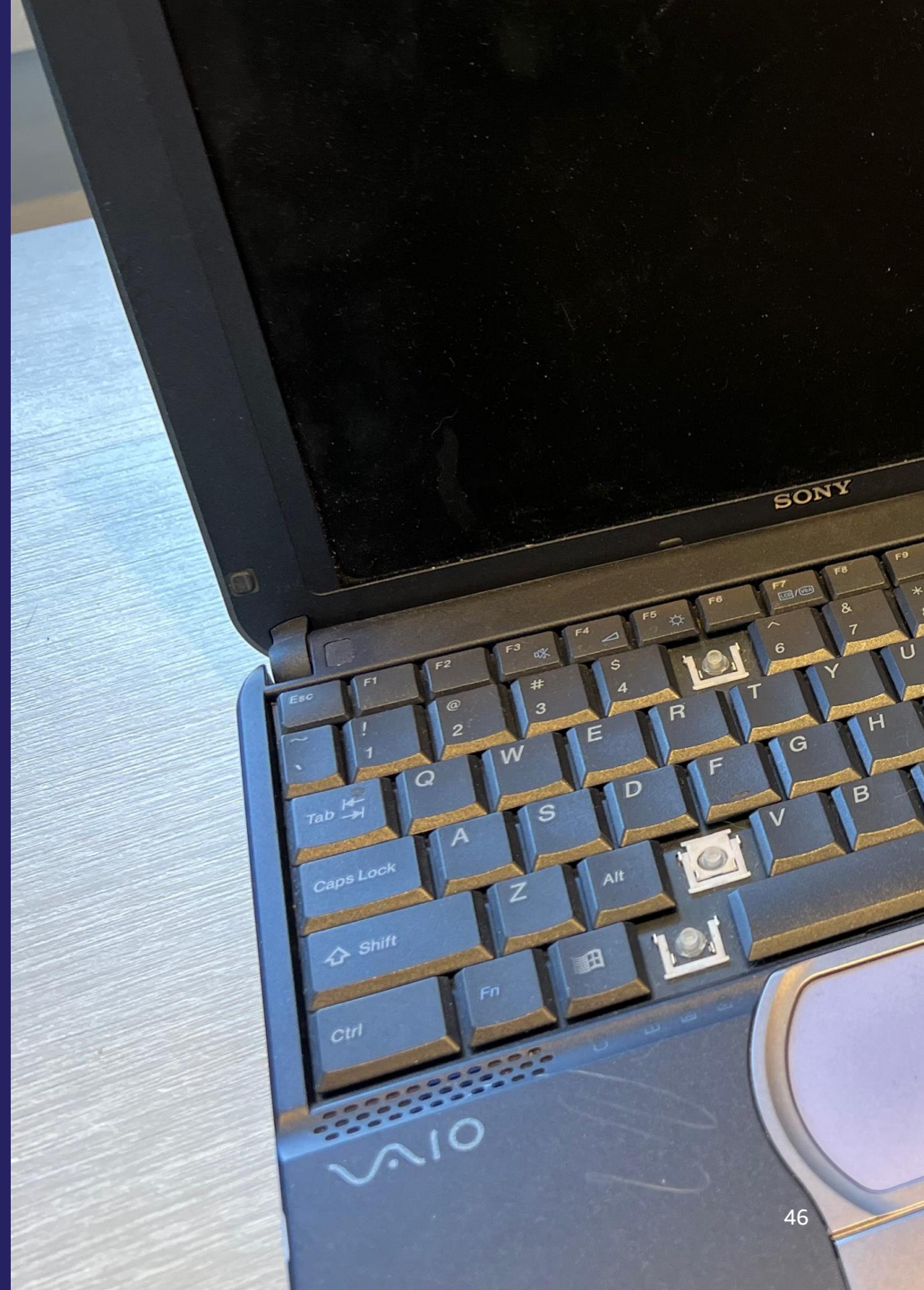
- Lightsail for Research is an AWS service
- There is nothing for customers to install
- Requires no cloud or IT skills to get started
- Simple to explain, understand, and use
- Offers bundled pricing, makes spending clear up front
- Has built-in cost controls, saving customers money

Why Lightsail for Research?

