Track 3 | Session 5

使用 Amazon EC2 打造企業計算平台與成本和容量優化

Jack Hsu
Partner Solutions Architect
Amazon Web Services

aws SUMMIT ONLINE

Agenda

Amazon EC2 foundations

Broadest and deepest platform for enterprise workloads

Enterprise workload examples

- High performance computing (HPC)
- Machine learning infrastructure
- Windows on AWS
- SAP on AWS
- VMware Cloud on AWS

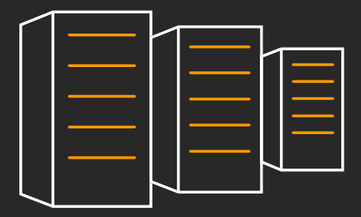
Optimizing Amazon EC2 cost and capacity

Amazon EC2 foundations

Broadest and deepest platform for enterprise workloads

aws SUMMIT ONLINE

Amazon Elastic Compute Cloud (Amazon EC2) foundations



Resources

Instances

Storage

Networking

Availability

Regions and Availability Zones

Load balancing

Automatic scaling

Management

Deployment

Monitoring

Administration

Purchasing options

On-Demand Instances

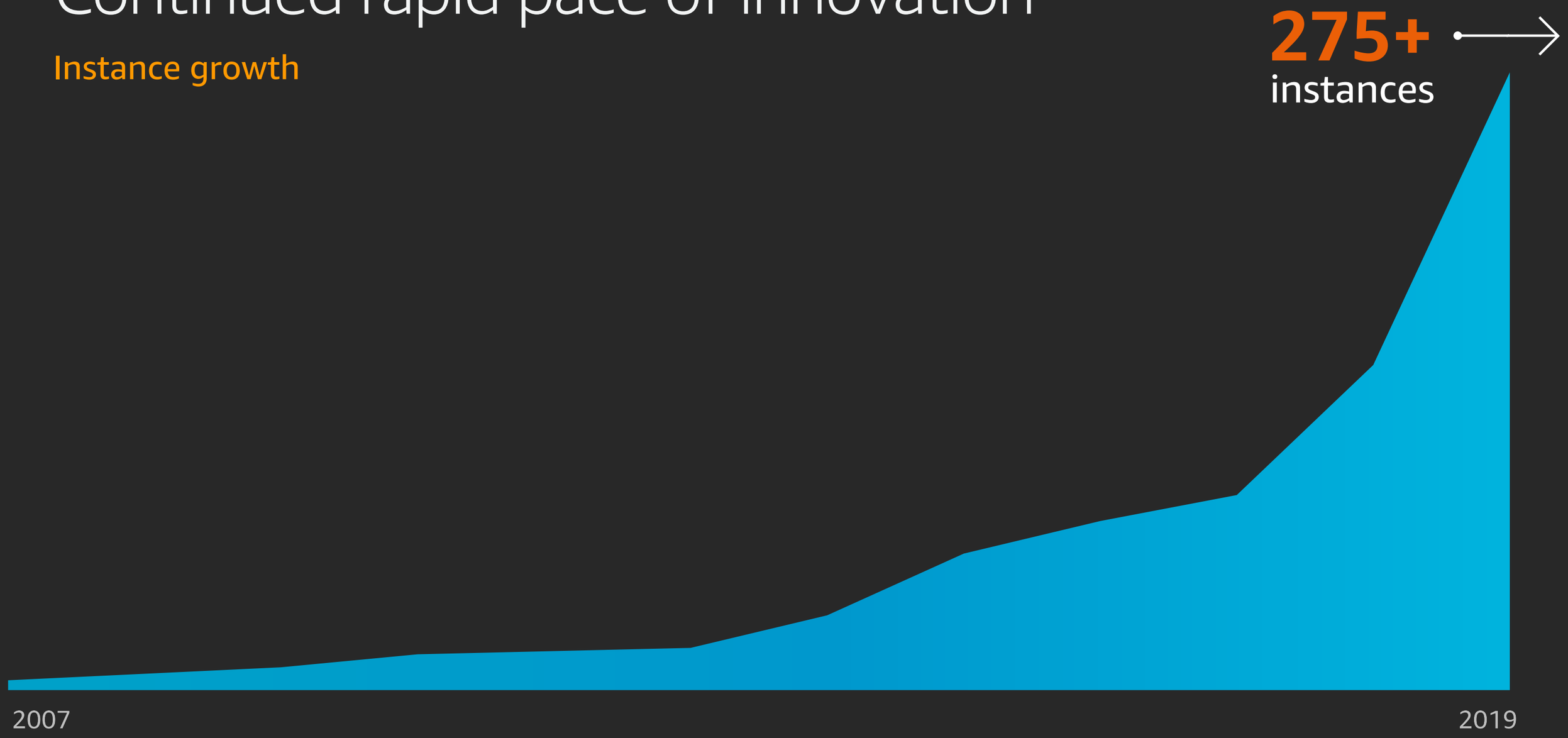
Reserved Instances

Spot Instance

Savings Plan

Continued rapid pace of innovation

Instance growth



Broadest and deepest platform choice

Workloads

General-purpose

Burstable

Compute-intensive

Memory-intensive

Storage (high I/O)

Dense storage

GPU compute

Graphics-intensive

Inference

+

Capabilities

Choice of processor
(AWS, Intel, AMD)

Fast processors
(up to 4.0 GHz)

High memory footprint
(up to 24 TiB)

Instance storage
(HDD and NVMe)

Accelerated computing
(GPUs and FPGA)

Networking
(up to 100 Gbps)

Bare metal

Size
(Nano to 32xlarge)

+

Options

Amazon Elastic Block Store

Elastic graphics

Amazon Elastic Inference

=

275+

instance types

for virtually every workload and business need

Accelerated computing workloads

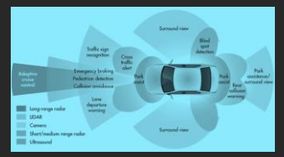
Applications that benefit from hardware acceleration

Machine learning/AI

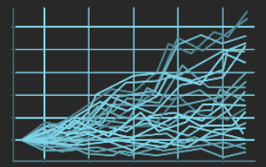
Image and video recognition

Natural language processing

Autonomous vehicle systems

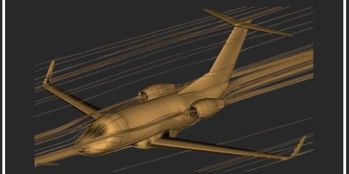


Personalization & recommendation



High-performance computing

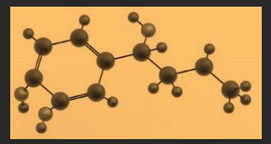
Computational fluid dynamics



Financial and data analytics

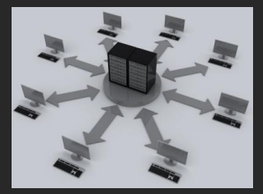
Genomics

Computational chemistry



Graphics

Virtual graphic workstation



3D modeling & rendering

Video encoding

AR/VR

Figure 1. Magic Quadrant for Cloud Infrastructure as a Service, Worldwide

Source: Gartner (July 2019)

