
AWS
re:Invent
NOV. 27 – DEC. 1, 2023 | LAS VEGAS, NV

CMP207-R

Capacity, availability, cost efficiency: Pick three

Lorenzo Winfrey
(he/him)
Sr. Business Development Specialist,
Flexible Compute
AWS

Eric Legault
(he/him)
Principal Engineer
Salesforce

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

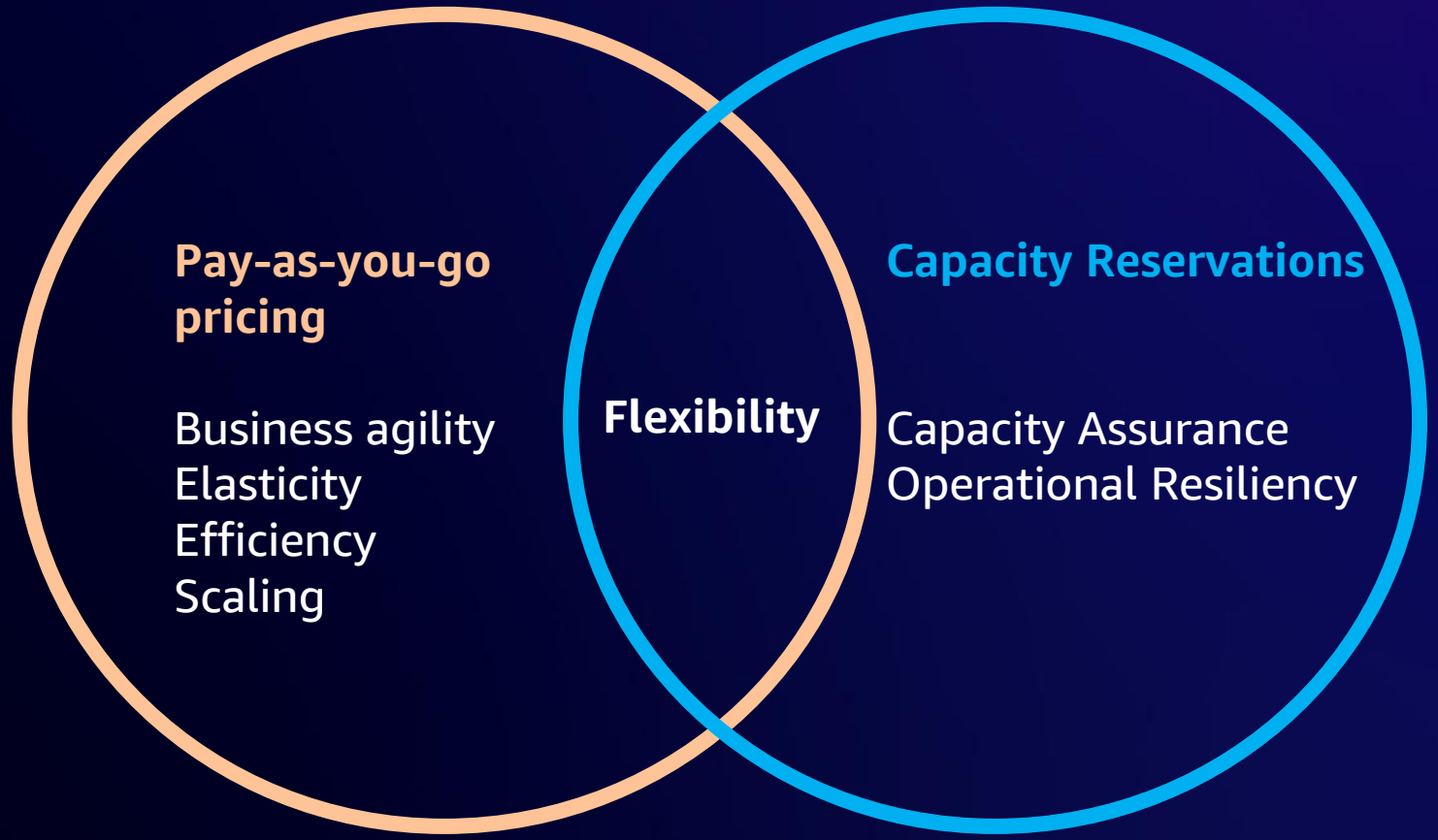
Capacity management success – optimize for availability



The key to successful capacity management is finding the right balance between performance, availability, and cost for your workloads

aws
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Every workload has unique requirements and objectives



aws
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

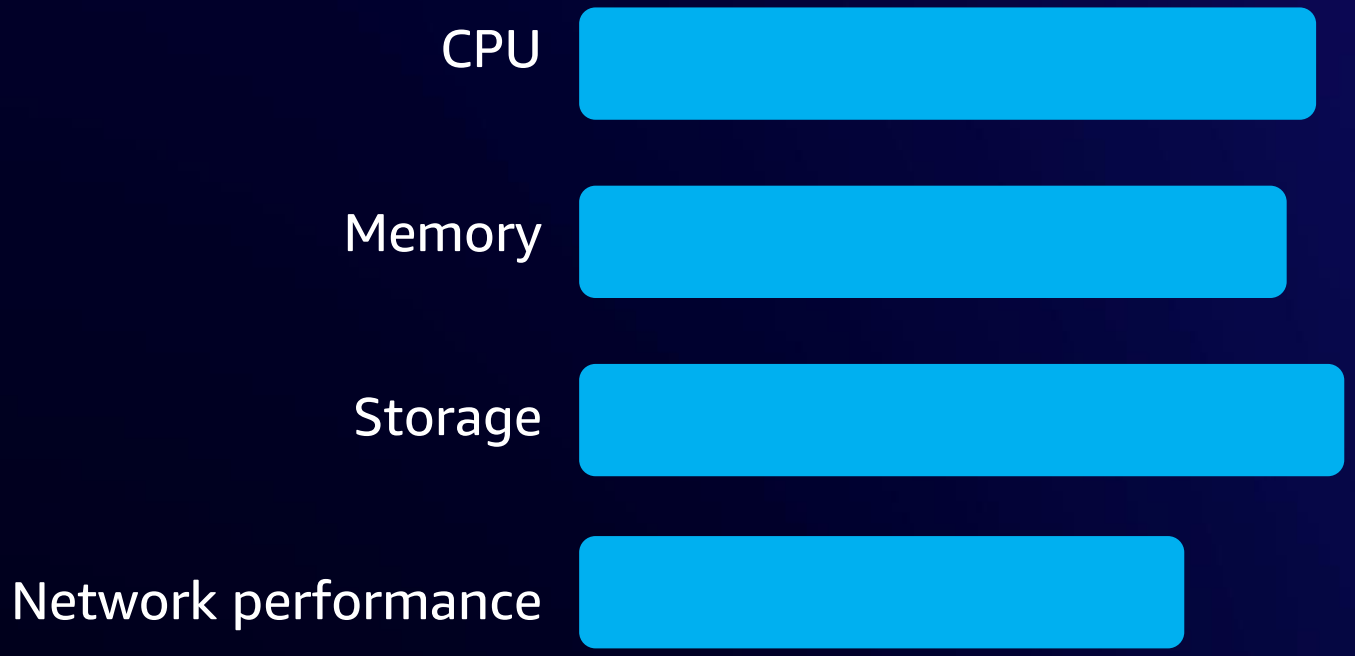
AWS Regions and Availability Zones (AZs)

