Reduce Inferencing Cost by up to 90% using Amazon Elastic Inference and Amazon EC2 Spot Instances

Chakravarthy Nagarajan
Solutions Architect – EC2 Spot

Agenda

- Challenges in scaling deep learning Inference at scale
- Amazon Elastic Inference – recap
- Amazon Elastic Inference with EC2 Spot - Architecture
- Amazon EC2 Spot Instances – Best practices
- Tools for Operational efficiency at scale
- Demo

aws

Challenges in Deep Learning Inference at scale

- Right sizing with flexibility
- Under utilization of compute resources(GPU/CPU/Mem)
- Elasticity
- Optimize compute cost
- Operational efficiency

Amazon Elastic Inference - recap

Amazon Elastic Inference

Reduce deep learning inference costs up to 75%

Lower inference costs

Match capacity to demand

Available between 1 to 32 TFLOPS per accelerator

KEY FEATURES

Integrated with Amazon EC2 and Amazon SageMaker and Amazon ECS

Support for TensorFlow, Apache MXNet, and ONNX

Single and mixed-precision operations

aws

Acceleration sizes tailored for inference

Accelerator Type	FP32 Throughput (TOPS)	FP16 Throughput (TOPS)	Accelerator Memory (GB)	Price($/hr) (US)
eia2.medium	1	8	2	$0.12
eia2.large	2	16	4	$0.24
eia2.xlarge	4	32	8	$0.34

Attach accelerators to any Amazon EC2 instance or Amazon SageMaker instance type or Amazon ECS Task

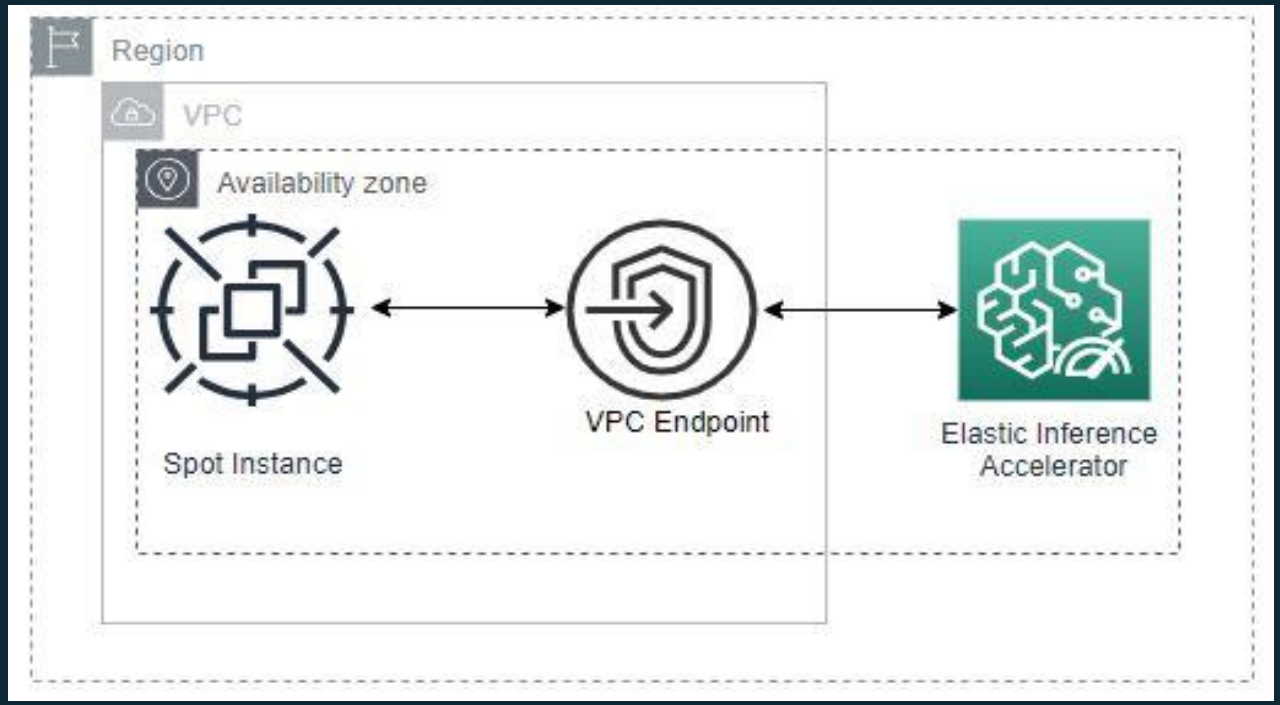
Available in N. Virginia, Ohio, Oregon, Dublin, Seoul and Tokyo (eia1* only)