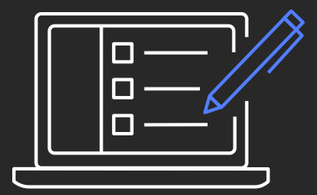
AWS recognized as a cloud leader for the 9th consecutive year

Gartner, Magic Quadrant for Cloud Infrastructure as a Service, Worldwide, Raj Bala, Bob Gill, Dennis Smith, David Wright, July 2019. ID G00365830.

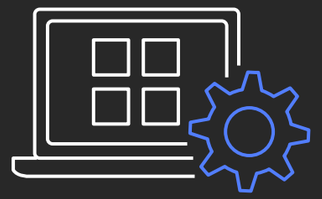
Enterprise workload examples

Compute platform optimized for enterprise apps

HPC



Machine learning



Windows workloads

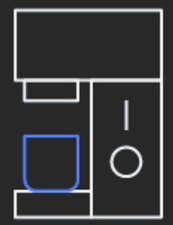
SAP

SAP

VMware Cloud on AWS

HPC impacts your life every day

Your morning coffee



The car you drive

The fuel you use

Knowing the weather

Your retirement portfolio

The movies you watch

The medicines you take

Addressing HPC technical requirements

Amazon FSx
for Lustre

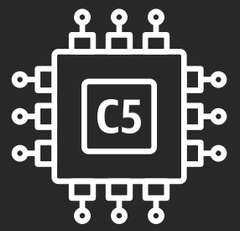
Amazon EC2
C5n instances

Elastic Fabric Adapter
+
100 Gbps networking

AWS
ParallelCluster

Run HPC in the cloud easily and securely, without compromising on price performance

Amazon EC2 C5n instances



Two key HPC-related features

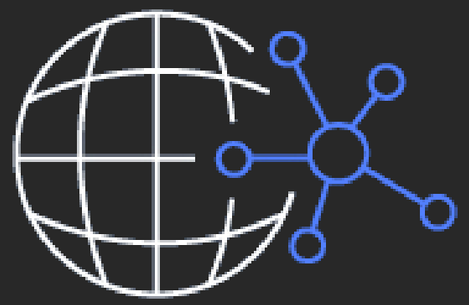
- More memory bandwidth
- 100G network throughput

Other 100G instances powered by the AWS Nitro System

P3dn, I3en, M5n, R5n, G4dn

Elastic Fabric Adapter (EFA)

SRD protocol

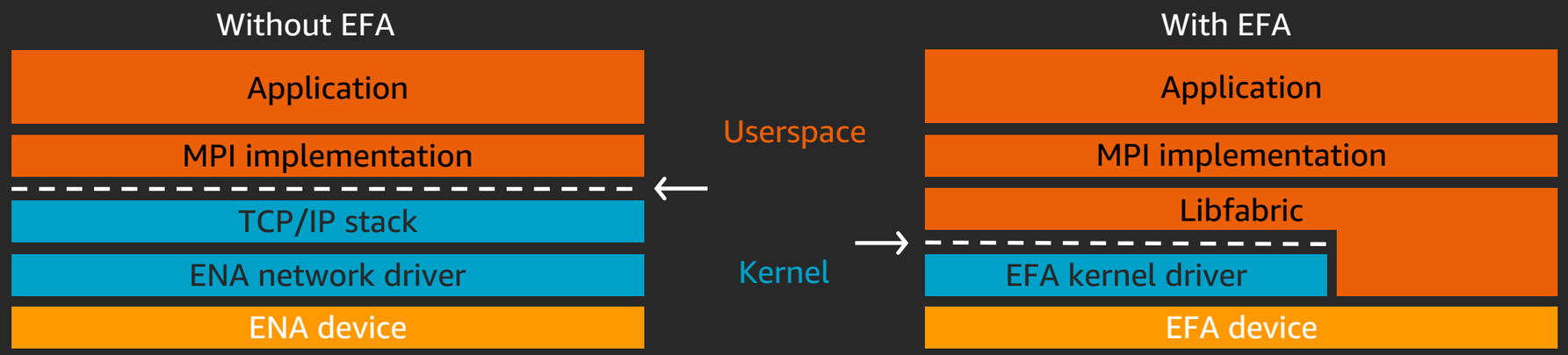


Proving myths about latency constraints wrong

CFD

Seismic

Weather modeling



MAXAR

Scale tightly coupled HPC applications on AWS

The AWS ML stack

Broadest and deepest set of capabilities

AI services

Vision			Speech		Language		Chatbots	Forecasting	Recommendations
Amazon Rekognition Image	Amazon Rekognition Video	Amazon Textract	Amazon Polly	Amazon Transcribe	Amazon Translate	Amazon Comprehend / Amazon Comprehend Medical	Amazon Lex	Amazon Forecast	Amazon Personalize

ML services

Amazon SageMaker

Ground Truth	Notebooks	Algorithms + AWS Marketplace	Reinforcement learning	Training	Optimization	Deployment	Hosting

ML frameworks + infrastructure

Frameworks	Interfaces
TensorFlow, mxnet, PYTORCH	GLUON, Keras

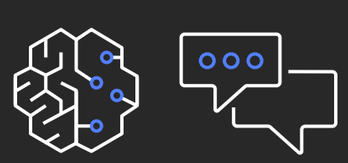
Infrastructure						
EC2 P3 & P3DN	EC2 G4	EC2 C5	FPGAs	AWS IoT Greengrass	Amazon Elastic Inference	AWS Inferentia