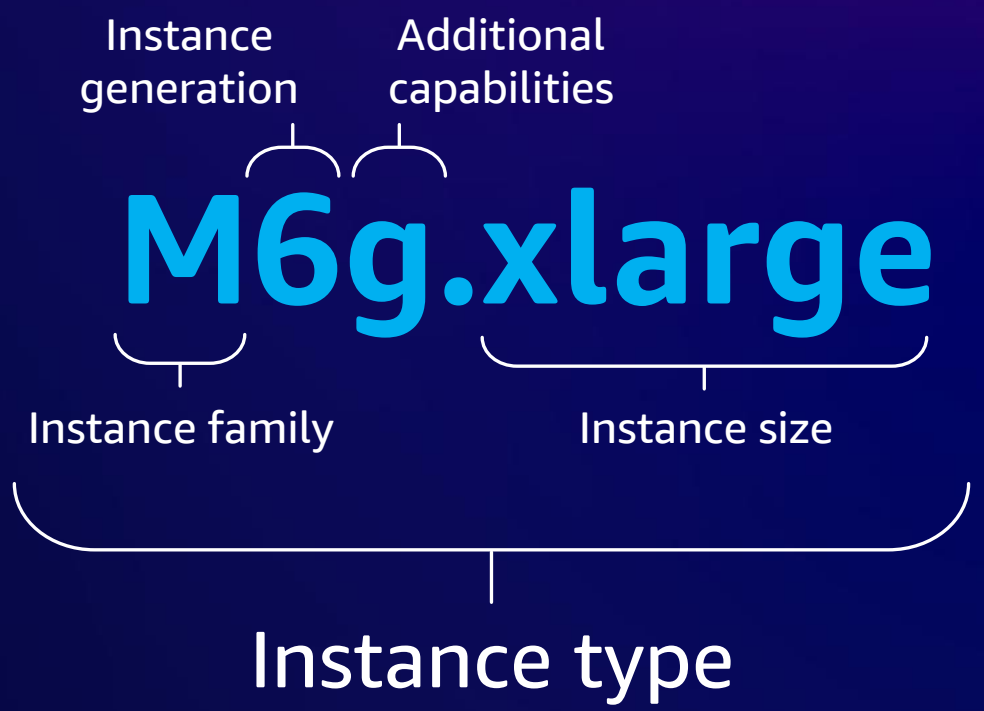


Announced Regions: 5 Regions and 15 AZs in Australia, Canada, Israel, New Zealand, and Thailand

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 instance characteristics





aws
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

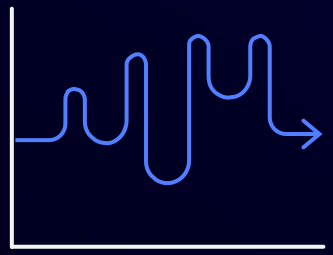
It’s all about flexibility

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS has multiple pricing models to optimize your costs

On-Demand

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



Stateful spiky workloads

Savings Plans

Significant savings for 1-year or 3-year hourly usage commitments



Committed and steady-state usage

Spot

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



Fault-tolerant, flexible, stateless workloads

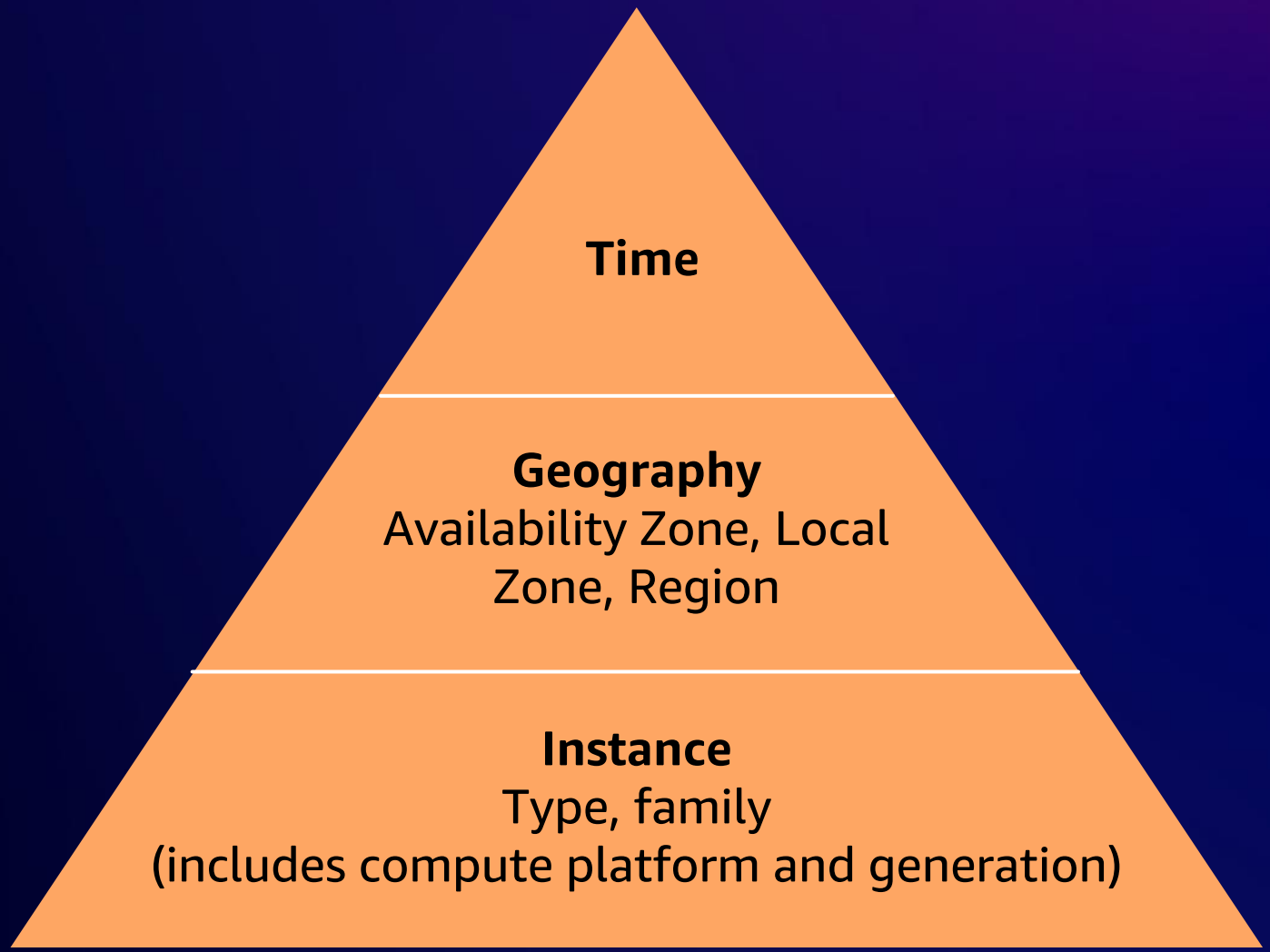
The best practice is to combine all purchase options

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Building flexibility into your workload

Flexibility is one of the most-effective ways to avoid insufficient capacity exception (ICE) issues



aws
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Flexibility in the cloud – Benefits



Better scalability via elasticity-related services and features



Higher availability by using more capacity pools, AZs, and regions



Lower cost through the right purchase options for the workload

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Select the instance types that fit your needs

650+

Instance types

for virtually every workload and business need

1. What processors can my workload use?
2. What are my workload's performance requirements?
3. What is my workload's consumption pattern?

A flexible list of instance types that fit **your workload**

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Flexibility toolset

Make use of the EC2 tools to get started and get the highest availability

Auto Scaling groups

A collection of EC2 instances that are treated as a logical grouping for the purposes of automatic scaling and management, across multiple instance types, CPU architectures, and purchase options

