Introduction to Amazon Cloud

Amazon EC2 and Spot Overview

Patrick Guha

Solutions Architect
Amazon Web Services

Agenda

- Introduction to Amazon Cloud
- AWS Global Reach
- Amazon EC2 Overview
- Amazon EC2 Spot Overview

What is cloud computing?

Cloud computing is the on-demand delivery of IT resources and applications over the Internet with pay-as-you-go pricing.

What is AWS?

AWS provides a highly reliable, scalable, low-cost infrastructure platform in the cloud that powers millions of businesses in over 245 countries and territories around the world.

Benefits

- Low Cost
- Elasticity & Agility
- Open & Flexible
- Secure
- Global Reach

How AWS can help your research

Science, not servers

Use compute when you need it to do large-scale analysis

Collaboration

Access data sets that span institutions

Share effort

Leverage work done by other scientists to save time

Reproduce research

A common platform for reproducing scientific analyses

State-of-the-art analytics

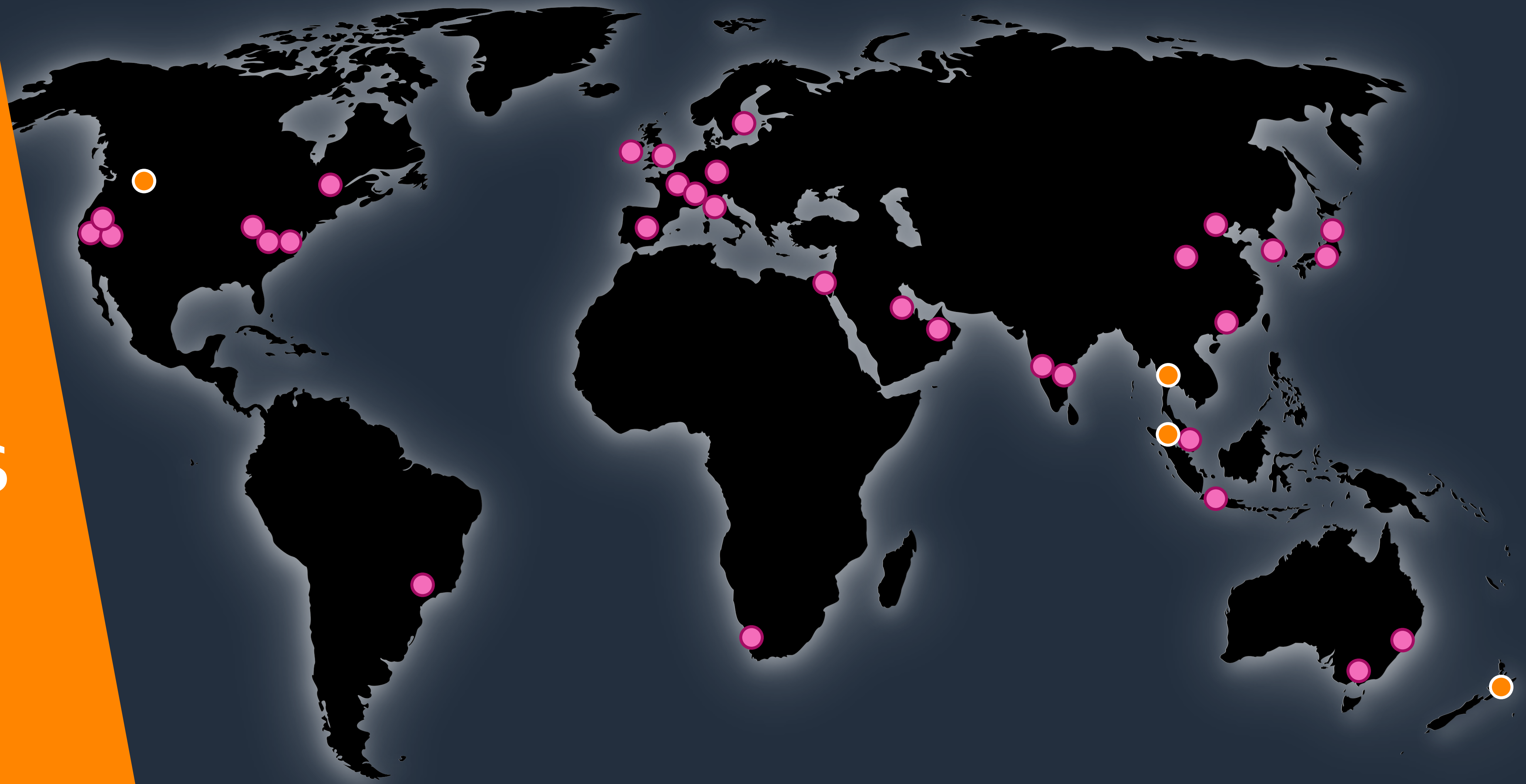
Use data science methods in your research