Researchers

Research lives on laptops…

- Papers & proposals written
- Data stored & visualized
- Analyses performed
- Collaboration platform

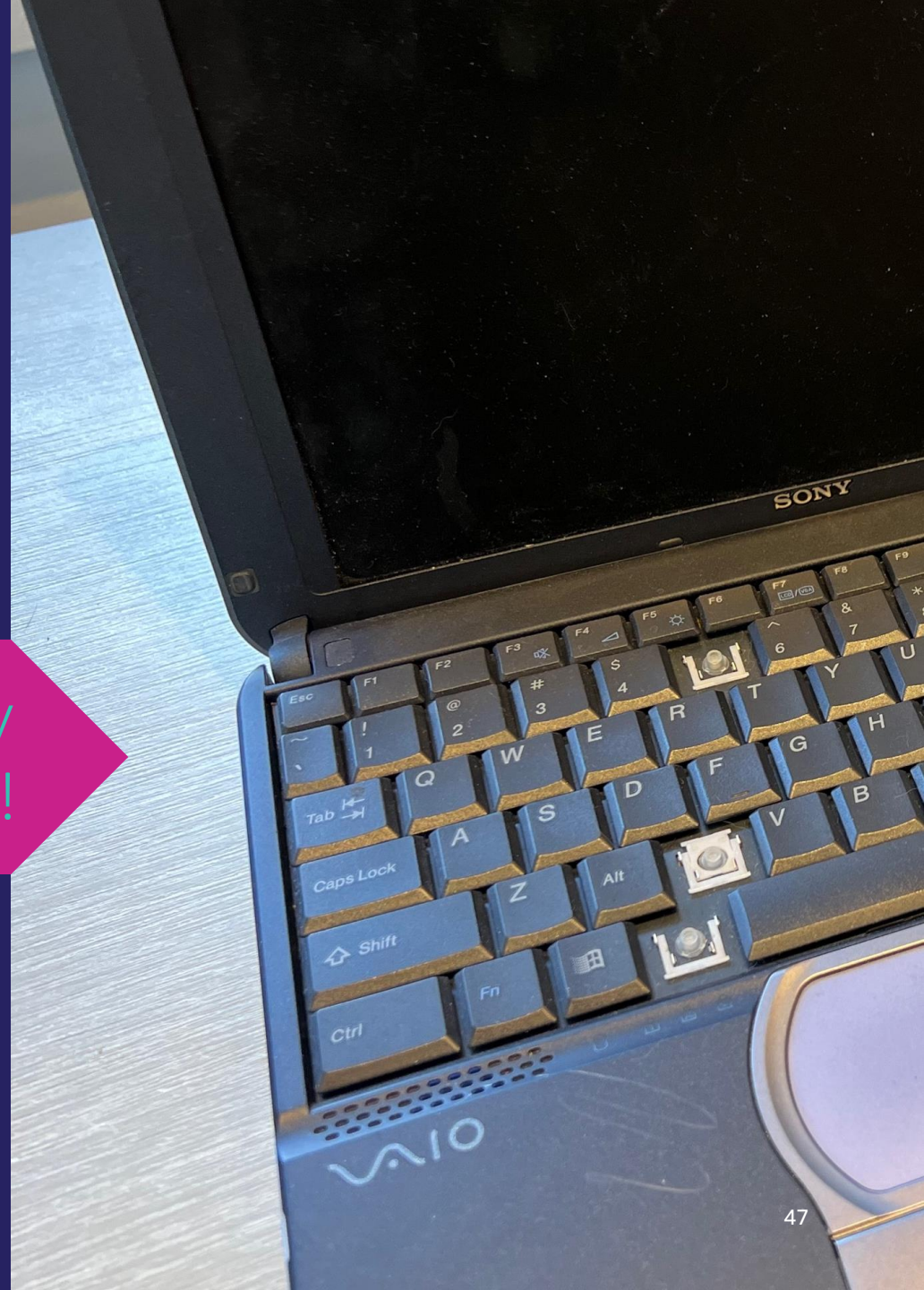


Researchers

Research lives on laptops…

- Papers & proposals written
- Data stored & visualized
- Analyses performed
- Collaboration platform

Increasingly demanding!