EC2 Fleet

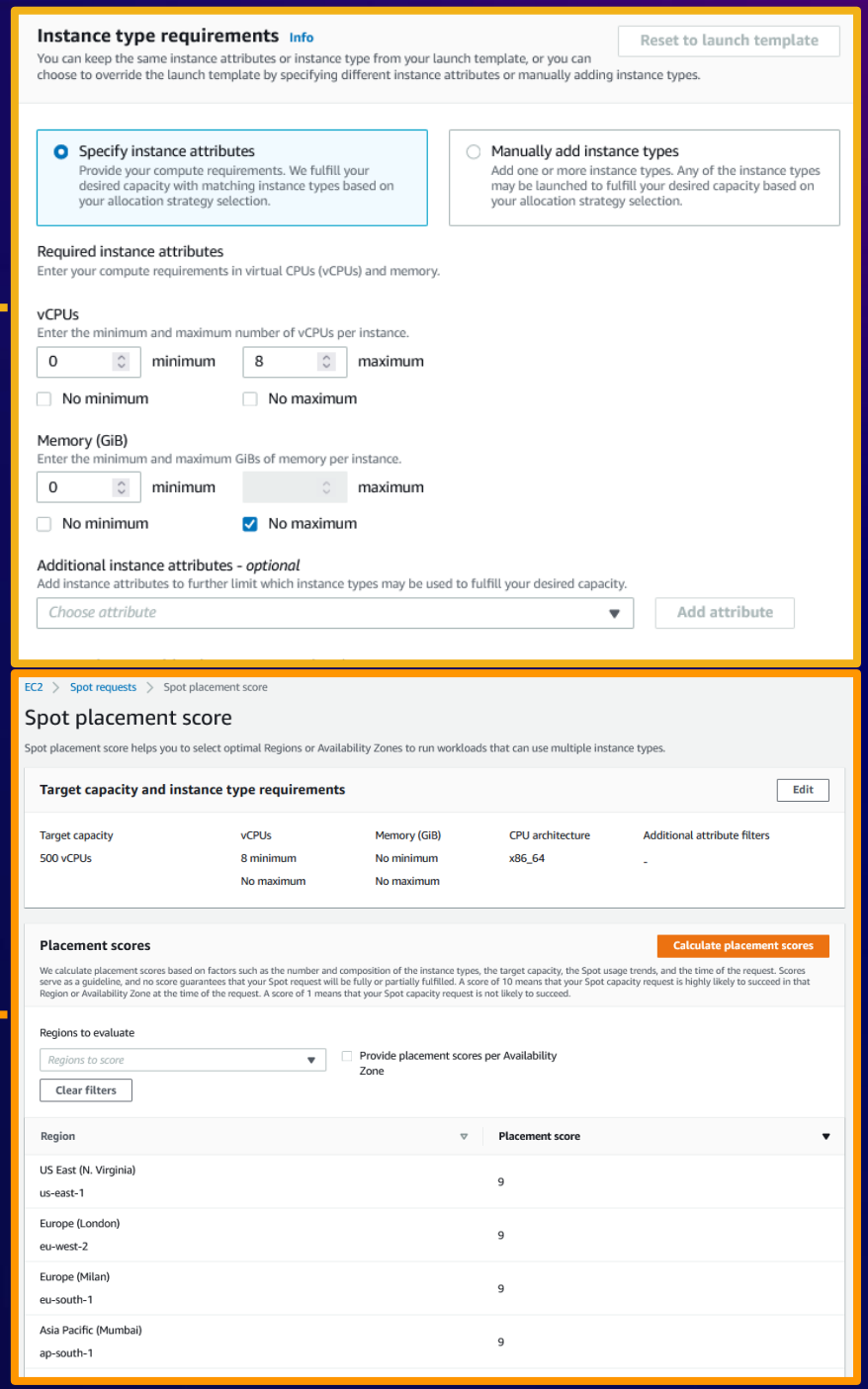
The configuration to launch a group of instances across multiple Availability Zones, purchase options, and instance types

Attribute-based instance selection

Rather than selecting specific instance types, this tool allows you to provision capacity on instance attributes (such as CPU/Memory) depending on requirements, also supporting new instances as soon as they become available

Spot placement score (SPS)

Find the optimal location (Reg/AZ) for your Spot workload; you provide the resource and capacity requirements, and SPS recommends locations (Reg/AZ) and the likelihood of getting the Spot capacity you need



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

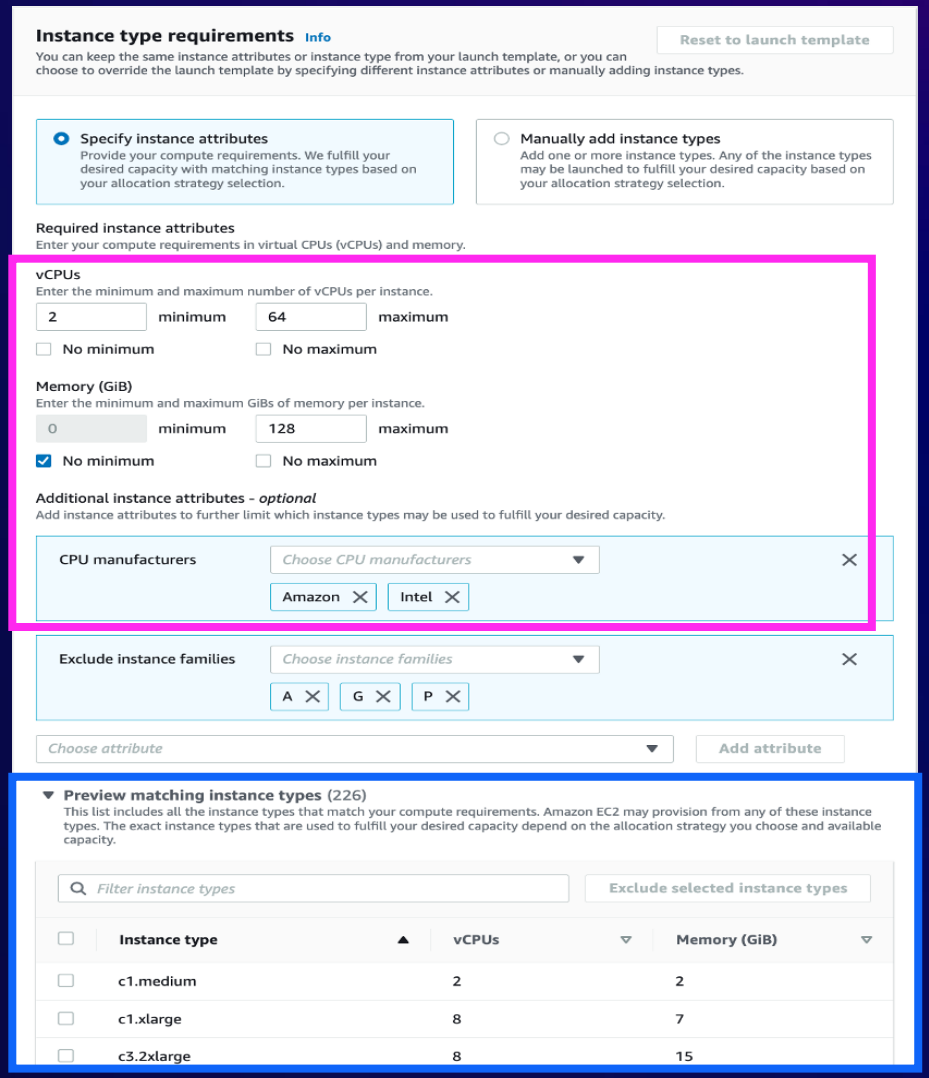
Diversify with attribute-based instance selection

PROVISION CAPACITY BASED ON INSTANCE ATTRIBUTES DEPENDING ON COMPUTE REQUIREMENTS, SUCH AS VCPU, RAM, STORAGE TYPE, AND MORE