*eia1 is our 1st EIA family with upto 4GB in memory

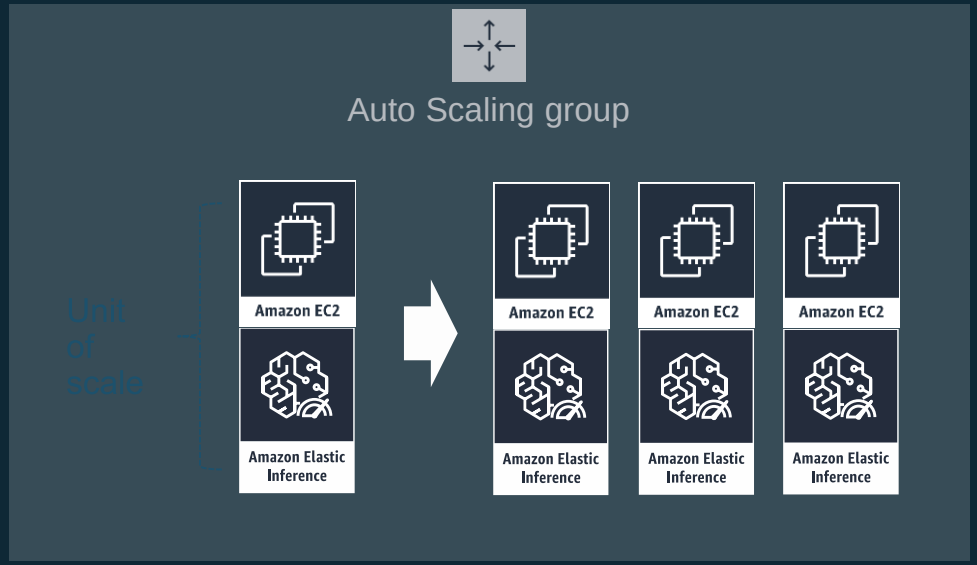
aws

Amazon Elastic Inference - Architecture

Amazon Elastic Inference with EC2 Spot Instances – Architecture



Scale capacity in EC2 Auto Scaling groups



Specify EI within launch templates

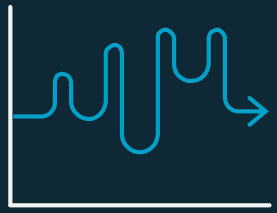
aws

Amazon EC2 Spot Instances

Amazon EC2 purchase options

On-Demand

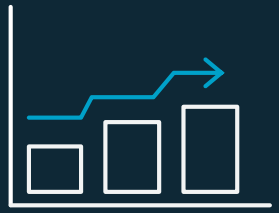
Pay for compute capacity by **the second** with no long-term commitments