Machine learning use cases

Applications that benefit from accelerated compute

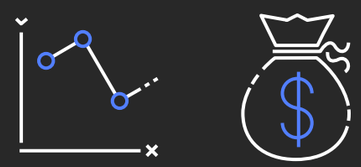
Machine learning/AI

Natural language processing



Image/Video analysis

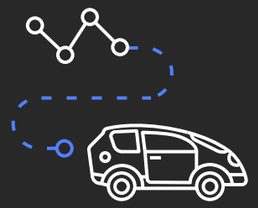
Financial services



Healthcare & life sciences

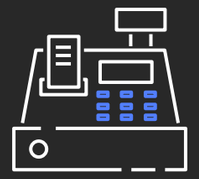
Manufacturing

Autonomous vehicle systems



Recommendation systems

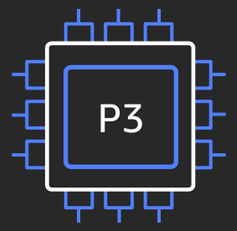
Retail

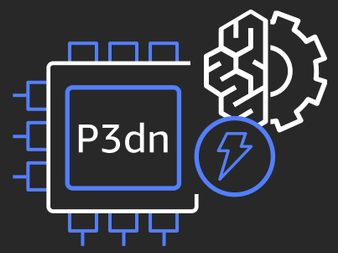


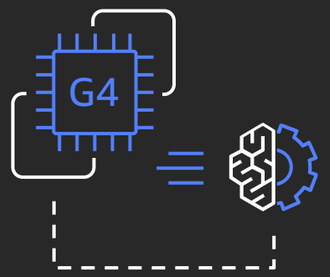
Travel & hospitality

Energy

Accelerated compute portfolio for machine learning







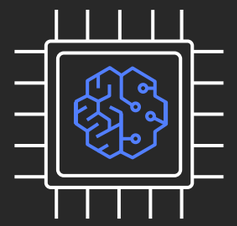
ML training

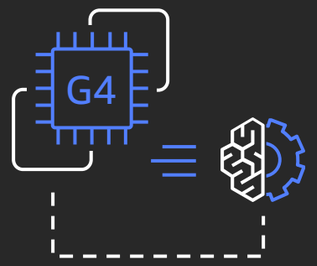
P3/P3dn GPU compute instance

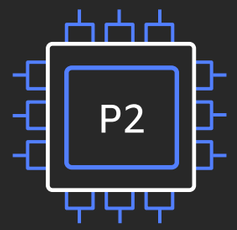
- Up to 1 petaflop of compute with 8x NVIDIA V100 GPUs
- Up to 256 GB of GPU memory
- Up to 100 Gbps of networking
- Designed to handle large distributed training jobs for fastest time to train

G4: GPU compute instance

- Up to 520 teraflops of compute with 8x NVIDIA T4 GPUs
- Cost-effective small-scale training jobs







ML inference

AWS Inf1 instance

- Up to 2,000 TOPS with 16x AWS Inferentia accelerators
- Lowest cost per inference in the cloud
- Designed for high throughput and low latency

G4: GPU compute instance

- Up to 1,030 TOPS of compute with 8x NVIDIA T4 GPUs
- Increased performance, lower latency, and reduced cost per inference compared to previous GPU-based instances

P2: GPU compute instance

- Up to 160 teraflops of compute with 16x NVIDIA K80 GPUs
- General-purpose GPU compute

Running Microsoft applications on AWS

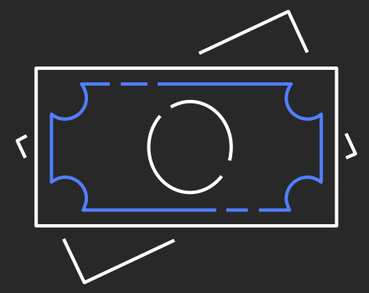
Self-managed using
Amazon EC2

- BYOL Windows & SQL Server
- Purchase EC2 Windows + BYOL SQL
- Purchase EC2 Windows + SQL Server

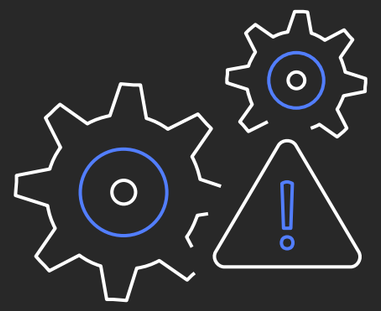
As a managed service using
Amazon RDS

- RDS for SQL Server

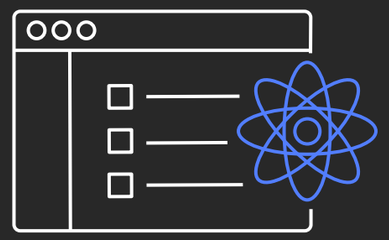
Customers have asked for ways to optimize TCO for Windows workloads



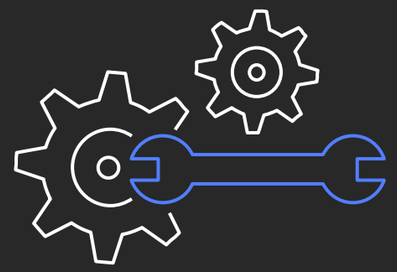
Leverage existing software licensing investments



Get elasticity of Amazon EC2 for BYOL instances that require Dedicated Hosts



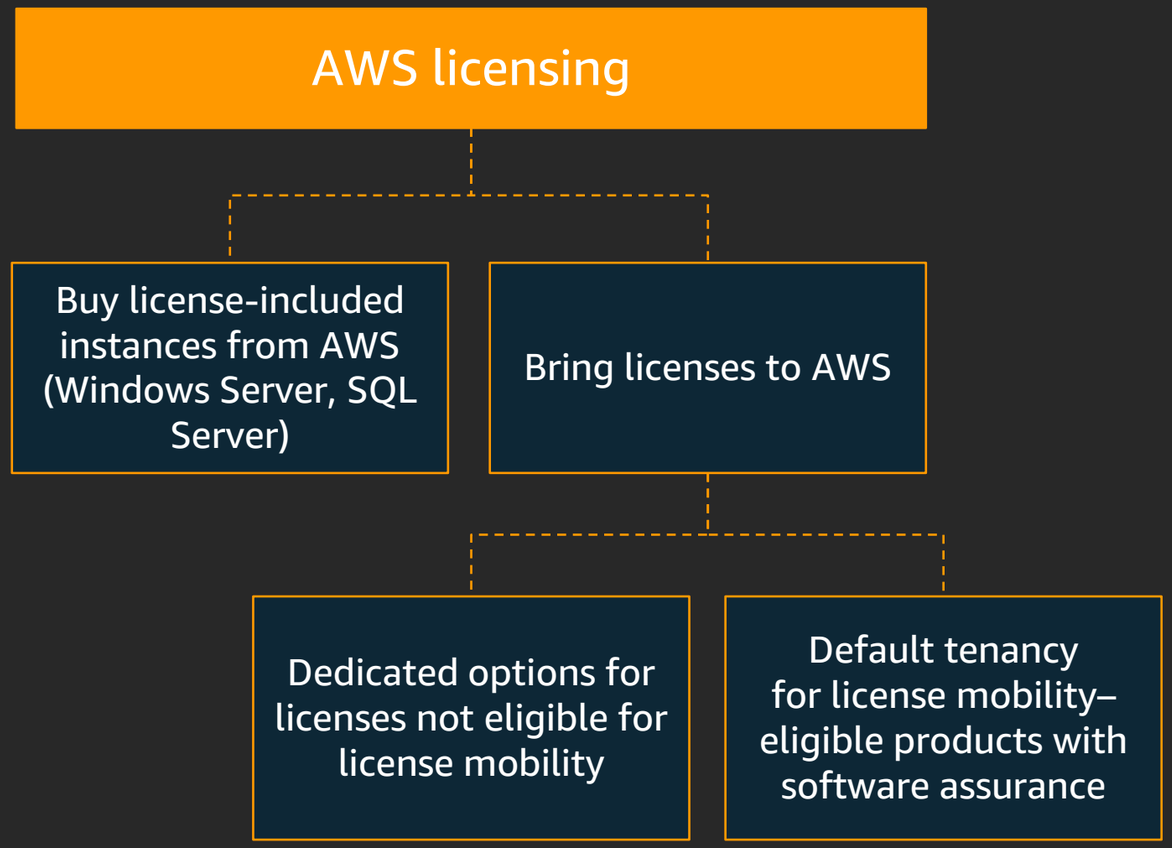
Improve visibility of license usage across hybrid environments