Benefits

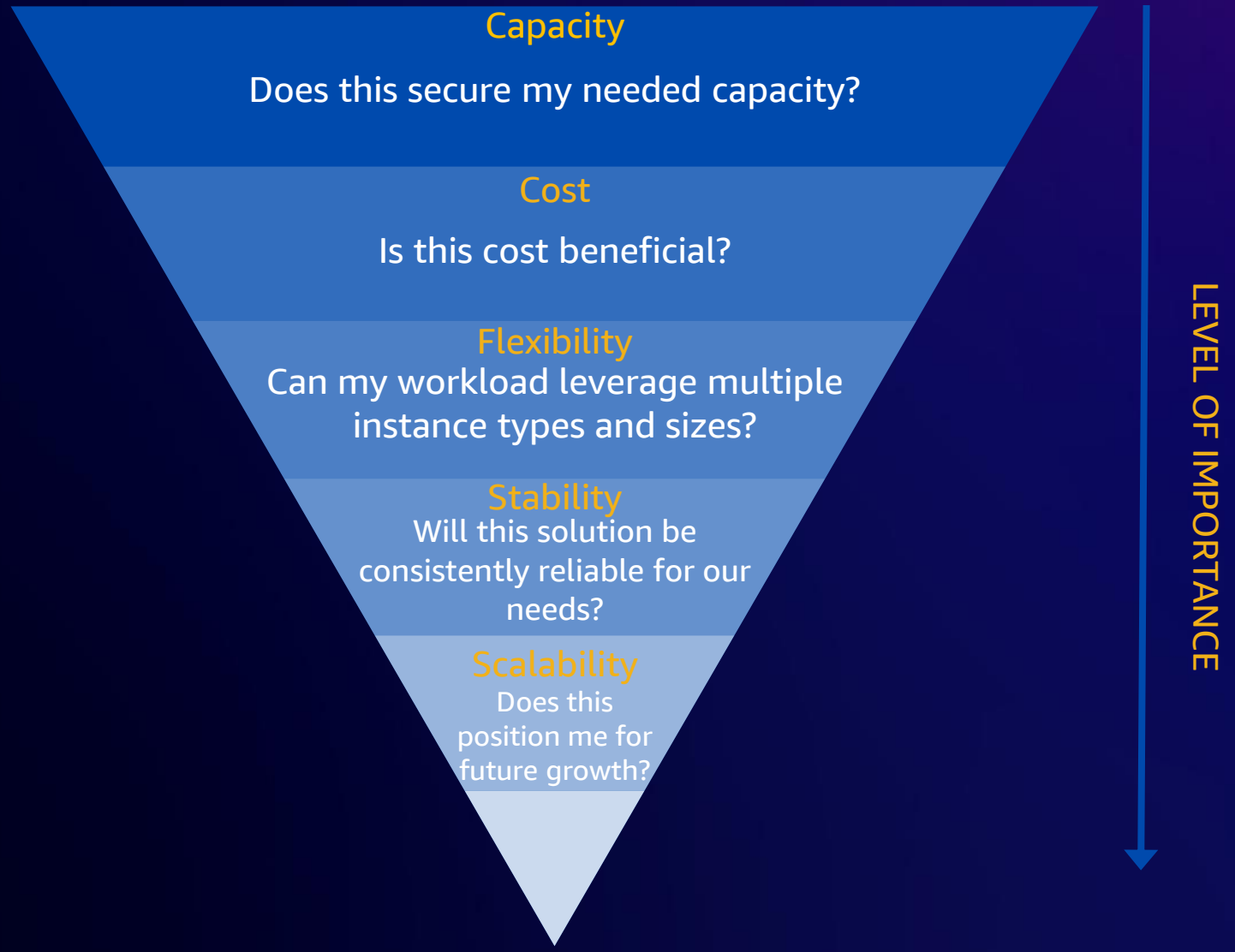
- Price protected
- Instance flexibility
- Simplifies provisioning
- Mixed instance groups
- Future proof
- Preview instances

- Attribute-based instance selection requires the use of launch templates
- On-Demand allocation strategy: “lowest-price”
- Spot allocation strategy: “price-capacity-optimized”



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What matters most to you?



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Savings Plans

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What are Savings Plans?



What are they?

A **flexible** pricing model that helps you save **up to 72%** for committed compute usage



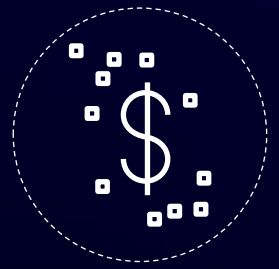
How do they work?

Commit to an hourly rate and the savings will **automatically** and simultaneously apply to your compute usage

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Types of Savings Plans



Compute Savings Plans

- Offers the **greatest flexibility**
- Discounts of up to **66%**
- Automatically applied to any usage across:

- **Region**
- **Instance family**
- **Instance sizes**
- **Tenancy**
- **Operating system**
- **Compute service options**



EC2 Instance Savings Plans

- Provides the **deepest savings**
- Discounts of up to **72%**
- Automatically applied to selected EC2 Instances and Regions across:

- **Instance sizes**
- **Operating system**
- **Tenancy**



SageMaker Savings Plans

- Up to 64% off eligible SageMaker machine learning instance usage
- Flexible across:

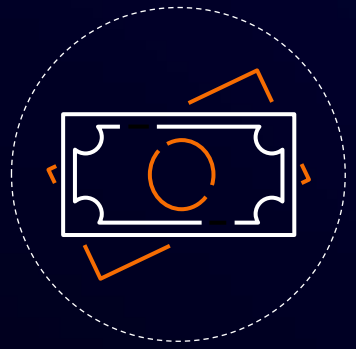
- **SageMaker ML usages**
- **Instance family**
- **Size**
- **Region**

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Three key takeaways

1. Save money on your steady-state usage



Reduce your compute bill for overall steady-state usage, regardless of instance family, Region, operating system, and tenancy

2. Reduce costs by modernizing workloads



Optimize costs when migrating workloads to newer instances or modernizing your application with AWS Lambda and AWS Fargate

3. Centralize and simplify cost management



Purchase Savings Plans in one AWS account, and AWS will automatically apply it across your AWS organization

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Capacity management and planning

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Function of capacity management

HIGHEST PERFORMANCE AT THE LOWEST COST



Performance



Cost

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Application use cases

Use case	Compute profile	Compute option
Modern applications, stateless batch/HPC (depends on framework)	Stateless, fault-tolerant, highly instance flexible and cost-driven	Apply flexibility and combine Spot instances, Saving Plans, and On-Demand
Stateful batch/HPC, machine learning (depends on framework)	High compute demand, spikey and unpredictable, might run for a short period of time	Work to identify the right balance between Capacity Reservations (for steady state/baseline) and flexibility across locations/instances
Databases, load testing, accelerated computing	Performance sensitive, require specialized hardware, non-flexible, predictable runtime patterns	Capacity Reservations

aws

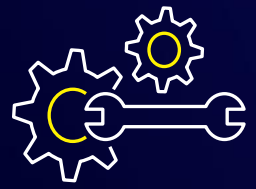
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Situations needing higher levels of availability

For business critical workloads



Disaster recovery



Maintenance



Regulatory requirements



Peak events

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

On-Demand Capacity Reservations

For steady-state workloads



- Manage capacity and discounts independently
- No commitment required; can be created and canceled as needed, if the capacity is available
- Reserve capacity by instance type, tenancy, Availability Zone, placement group
- Capacity held and charged whether or not you run instances
- Active until canceled by the customer
- Share reservations across accounts

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The capacity management lifecycle

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Capacity management lifecycle

Changing business needs

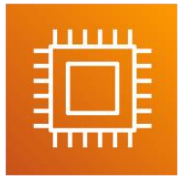
Plan



Gather information

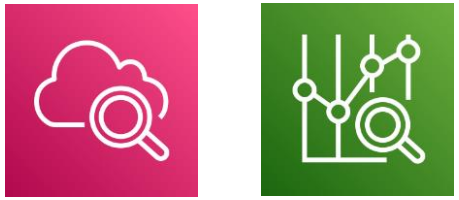
Pre-optimize

Acquire



Launch or procure instances

Monitor



Discover the insights you need to take action

Optimize



Optimize for cost, availability, and performance

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Capacity planning at Salesforce UIP

How we save millions by leveraging multiple EC2 provisioning options

Eric Legault, Principal Engineer
elegault@salesforce.com
(He/Him)

What is Salesforce UIP project ?