Researchers

Modern research needs more…

- Compute: speed/cores/GPU
- Memory: limited problem size
- Time: long running/multiple analyses

Researchers

Most researchers exist in an IT abyss…

Little research IT support and enablement

Want…

to focus on their research

simple access to resources

They are budget conscious

Value to Central/Research IT

Central and Research IT

Quick demo

Platform as a Service

Need to Scale?

python

Dask builds on the existing Python and

What is Dask?

DASK

Dask helps data scientists and data

What is Coiled?

Coiled

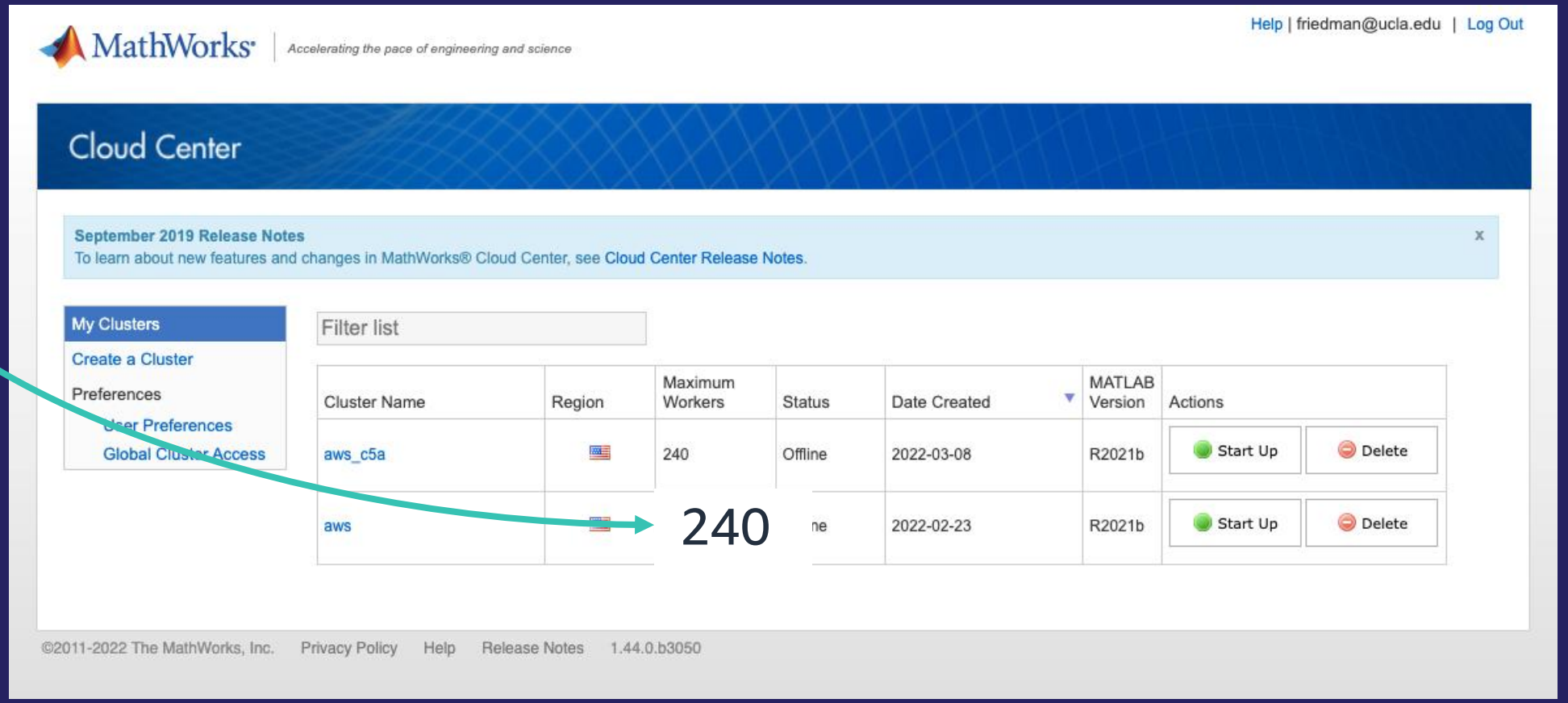
Coiled helps organizations adopt Dask ***in AWS***