Simplify upgrade management experience

Flexible options for Microsoft licenses on the AWS Cloud

1. Flexible **pay-as-you-go** licensing choices
2. Bring your **license mobility** benefits to AWS
3. Bring licenses to AWS **without paying for software assurance**



Our experiences with customers point to 4 dominant modernization pathways

AWS Lambda with .NET and PowerShell

ancestry

.NET to Serverless

SQL Server on Windows to Amazon Aurora

GrubHub

Amazon ECS for Windows

Zocdoc

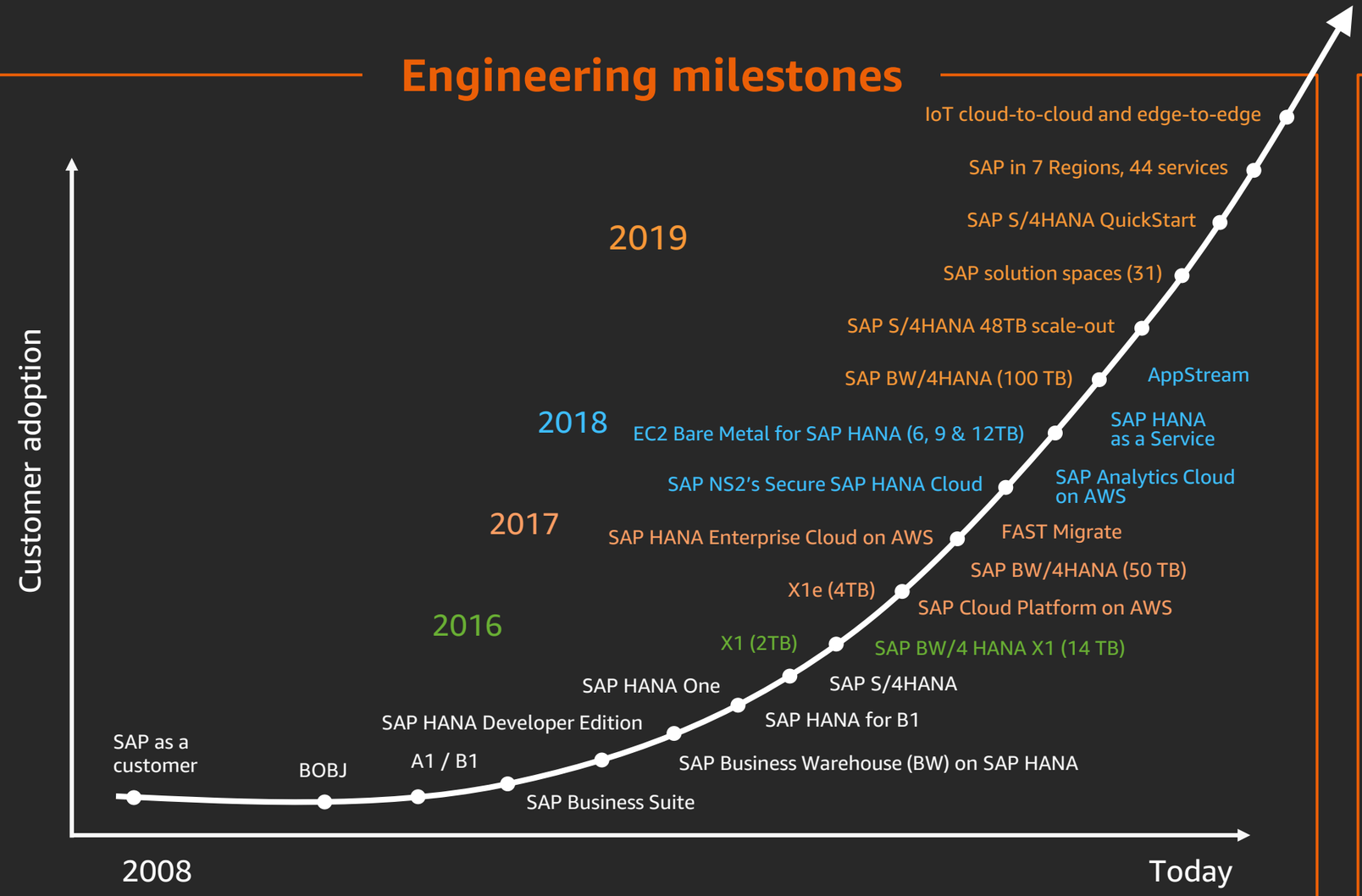
.NET to Containers

SQL Server on Windows to Linux

Decisiv

SAP on AWS: Unmatched pace of innovation

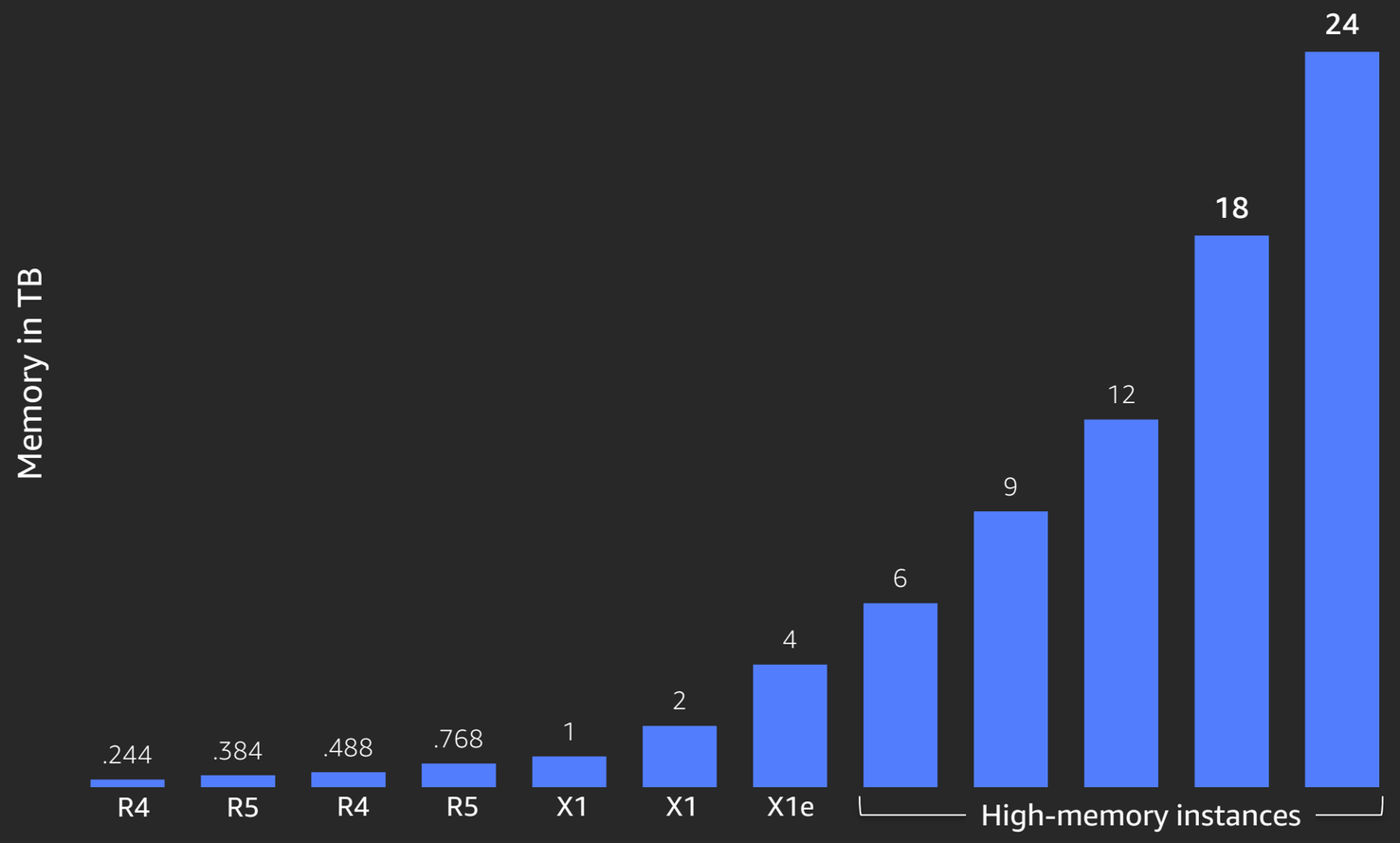
Engineering milestones