Unified Intelligence Platform (UIP) - Internal Data Lake

UIP

- 100PB+ Data Lake
- 200-250 TB processed per day
- 1 Trillion+ events
- Compute engines: Apache Spark / Trino
- 15+ clusters / 3000+ nodes per month
- $1 Million saved per month using our current EC2 provisioning strategy

Balancing Capacity / Availability / Cost

- Analytic workloads - Fault tolerant by using Spark
- Service Level Agreements (SLAs) - Jobs
- Non SLAs - Jobs
- Ingestion processing in 5 minutes windows
- Willing to trade performance for cost in some cases



Provisioning

- Capacity reserved for minimum requirements
- Saving plan – On Demand / Flexible requirement
- Spot for everything else



Regions

Two Regions





Availability Zones

Three Zones (A-B-C)



EC2 Fleets characteristics



Memory Fleets	Balanced Fleets	Compute Fleets	Mixed Fleets
R Families	M Families	C Families	R / M Families
r6i	m6i	c6i	r6i
r6id	m6id	c6id	r6id
r6in	m6in	c6in	r6in
r6idn	m6idn	c6idn	r6idn
			m6i
			m6id
			m6in
			m6idn

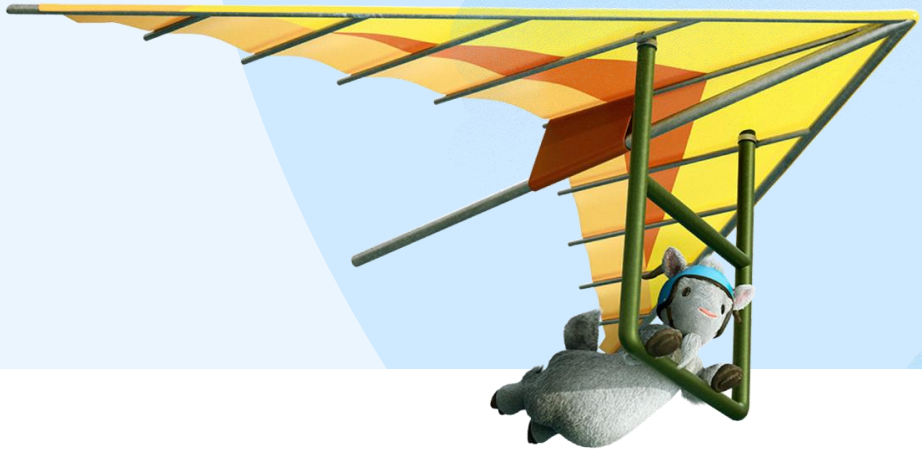


Also to consider :

- Local SSD / EBS
- Storage volume
- Network
- EBS Bandwidth
- AMD
- Intel

“Using AWS helped us to play with many different configurations of machines and try different scenarios. It’s just a matter of changing the configuration and a couple of hours later we could see if it actually worked or not”

“In terms of building our fleet, I think the real benefit here is how many instance types you can use”

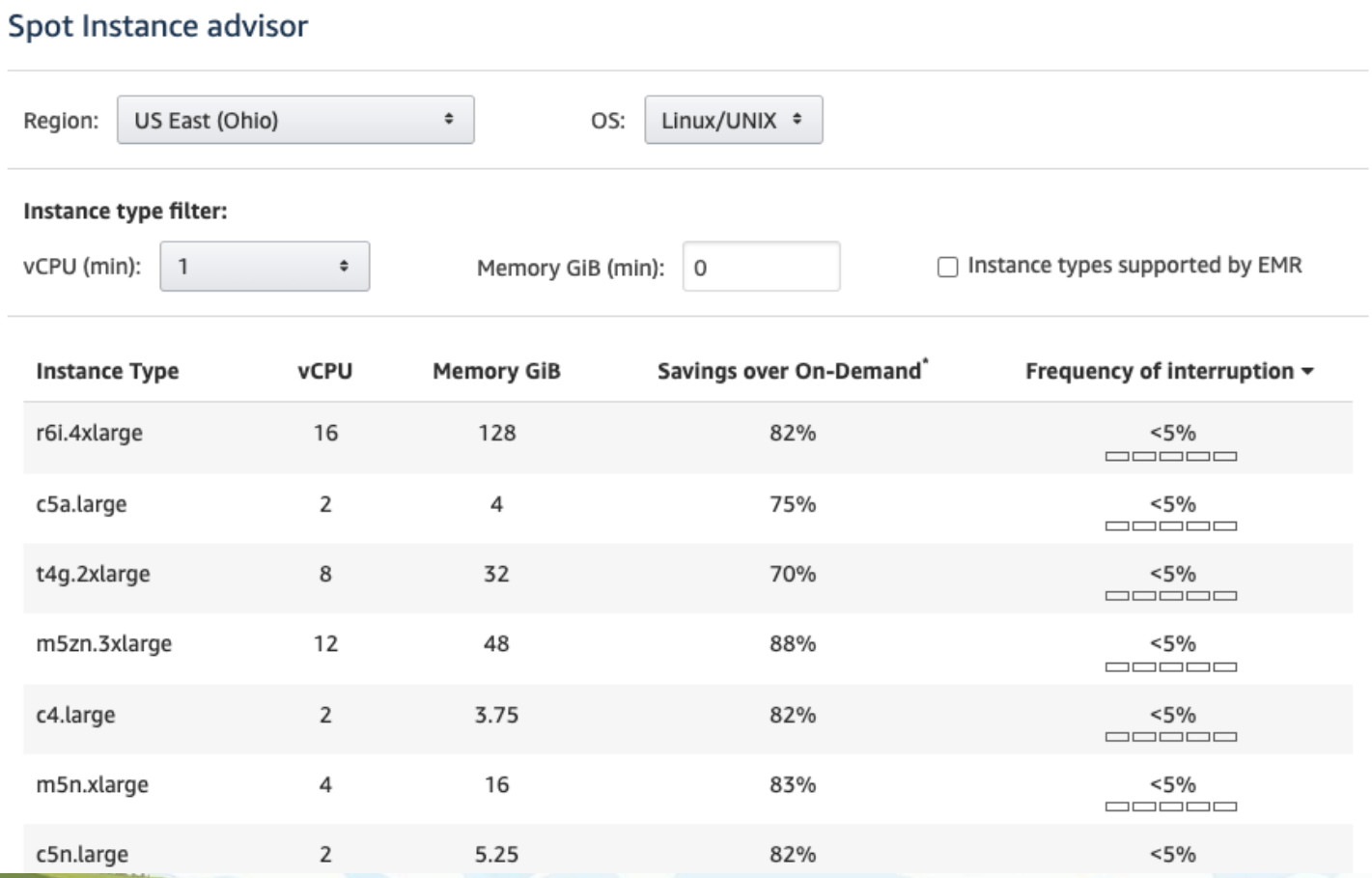


Elasticity

- Autoscaling clusters + 5% for Reclaims
- Leverage Compute Saving plan in follow the Sun mode
- Use more Spot when needed
- 60+ instance types
- Two+ EC2 generations