Security

A collection of tools to protect data and privacy

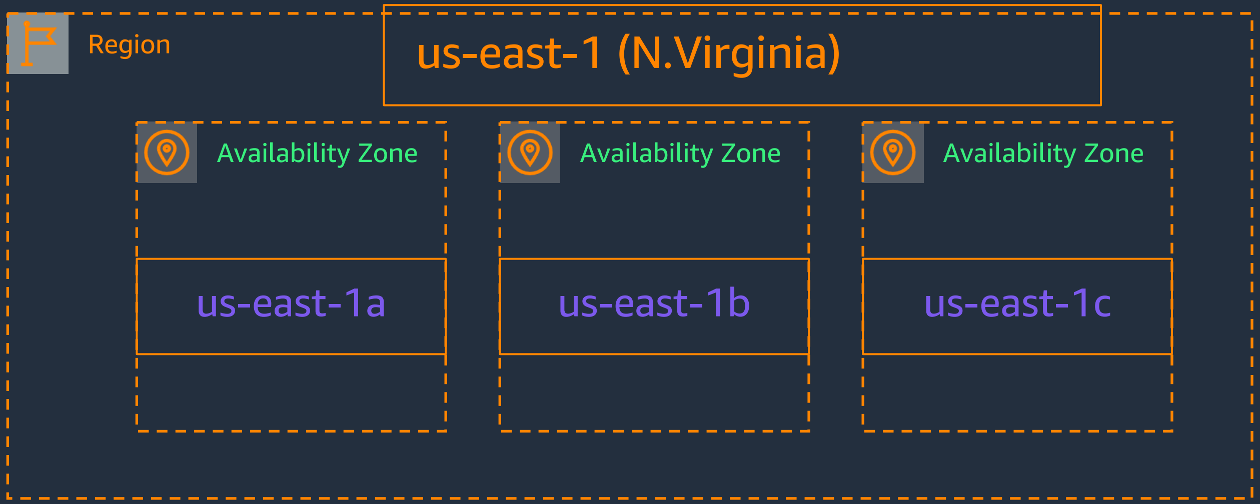
AWS Global Reach

32 Regions



Availability Zones

- Each AWS Region consists of multiple, isolated, and physically separate AZs within a geographic area
- An Availability Zone (AZ) is one or more discrete data centers with redundant power, networking, and connectivity in an AWS Region
- High throughput, low latency (< 10 ms) network between Availability Zones
- All traffic between AZs is encrypted
- Physical separation with 100 km (60 miles)



Region

us-east-1 (N.Virginia)

Availability Zone

us-east-1a

Availability Zone

us-east-1b

Availability Zone

us-east-1c

Amazon EC2 Overview

Amazon Elastic Compute Cloud (Amazon EC2)