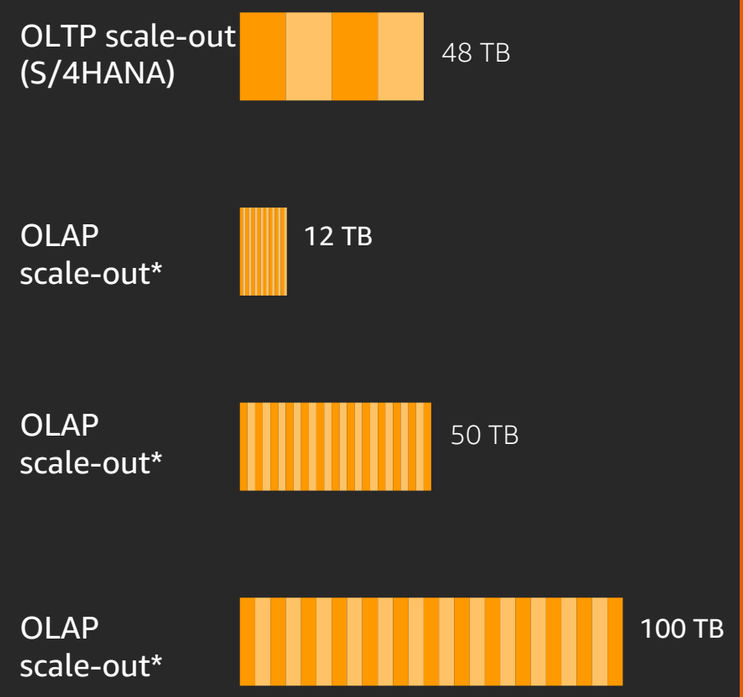
SAP on AWS

Amazon EC2 instances for SAP HANA

Scale-up options



Scale-out options

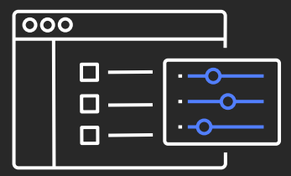


*BWoH, BW/4HANA, and data mart

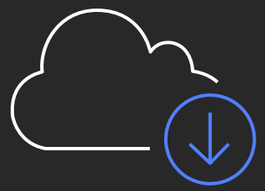
VMware Cloud on AWS

VMware software-designed data center (SDDC) technologies you know and trust, delivered as a service on the world's most popular public cloud

Rich VMware SDDC delivered as a cloud service on AWS

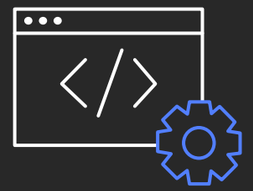


Consistency and familiarity of VMware technologies



Easy workload portability and hybrid capabilities

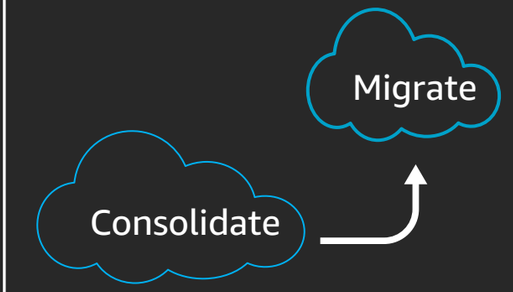
Direct access to the power of native AWS services