aws partner network

MATLAB + Cloud Center

Create cloud parallel pool* from your laptop

What?



https://cloudcenter.mathworks.com/

* Parallel Computing Toolbox Required

Research Storage

The universe of research data and its challenges

Growing Exponentially

From new sources

Increasingly diverse

Used by many researchers

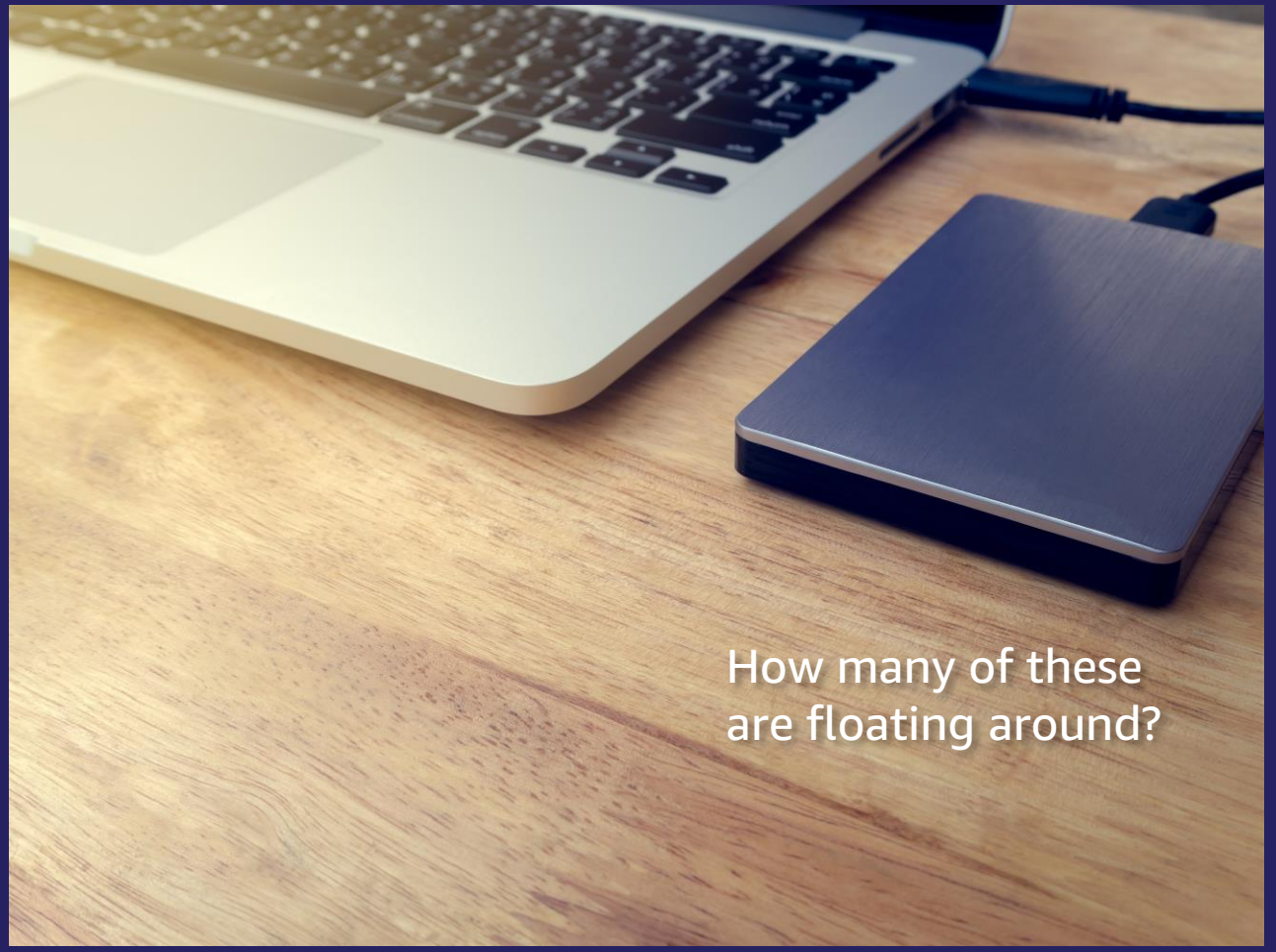
Analyzed by many applications

Administrative overhead