Virtual server instances in the cloud

AMAZON EC2

Linux | Windows | Mac

Arm and x86 architectures

General purpose and workload optimized

Bare metal, disk, networking capabilities

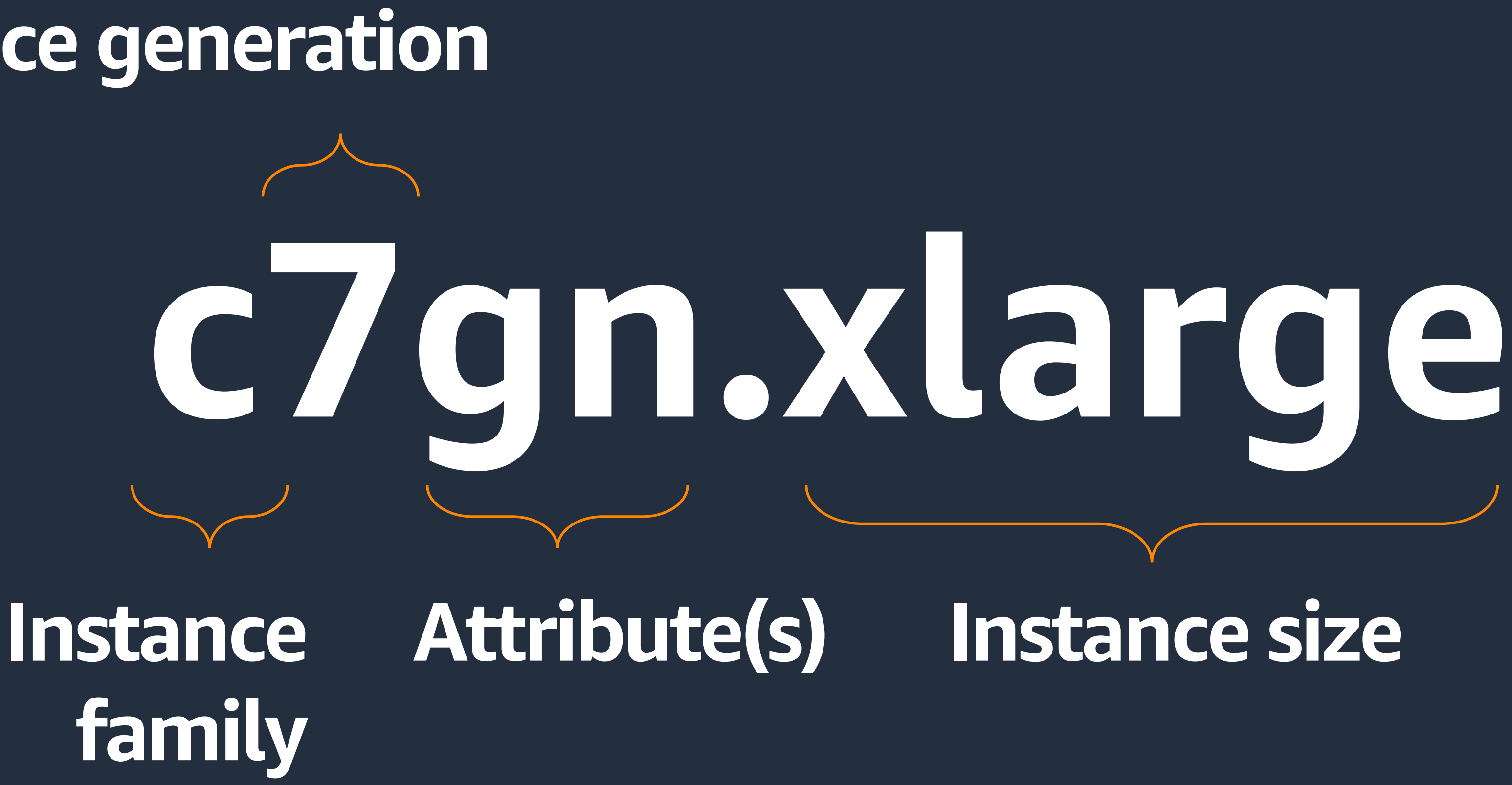
Packaged | Custom | Community AMIs

Multiple purchase options: On-Demand, Spot instances, Reserved Instances, Savings Plans, Dedicated Hosts

Instance Types

	General Purpose		Compute Optimized		Memory Optimized				Accelerated Computing			Storage Optimized		
	Burstable performance	General Purpose	Compute Intensive	Compute + network up to 100 Gbps*	Memory Optimized	In-memory	Memory Intensive	Compute and Memory Intensive	Graphics Intensive	General Purpose GPU	FPGA	High I/O	Dense Storage	Big Data Optimized
intel	T3	M5	C5	C5n	R5	X1	X2iedn		G3	P2	F1	I3en	D3	H1
Local storage (NVMe SSD)		M5d	C5d		R5d			Z1d				I3		
AMD	T3a	M5a			R6a				G5					
metal		M5	C5		R5	u-24tb1		Z1d				I3		
AWS Graviton	T4g	M7g	C7g	C7gn	R7g	X2gd			G5g			Im4gn		