Spiky workloads, to define needs

Reserved Instances

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



Committed & steady-state usage

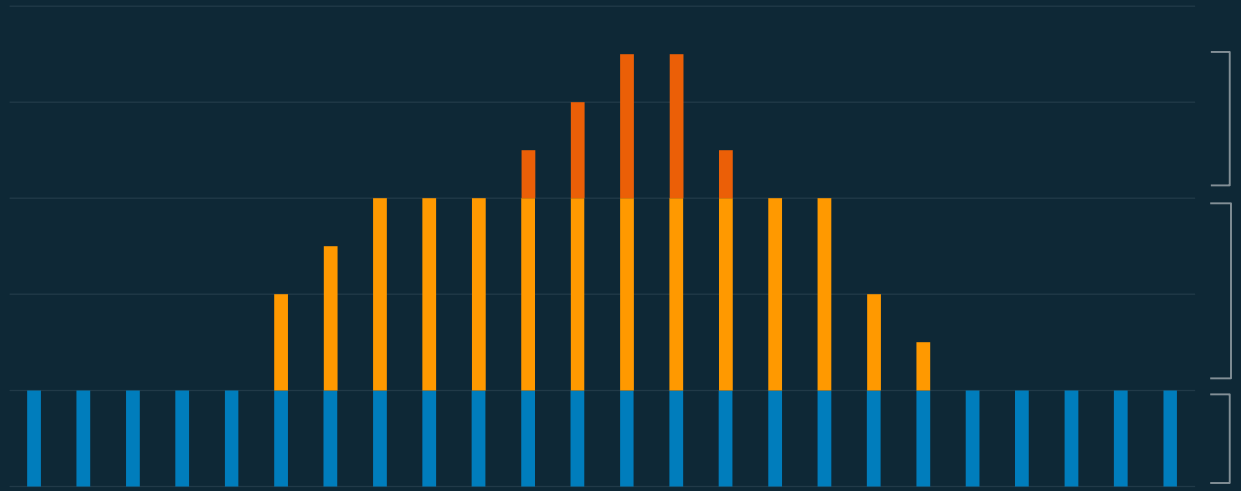
Spot Instances

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

Fault-tolerant, flexible, stateless workloads

aws

Combine purchase options to optimize costs



Scale using Spot for fault-tolerant, flexible, stateless workloads

On-Demand, for new or stateful spiky workloads

Use **RIs** **for known, steady-state workloads**

aws

How do Amazon EC2 Spot Instances work?

70 - 90% off

Spot infrastructure

Is same as On-Demand and RIs

Spot pricing

Smooth, infrequent changes no spikes, more predictable

Interruptions

Only happen when OD needs capacity (**no bidding**)

Diversify

Choose different instance types, size and AZ in a single fleet

aws

Amazon EC2 Spot instance pools explained

C4	1a	1b	1c	On Demand
8XL	$0.50	$0.27	$0.29	$1.76
4XL	$0.21	$0.30	$0.16	$0.88
2XL	$0.08	$0.07	$0.08	$0.44
XL	$0.04	$0.05	$0.04	$0.22
L	$0.01	$0.01	$0.04	$0.11

Each instance family

Each instance size

Each Availability Zone

In every region

Is a separate ***Spot pool***

Amazon EC2 Spot integrations

Auto Scaling

AWS Batch

Amazon Elastic Container Service

Amazon Elastic Container Service for Kubernetes

Amazon EMR

Amazon SageMaker

AWS CloudFormation

AWS Thinkbox Deadline

aws

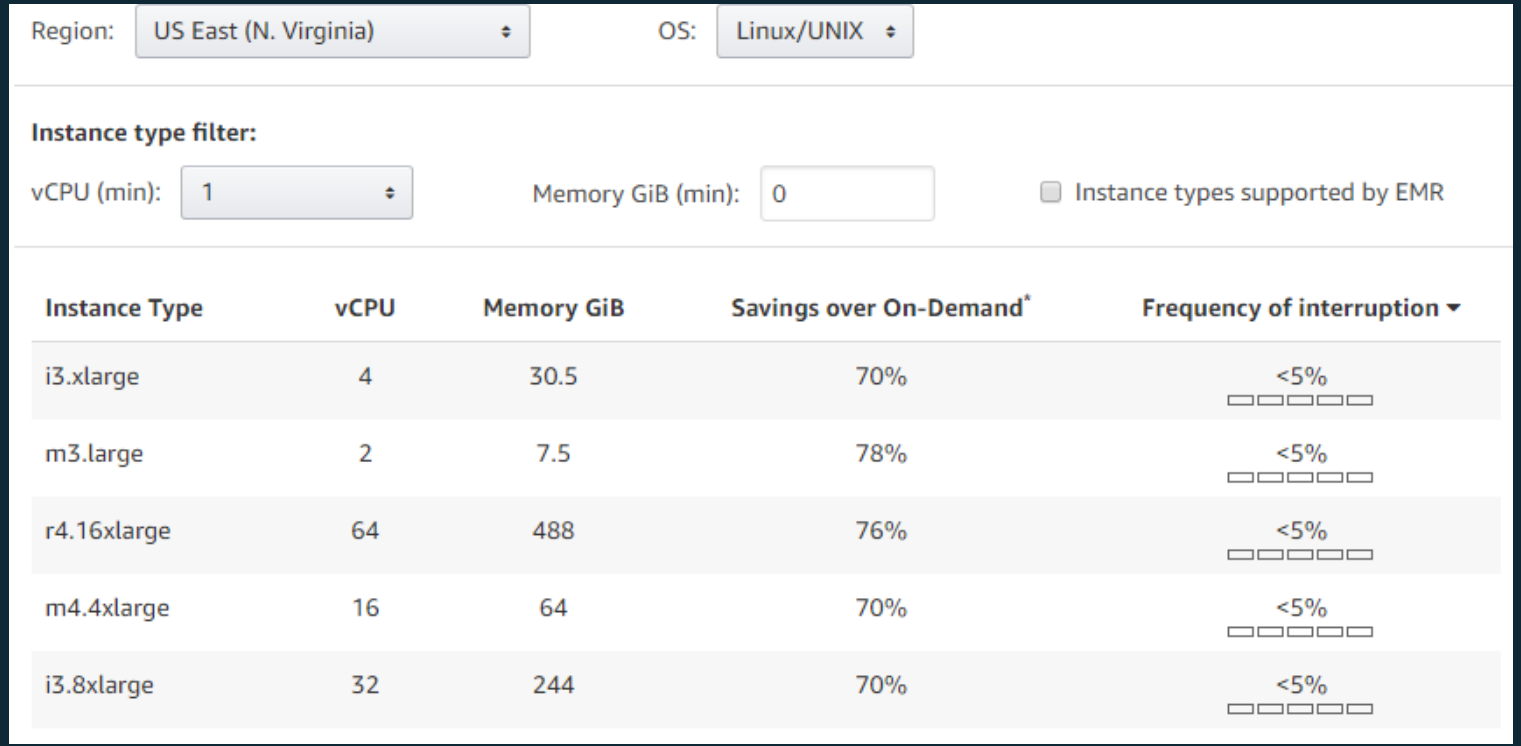
Flexibility is key to successful adoption