Identify Spot instance type that are more at risk of reclaim

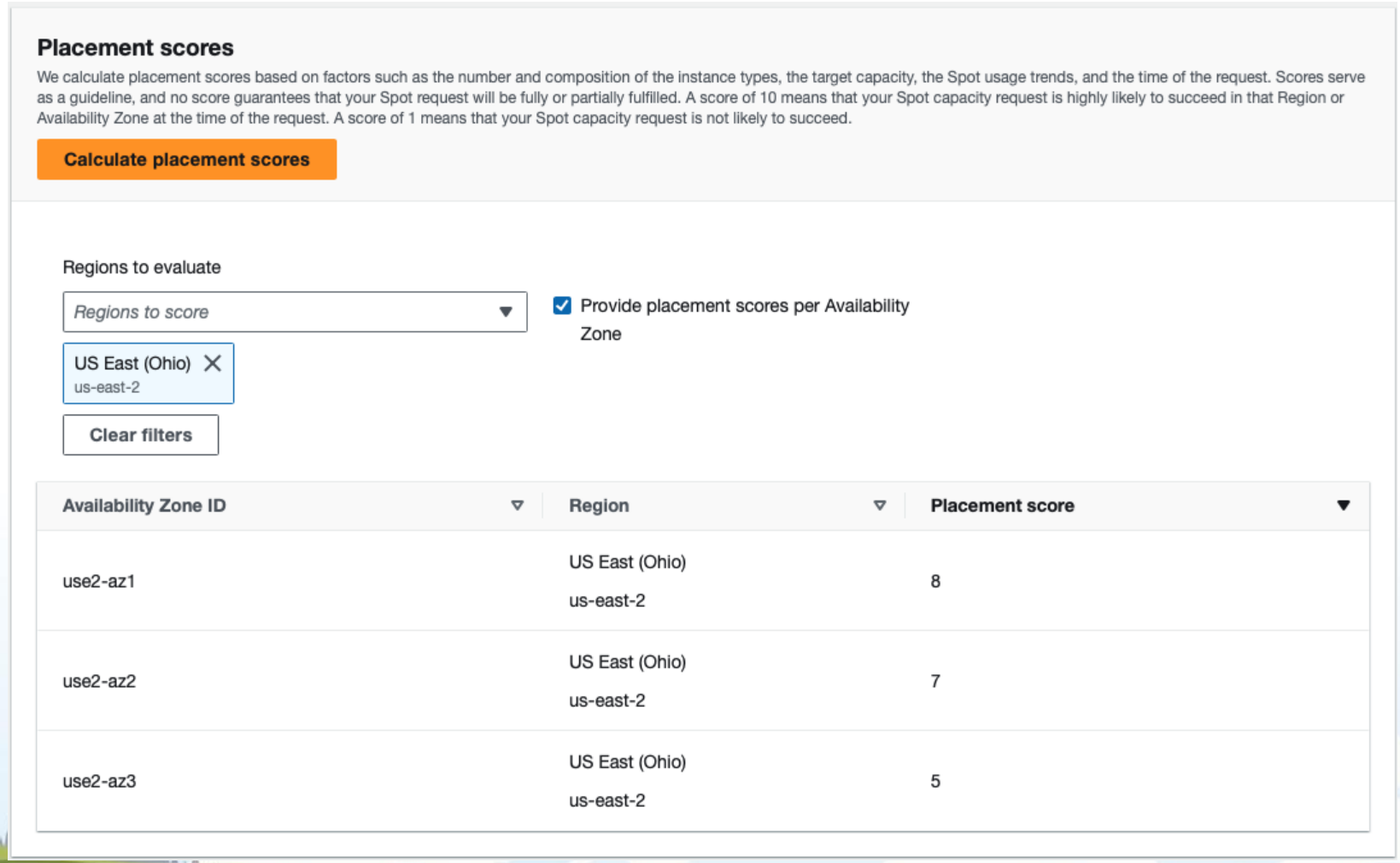
Spot instance advisor allows you to know frequency of interruptions (Reclaim) as well as average savings over On-Demand per region:
https://aws.amazon.com/ec2/spot/instance-advisor/



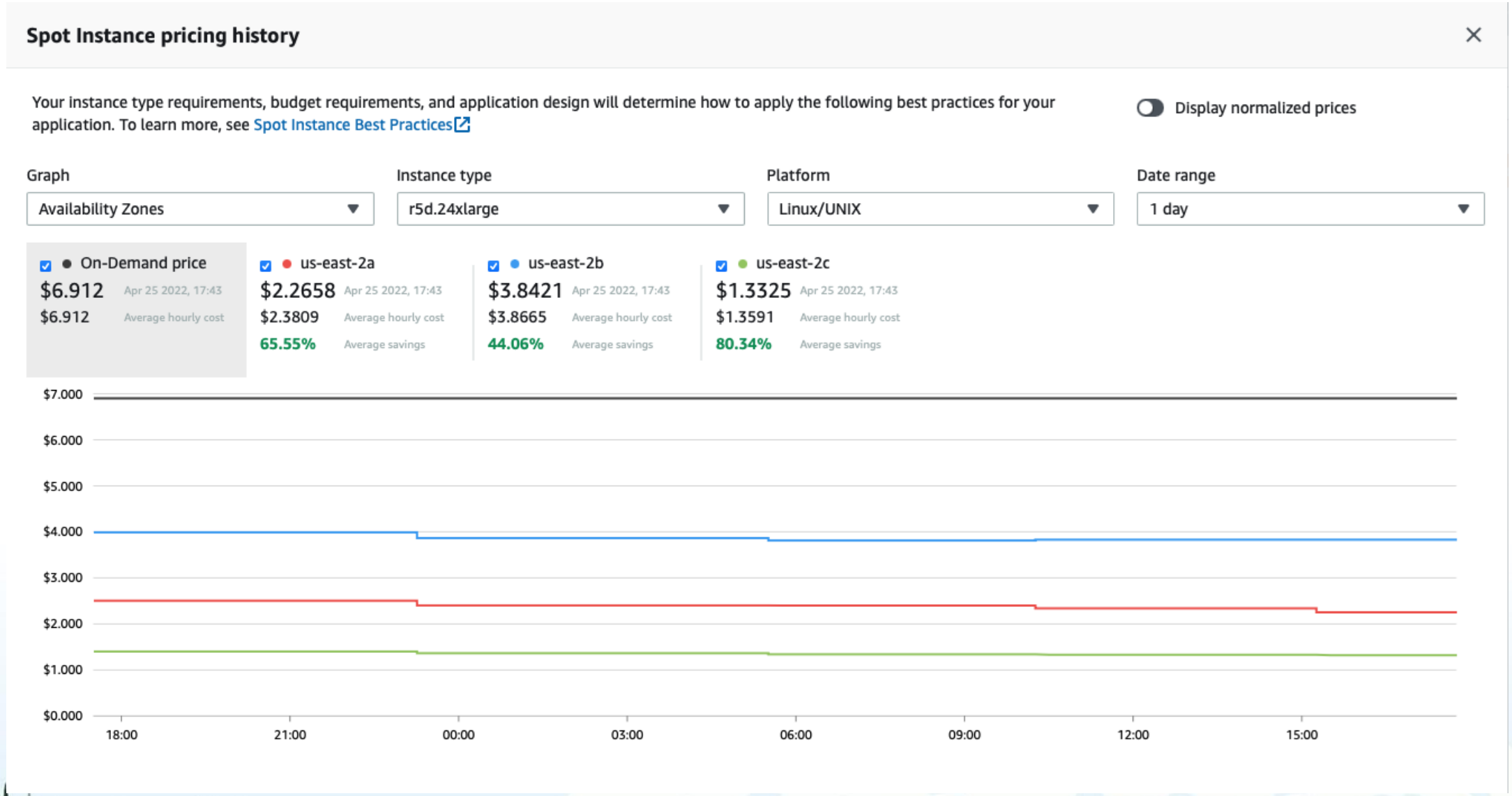
Instance Type	vCPU	Memory GiB	Savings over On-Demand*	Frequency of interruption
r6i.4xlarge	16	128	82%	<5%
c5a.large	2	4	75%	<5%
t4g.2xlarge	8	32	70%	<5%
m5zn.3xlarge	12	48	88%	<5%
c4.large	2	3.75	82%	<5%
m5n.xlarge	4	16	83%	<5%
c5n.large	2	5.25	82%	<5%