Instance Naming



Instance generation

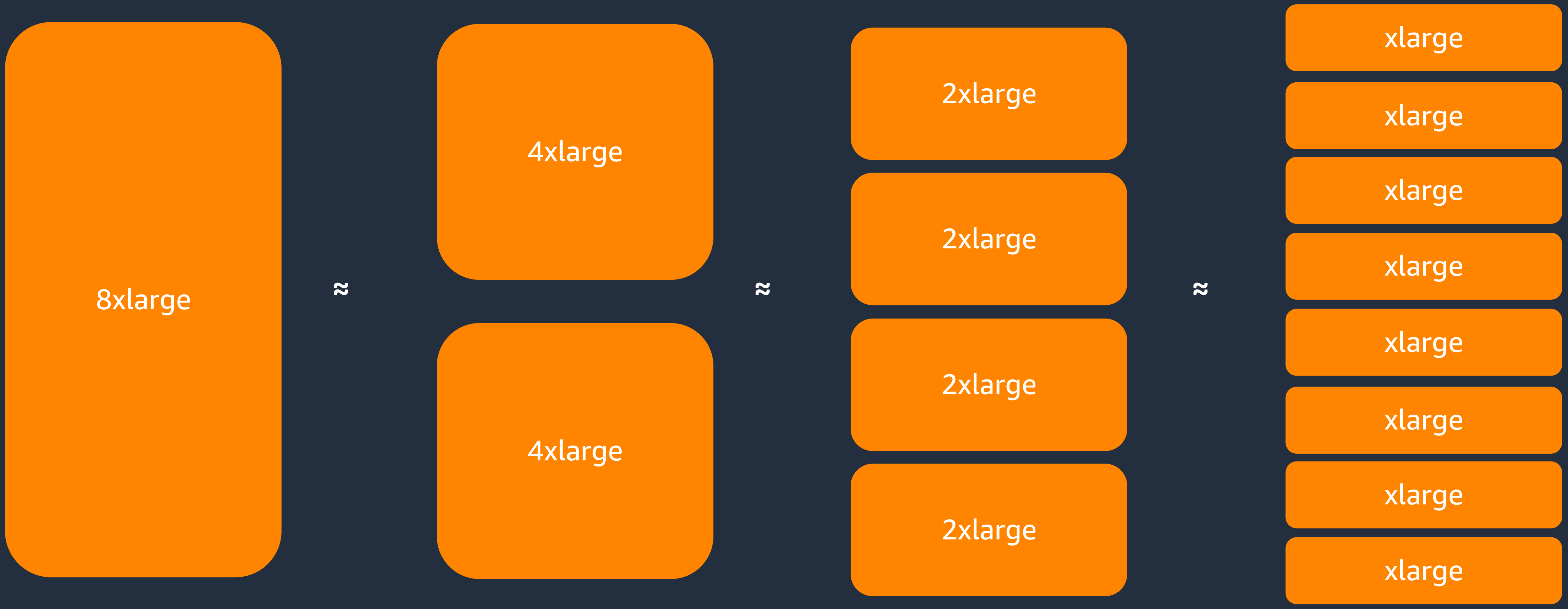
c7gn.xlarge

Instance family

Attribute(s)

Instance size

Instance Sizing



8xlarge

≈

4xlarge

4xlarge

≈

2xlarge

2xlarge

2xlarge

2xlarge

≈

xlarge

xlarge

xlarge

xlarge

xlarge

xlarge

xlarge

xlarge

c7gn.8xlarge

2 – c7gn.4xlarge

4 – c7gn.2xlarge

8 – c7gn.xlarge

Choose your processor and architecture

Intel® Xeon® Scalable (Skylake) processor

NVIDIA V100 Tensor Core GPUs

AMD EPYC processor

AWS Graviton Processor (arm)

FPGAs for custom hardware acceleration

Right compute for the right application and workload

What’s a virtual CPU? (vCPU)

- A vCPU is typically a hyper-threaded physical core*
- Divide vCPU count by 2 to get core count
- On Linux, “A” threads enumerated before “B” threads
- On Windows, threads are interleaved

- Cores by Amazon EC2 & RDS DB Instance type: https://aws.amazon.com/ec2/physicalcores/

* *CPU Optimizing options allow disabling hyperthreading and reduce number of cores*

Memory and Storage

What's a GiB?