Instance flexible

Time flexible

AZ flexible

Select the best instance types with Spot Instance Advisor



https://aws.amazon.com/ec2/spot/instance-advisor/

What about interruptions?

Minimal interruptions

Over **95%** of the instances were not interrupted in the last 3 months

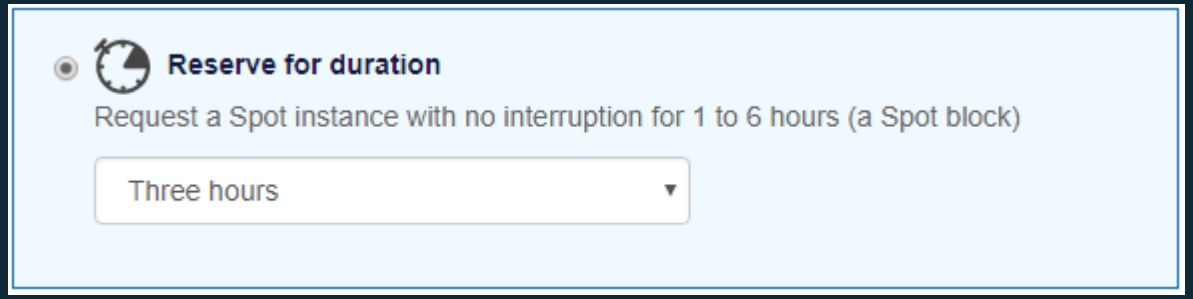
Leverage Spot Instance Advisor to select the suitable instance types to your workload with the lowest interruption rates

Catch the interruption notifications by polling EC2 metadata or using Cloudwatch events and automate response to interruptions:

- ELB connection draining and graceful shutdown
- Checkpointing
- Cordon your node and drain containers

EC2 Spot Blocks

- Defined duration workload without interruptions
- 1-6 hours
- Lower discounts compared to Spot

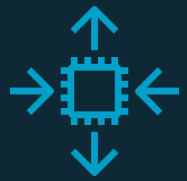


aws

Tools for Operational efficiency at scale

Provisioning EC2 Spot Capacity

Launch Templates



Auto Scaling Groups

EC2 Fleet

Lets see how this all works together to automatically optimize scale, performance, and cost behind the scenes