Identify optimal location where to launch your clusters

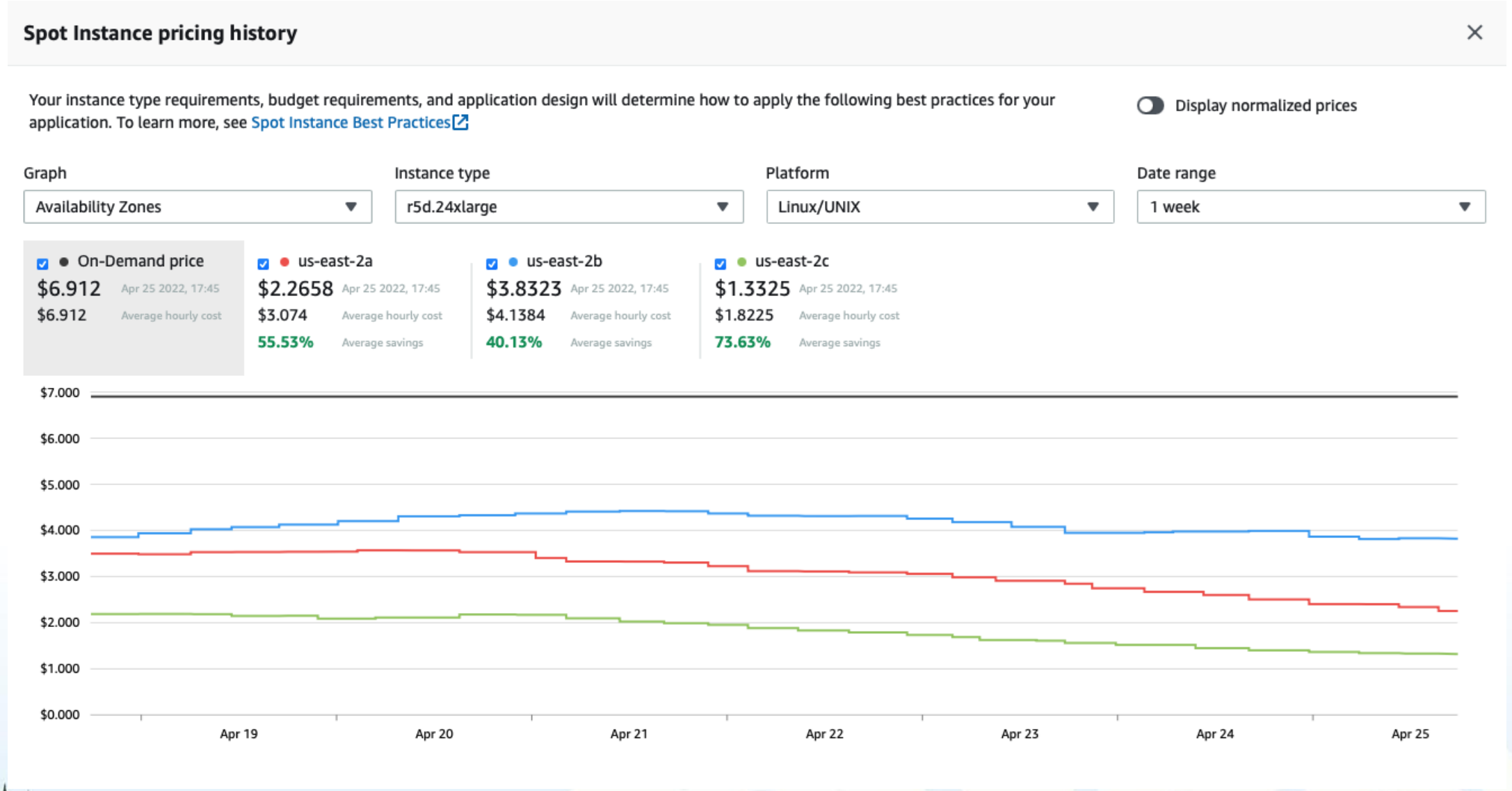
Amazon EC2 Spot placement score allows you to find the optimal location for your Spot workloads.



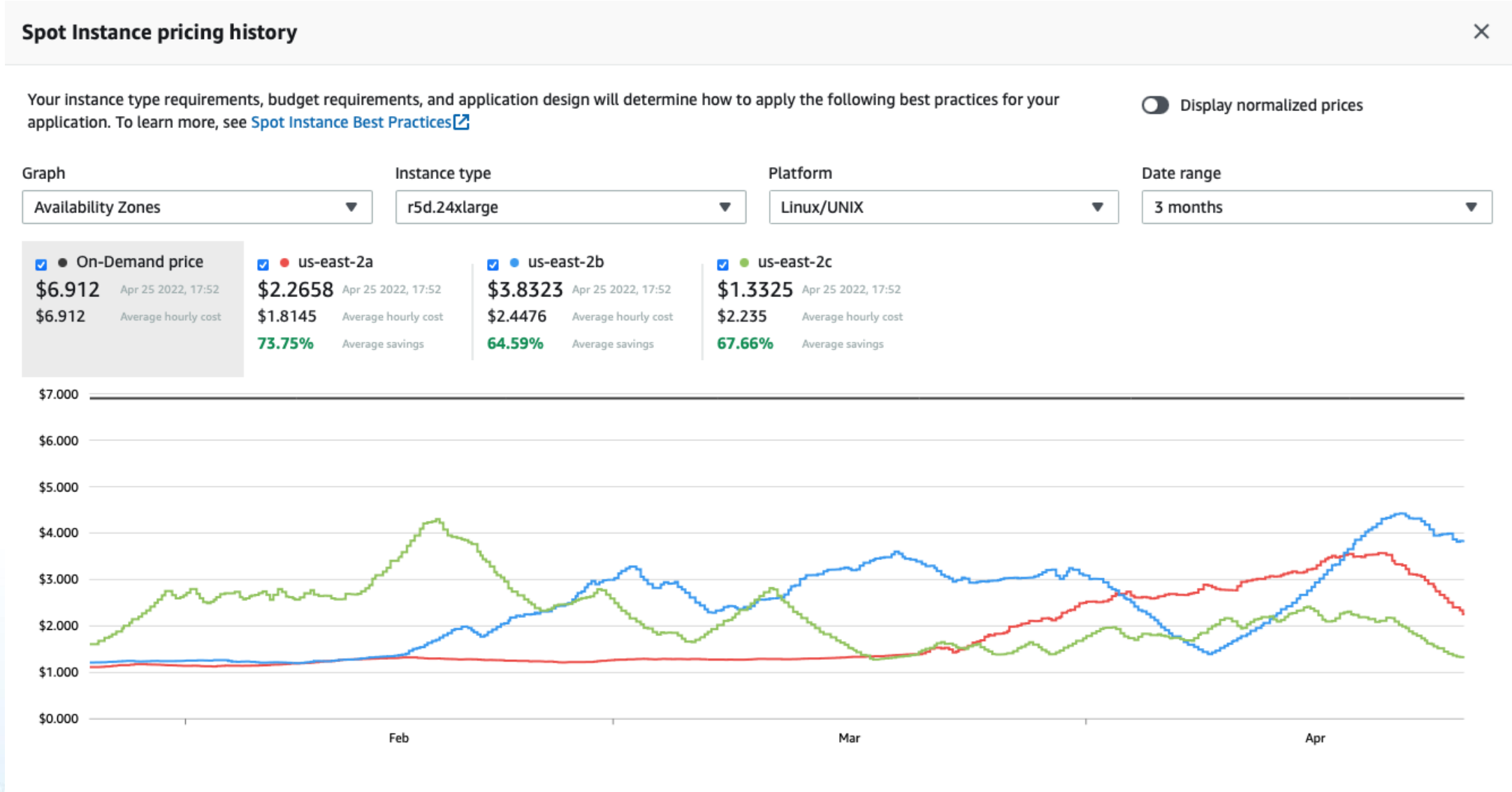
Spot Pricing History - 1 Day



Spot Pricing History - 1 Week



Spot Pricing History - 3 Month



Salesforce UIP Uses Amazon EC2 Spot Instances to Reduce Data Processing Times by More than 90% and Save Over $1 Million Monthly

$1 million
saved per month

200–250 TB
processed for the cost of 100 TB previously

4 hours to 15 minutes
processing time reduction

Configuration tests
explored and tested within hours

Overview

The Unified Intelligence Platform Team (UIP) of Salesforce manages a petabyte-level data lake, and it was looking to innovate the analysis and processing of data, with an eye toward cost savings and greater efficiency. Using Amazon Web Services (AWS) for a mix of instance-provisioning models from Amazon Elastic Compute Cloud (Amazon EC2), which provides secure and resizable compute capacity for virtually any workload, the UIP Team was able to build a scalable, elastic compute infrastructure. Its remodeled compute infrastructure takes less time to process twice as much data while saving the company over $1 million monthly.