- Memory is presented as GibiBytes (GiB) and not Gigabytes (GB)
- 256 GiB = 275 GB

What about storage?

- Storage is independent of compute
- You allocate drives known as Amazon Elastic Block Store (EBS) volumes
- Amazon EBS volumes support up to 64 TiB per volume
- Some instance types provide physically attached (ephemeral) storage

EC2 Operating Systems

- Windows Server 2012/2012 R2/2016/2019/2022
- Amazon Linux (NEW: Amazon Linux 2023)
- Debian
- SUSE
- CentOS
- Red Hat Enterprise Linux (RHEL)
- Ubuntu
- Mac, including M1 Mac instances

Visit the AWS Marketplace for more Operating Systems

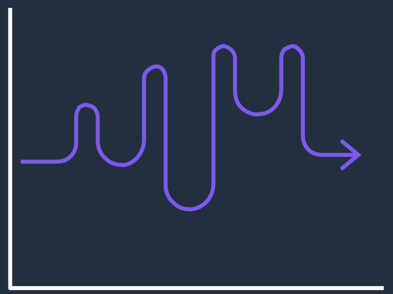
What is an Amazon Machine Image (AMI)?

- Provides the information required to launch an instance
- Launch multiple instances from a single AMI with the same configuration
- An AMI includes the following:
  - One or more Amazon Elastic Block Store (Amazon EBS) snapshots, or a template for the root volume (operating system, applications)
  - Launch permissions that control which AWS accounts can use the AMI
  - Block device mapping that specifies volumes to attach to the instance

Amazon EC2 purchase options

On-Demand

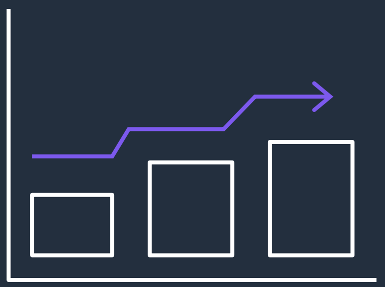
Pay for compute capacity by **the second** with no long-term commitments