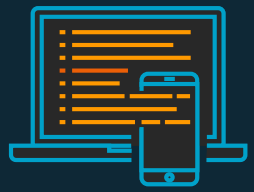
aws

Launch Templates

Increase productivity

Simplified permissions

Governance & best practices



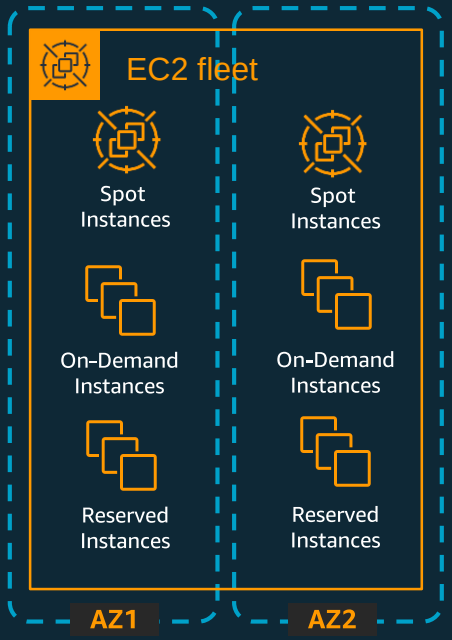
Consistent experience

Launch Templates are supported in EC2 Auto Scaling groups, EC2 Fleet and AWS Batch

aws

Amazon EC2 Fleet

*Allows to **synchronously** provision capacity across different instance types, AZs, and purchase options with a single API*



Use all three purchase options to optimize costs

Automatic optimization behind the scenes with machine learning

Benefits

Reduce costs

Increase operational efficiency

Reduce development effort

Key features

Flexible capacity allocation

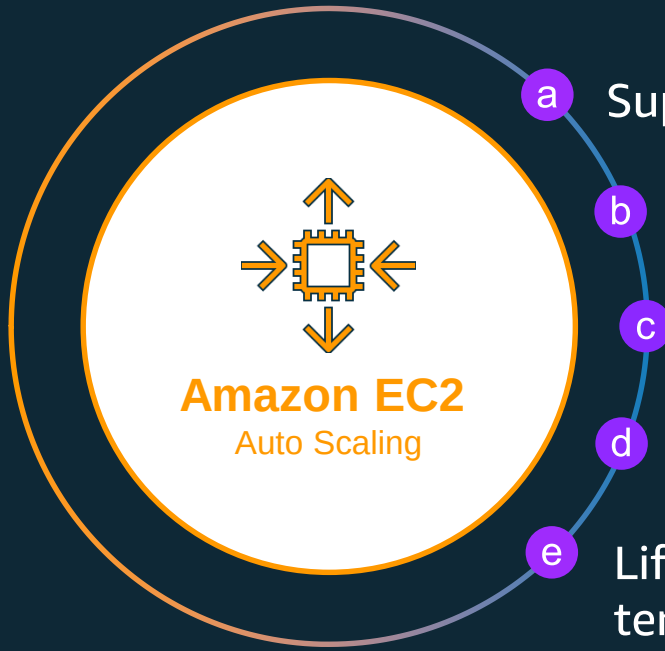
Massive scale

Simplified provisioning

Instant Fleets: Drop-In replacement for RunInstances

aws

Amazon EC2 Auto Scaling groups



a. Support for mixed instance types & Purchase Opt.

b. Replacement of unhealthy & interrupted instances

c. Balanced capacity across availability zones

d. Elastic Load Balancer integration

e. Lifecycle hooks, termination policies, instance termination protection, scaling processes…

aws

Allocation Strategies differences between ASG and EC2/Spot Fleet

EC2/Spot Fleet spot allocation strategies:

- lowestPrice
  - InstancePoolsToUseCount
- diversified
- capacityOptimized

Auto Scaling groups spot allocation strategies:

- lowest-price (across N pools)
- capacity-optimized

Amazon Elastic Inference and Amzon EC2 Spot - Best Practices

- What is your target latency SLA for your application, and what are you constraints?
- Convert to Fp16 for lower latency and higher throughput.
- Optimize further Inference cost by using Elastic Inference with EC2 Spot instances
- Be instance type agnostic and let ASG/Fleet provide the required capacity according to the allocation strategy
- Adopt Launch Templates to benefit from new ASG and Fleet features
- Architect for fault-tolerance to be Spot compatible and increase your availability
- De-couple storage and compute

aws

Demo

https://github.com/awslabs/ec2-spot-labs/tree/master/ec2-spot-elastic-inference

Thank You!

Chakravarthy Nagarajan(chakravn@amazon.com)

https://aws.amazon.com/ec2/spot/

https://github.com/awslabs/ec2-spot-labs