Salesforce uses AWS for its data workflows, while AWS uses Salesforce for its customer relationship management. This partnership makes it simpler for developers using both technologies to build and launch customer applications, use AWS services natively within Salesforce, and securely connect data and workflows across both Salesforce and AWS.



https://aws.amazon.com/solutions/case-studies/salesforce-case-study/


THANK YOU
D♥TA
AI+DATA+CRM

And before you go

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Introduction to Amazon EC2 Capacity Blocks

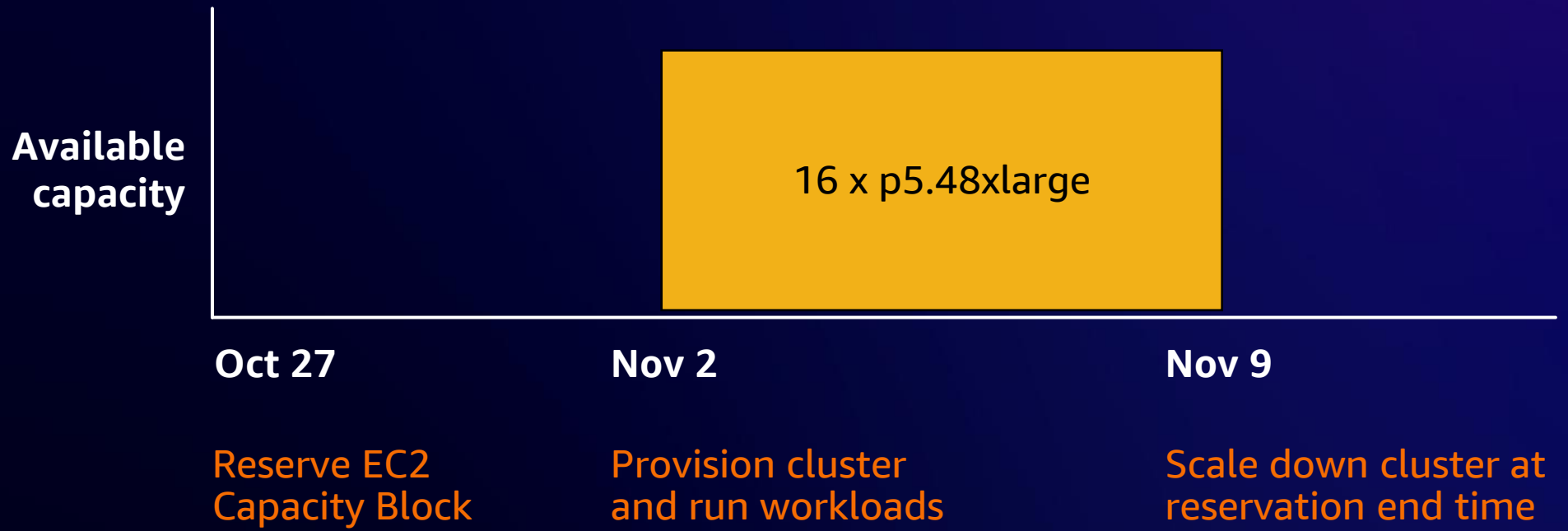
Better plan your ML projects around capacity availability

Pay for only the compute time that you need—no long-term commitment is required

Get low-latency, high-throughput network connectivity through colocation in Amazon EC2 UltraClusters

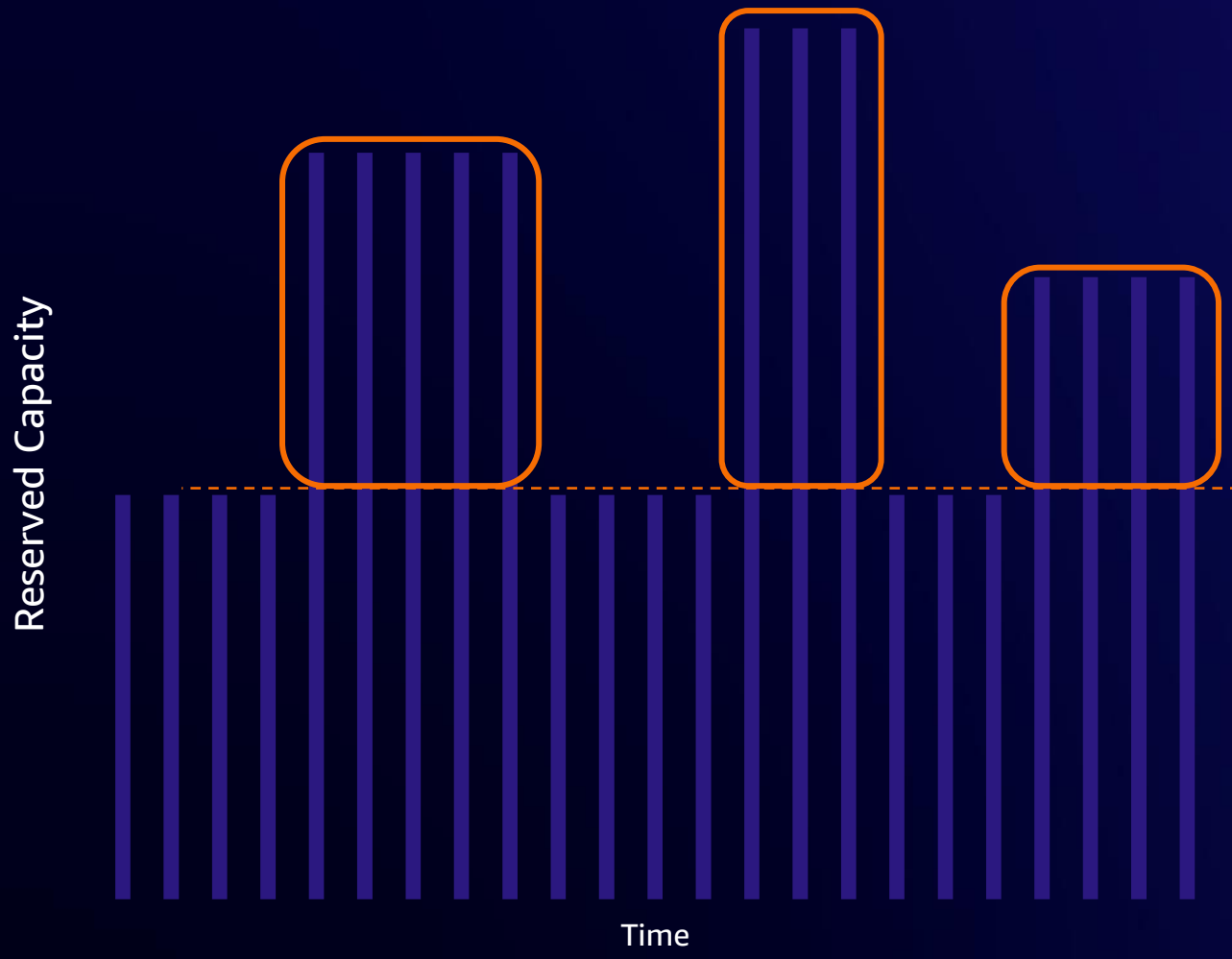
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

EC2 Capacity Block lifecycle example



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Supplement On-Demand Capacity Reservations with EC2 Capacity Blocks



Capacity assurance for short-duration needs that exceed steady-state reserved capacity

EC2 Capacity Blocks

Capacity assurance for steady-state usage over long term

ODCR + Savings Plans

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key use cases



Train or fine-tune ML models per your schedule

Reserve blocks of GPU capacity for training or fine-tuning ML models



Build **generative AI workloads** efficiently

Secure GPU compute capacity up to eight weeks in advance by reserving the right amount of capacity for each phase of your deployment



Conduct **research experiments** uninterrupted

Get uninterrupted access to the capacity that you reserved to complete experiments and prototyping

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key takeaways

- The key to finding the right balance between capacity, availability, and cost is **flexibility**.
- When it comes to selecting the right purchasing option, remember its not an either/or decision, it’s an **and** decision. Combine them to maximize flexibility and cost efficiency.
- Use the **capacity management lifecycle (PAMO)** in all phases of your cloud journey to make sure you are considering all relevant factors.
- Finally, if you need more flexibility and agility when it comes to GPU usage, don’t forget to check out **EC2 Capacity Blocks**.

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Continue your compute learning

Learn at your own pace



Expand your knowledge with learning plans via **AWS Skill Builder**

Increase your knowledge



Use our **AWS Ramp-Up Guides** to build your compute knowledge

Earn your Compute badge



Demonstrate your knowledge by achieving **digital badges**

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!



Please complete the session survey in the mobile app

Lorenzo Winfrey

winfreyl@amazon.com

linkedin.com/in/lorenzowinfrey

Eric Legault

elegault@salesforce.com

linkedin.com/in/ericlegault1

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.