Existing and new apps with containers and VMs

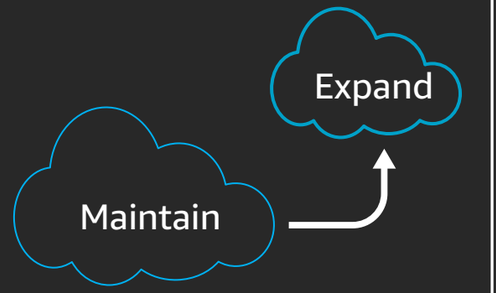
VMware Cloud on AWS Customer Use Cases

1 Cloud Migrations



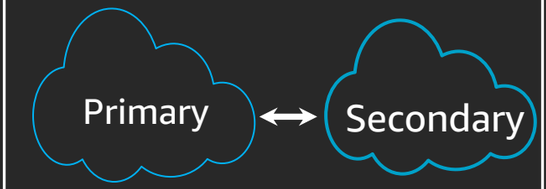
- Application Specific
- Data Center Wide
- Infrastructure Refreshes

2 Data Center Extension



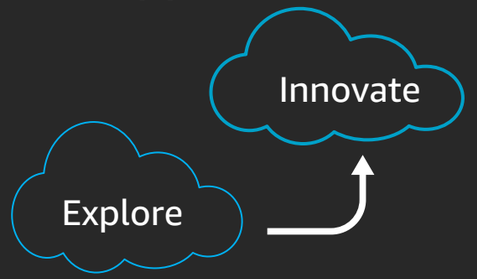
- Footprint Expansion
- On-Demand Capacity
- Test / Dev

3 Disaster Recovery



- For Additional Workloads
- DR Data Center Replacement
- Add or Modernize DR solutions

4 Next-Generation Applications



- Application Modernization
- Next-Gen App Build Out
- Enterprise Workloads

Optimizing Amazon EC2 cost and capacity

aws SUMMIT ONLINE

Optimizing Amazon EC2 cost and capacity

We continue to innovate for our customers

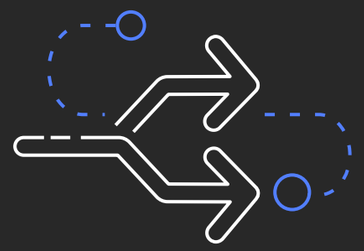
Pricing

Achieve optimal price/performance with different purchase models

Capacity

Capacity management made easy on the broadest and deepest compute platform

Guidance

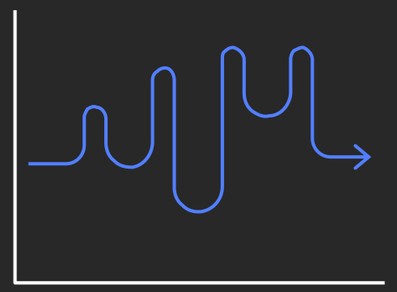


Cost and capacity recommendations enable ease of use and save time

Amazon EC2 purchasing options

On-Demand

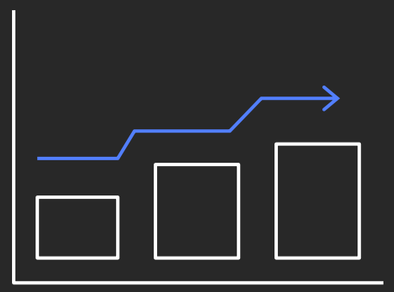
Pay for compute capacity by **the second** with no long-term commitments



Spiky workloads to define needs

Reserved Instances (RIs)

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



Committed and steady-state usage

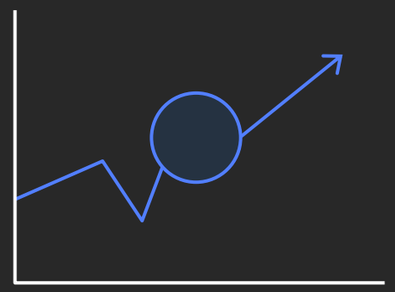
Savings Plans

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

Flexible access to compute

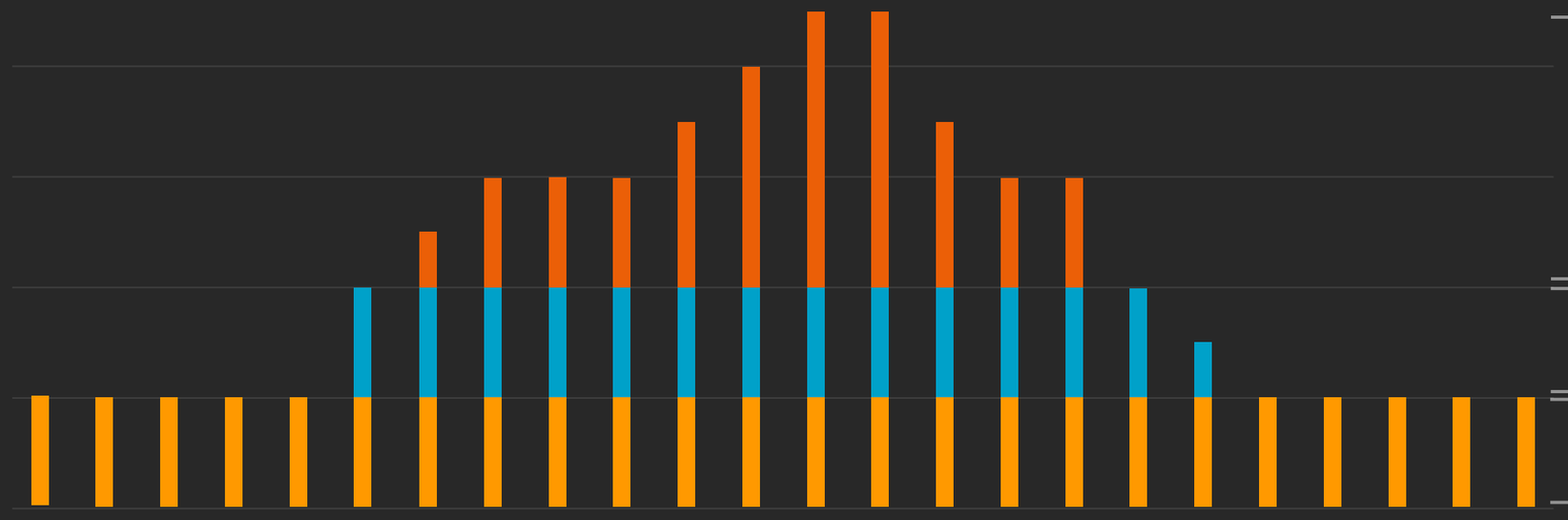
Spot Instances

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



Fault-tolerant, flexible, stateless workloads

To optimize Amazon EC2, combine purchase options



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

Use **On-Demand** for new or stateful spiky workloads

Use **RIs or a Savings Plan** for known, steady-state workloads

Types of Savings Plans

Compute Savings Plans

Offer the greatest flexibility, up to 66% off (same prices as Convertible RIs)