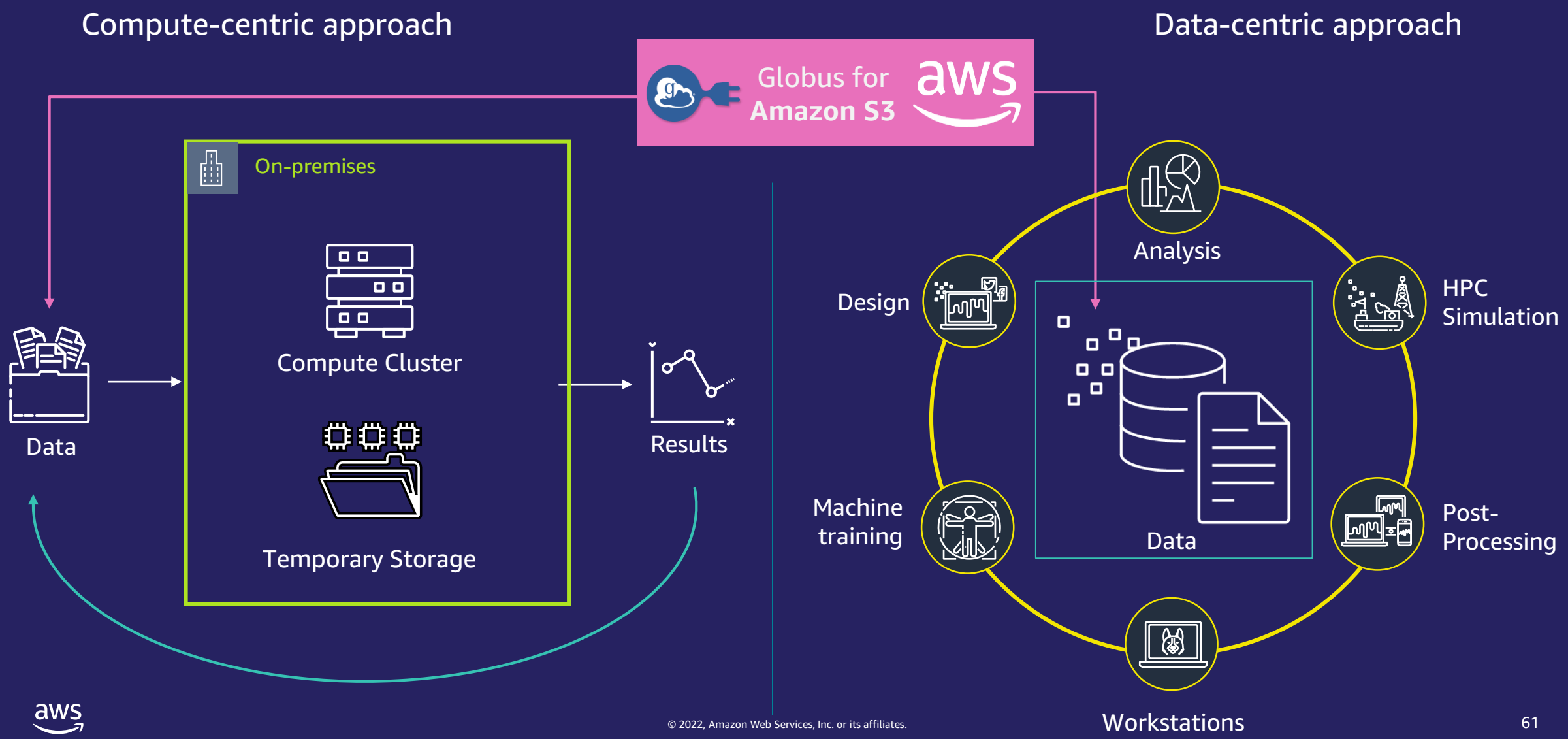
Lack of scalability

Lack of Agility

...and then, Data Silos

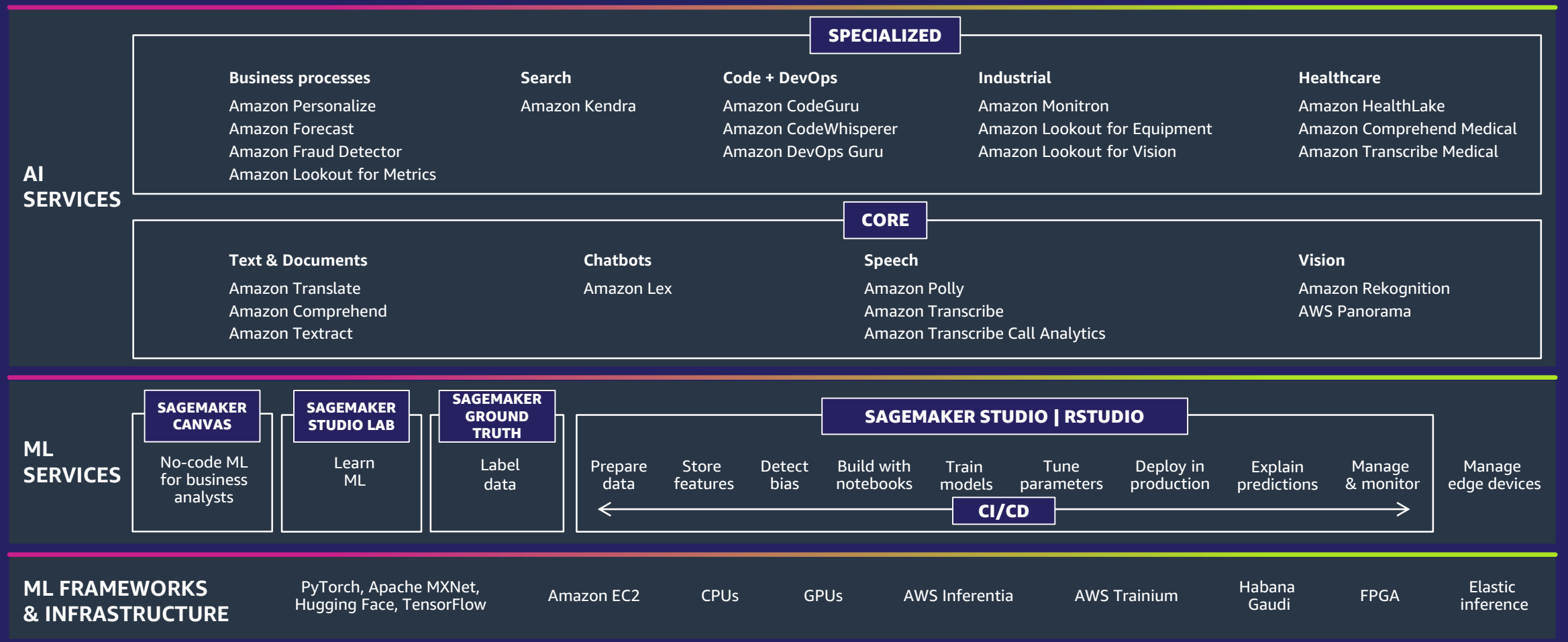


AWS enables a data oriented approach to research



Machine Learning

Broadest and Most Complete Set of ML Capabilities