Spiky workloads, to define needs

Reserved Instances

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



Committed and steady-state usage

Savings Plans

Same great discounts as Amazon EC2 RIs with **more flexibility**

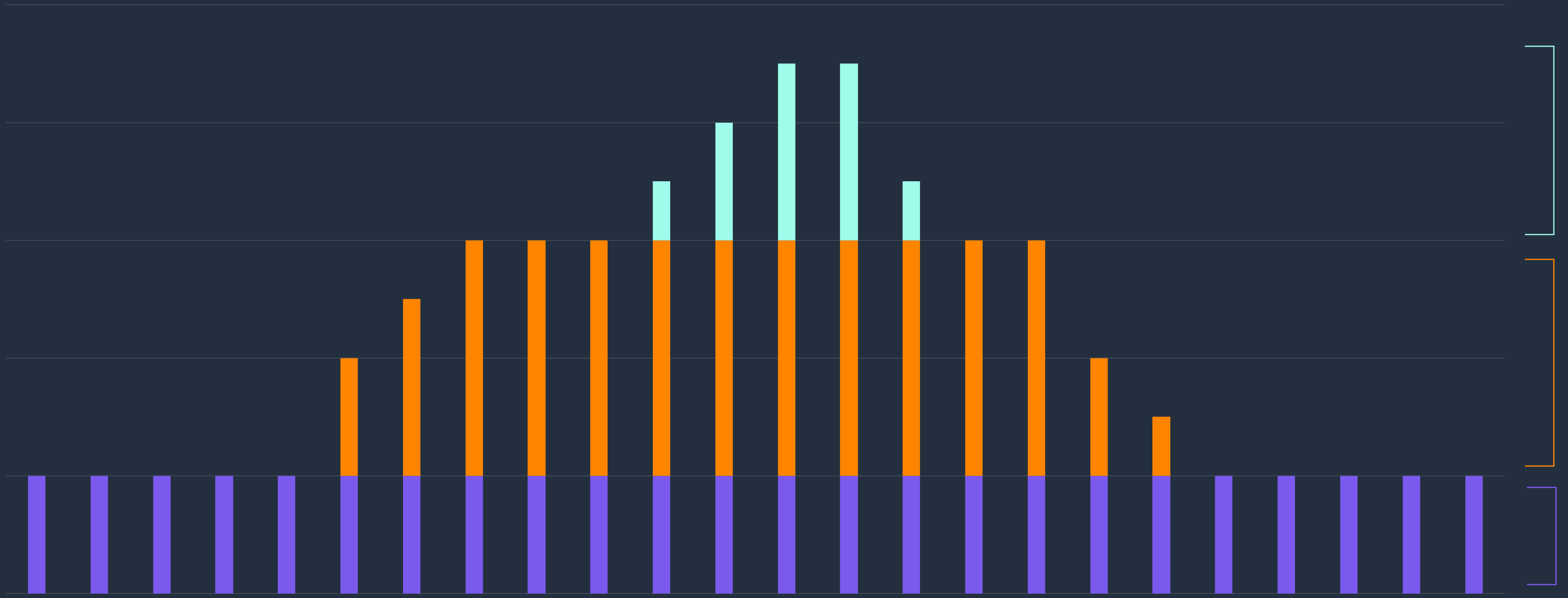
Committed flexible access to compute

Spot Instances

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

Fault-tolerant, flexible, stateless workloads

Simplifying capacity and cost optimization



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

Amazon EC2 Spot Overview

Amazon EC2 Spot

Spare Amazon EC2 capacity with savings of up to 90% over On Demand

Faster results

Increase throughput up to 10x while staying in budget

Easy to use

Launch through AWS services or integrated third-parties

Spot is ideal for workloads such as

Big data

Simulations

AI/ML Training

HPC

Spot is ideal for:

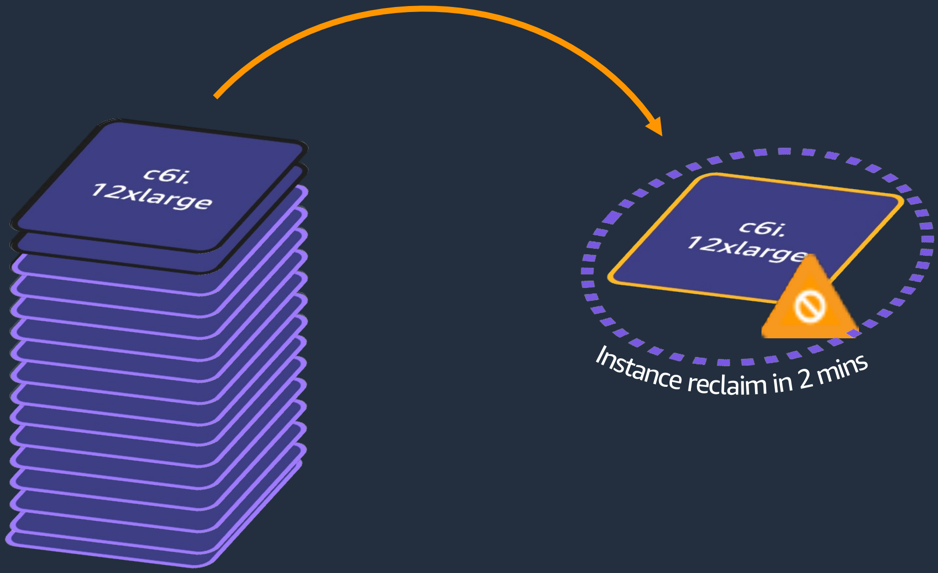
- Fault-tolerant
- Flexible
- Loosely coupled
- Stateless workloads

Or containerized workloads

EC2 Spot Interruptions

- By the nature of Spot as spare-capacity, instances can be reclaimed if needed by On-Demand
- AWS provides 2-minute notifications to enable you to handle the response in an automated way
- Diversification across instances reduces interruptions

- Historically, ***95%*** of the Spot instances launched in the last 3 months completed without interruption



A better way to leverage Spot?

- An up to 90% discount on EC2 is great, but you won’t see cost benefits if you have to re-run your job after Spot reclamations
- Not all software comes with memory checkpointing built-in

- 3rd Party AWS Partners, like MemVerge, provide software to solve this problem

MemVerge

Thank you!

Patrick Guha

patrguha@amazon.com
www.linkedin.com/in/patrickguha