Flexible across

- ✓ Instance family: e.g., Move from C5 to M5
- ✓ Region: e.g., Change from EU (Ireland) to EU (London)
- ✓ OS: e.g., Windows to Linux
- ✓ Tenancy: e.g., Switch Dedicated tenancy to Default tenancy
- ✓ Compute options: e.g., Move from EC2 to Fargate

EC2 Instance Savings Plans

Provide the lowest prices, up to 72% off (same as Standard RIs) on the selected instance family (e.g., C5 or M5), in a specific AWS Region

Flexible across

- ✓ Size: e.g., Move from m5.xl to m5.4xl
- ✓ OS: e.g., Change from m5.xl Windows to m5.xl Linux
- ✓ Tenancy: e.g., Modify m5.xl Dedicated to m5.xl Default tenancy

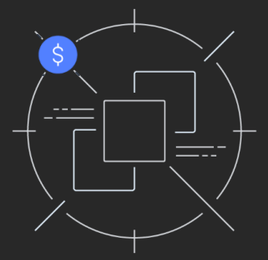
Spot, On-Demand capacity reservations, and Savings Plan together

Savings Plan

+

On-Demand capacity reservations

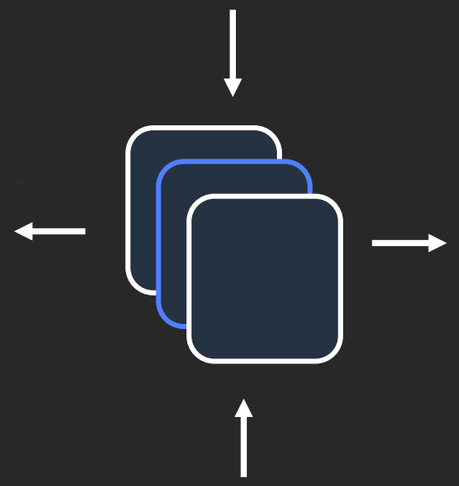
+



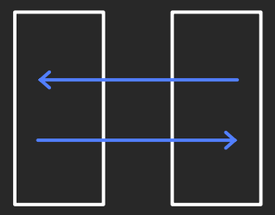
Spot Instances

=

Cost-effective, scalable compute



Save up to 90% using EC2 Spot Instances



Instances

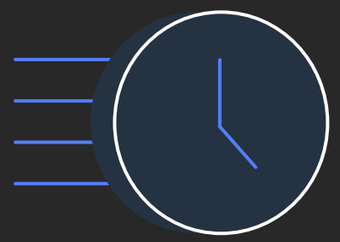
Same infrastructure as On-Demand and RIs

Pricing

Smooth, infrequent changes, more predictable

Usage

Choose different instance types, sizes, and AZs in a single fleet or EC2 Auto Scaling group