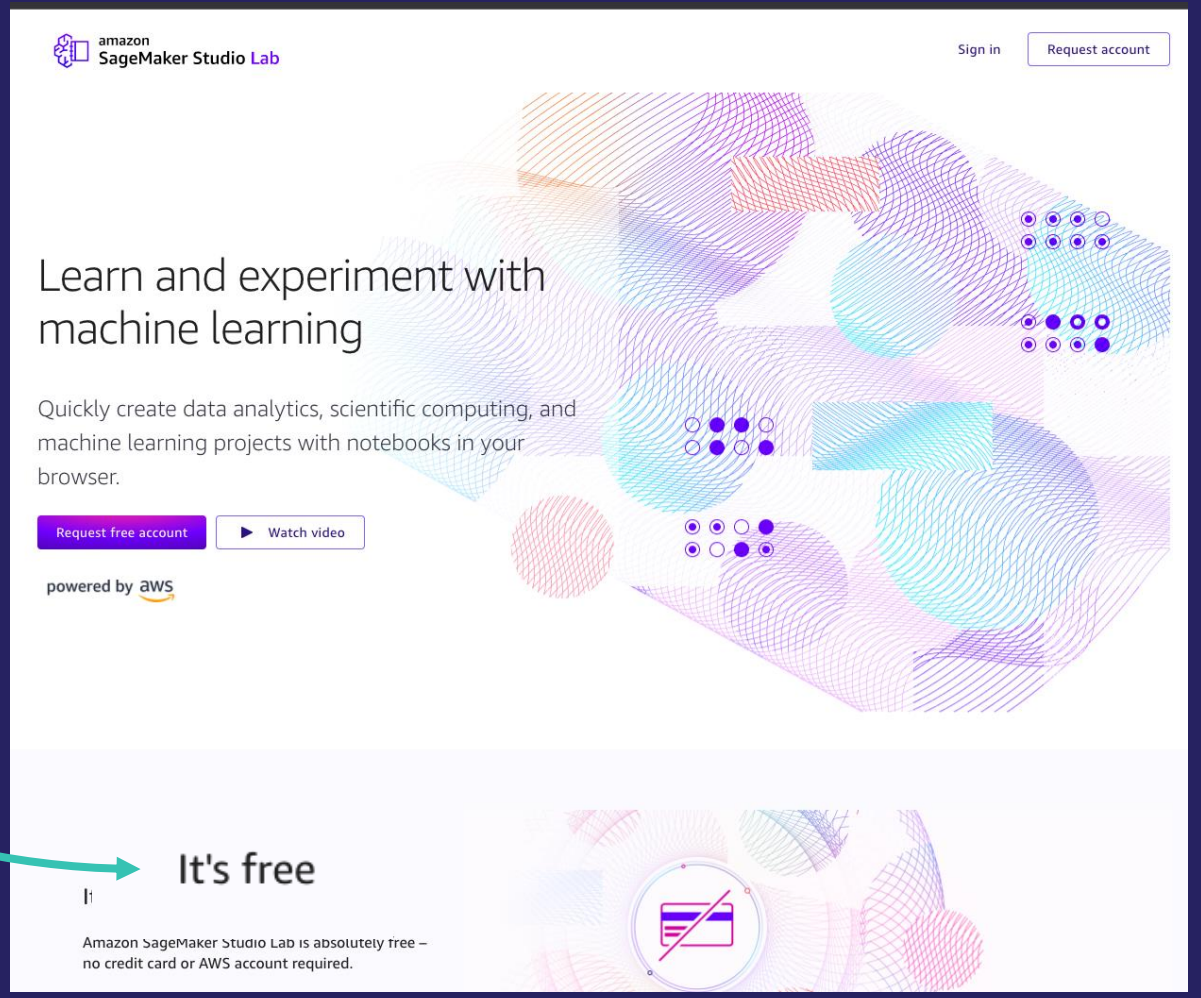


Sagemaker Studio Lab

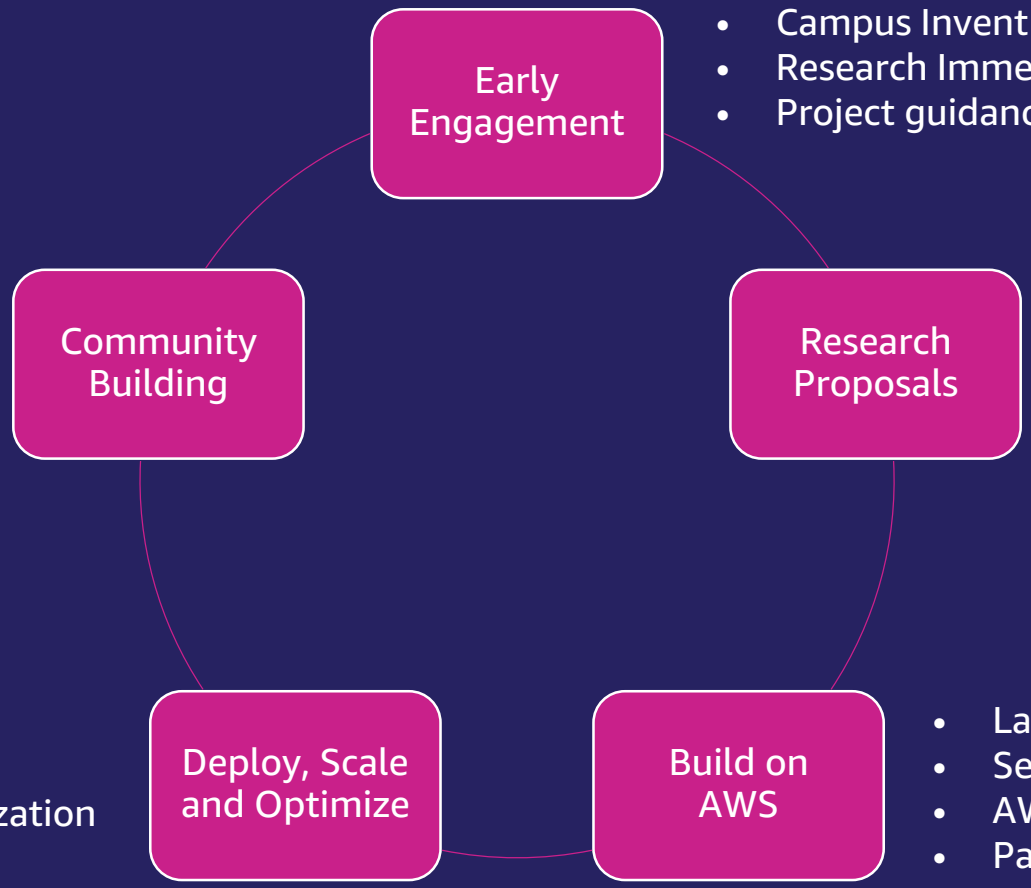
- Notebooks? Yes!
- GPUs? Yes!
- This?

- Sessions: 12h CPU, 4h GPU
- 16GB RAM, 15GB Persistent Storage

https://studiolab.sagemaker.aws/



How we work with researchers



- Campus Invent
- 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

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

scofri@amazon.com