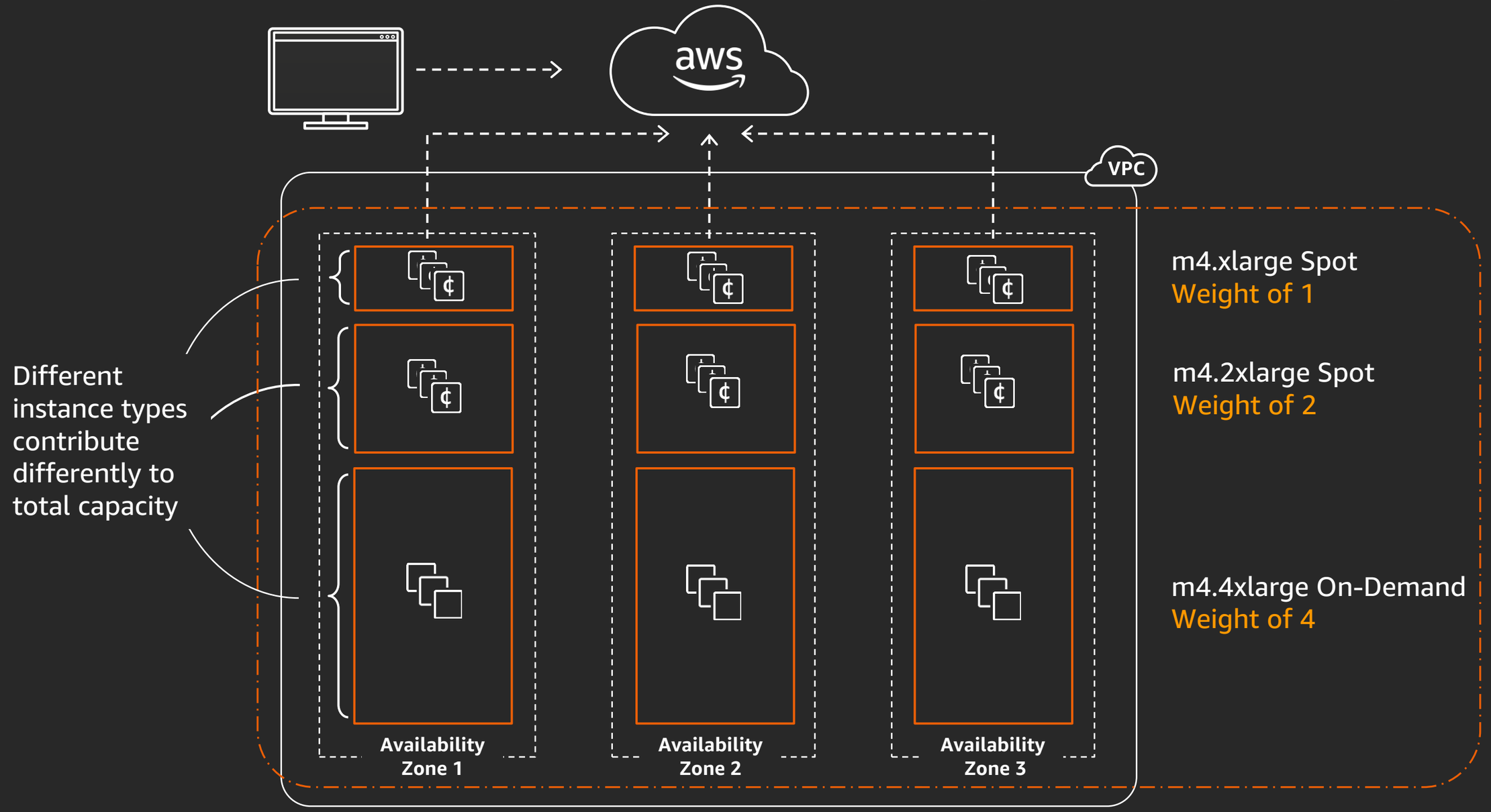


Capacity

Interruptions only happen if OD needs capacity

Pricing is based on long-term supply and demand trends; **no bidding!**

Now: Spot, On-Demand, and RIs in a single ASG with weights



Simplifying compute optimization

AWS Compute Optimizer

Identify optimal AWS compute resources for your workloads

AWS resources metadata

Amazon CloudWatch metrics

Mettle scans your AWS infrastructure and uses machine learning to automatically identify optimal AWS resources for your workloads

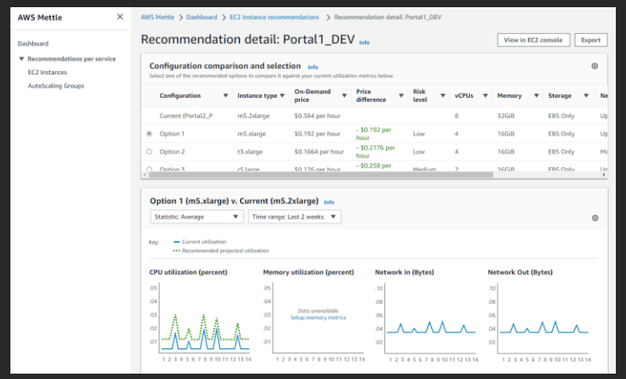
Identifies workload characteristics and profile based on the data gathered

Matches the resource requirements of your workloads to optimal AWS resources with recommendations

EC2 instance

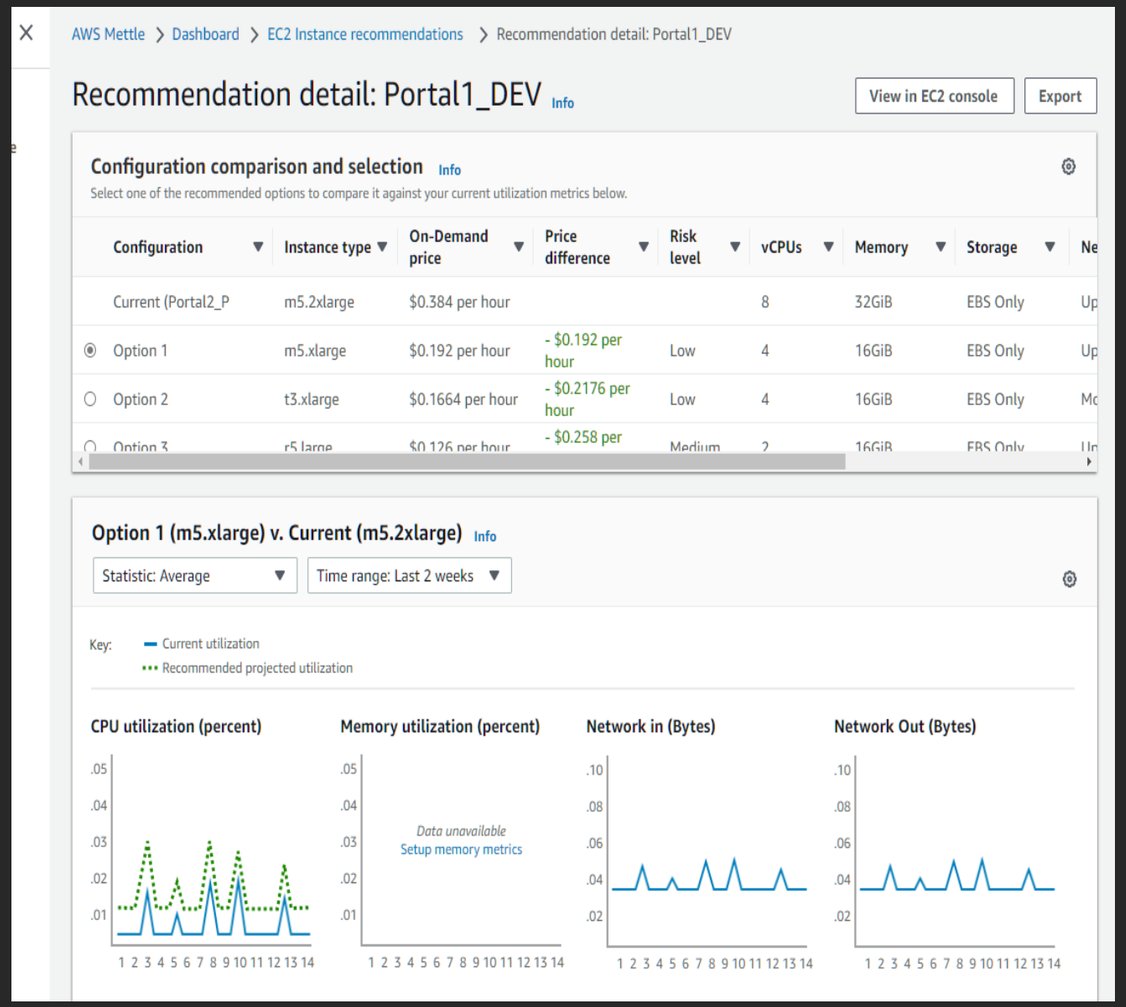
EC2 Auto Scaling groups

Helps you visualize what-if scenarios based on the recommended resources



Easy to choose with AWS Compute Optimizer

New services that recommend optimal AWS compute resources to reduce costs up to 25%



Recommends optimal EC2 instances

Optimizes performance and reduces costs by making recommendations to help you right-size compute to your workloads

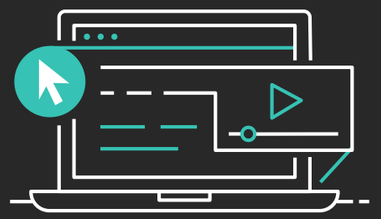
Analyzes Amazon CloudWatch metrics and considers Auto Scaling group configuration for intuitive and actionable recommendations

Up to three recommendations per workload

Available at no additional charge

Learn compute with AWS Training and Certification

Resources created by the experts at AWS to help you build cloud compute skills



20+ free digital courses cover topics related to cloud compute, including introduction to the following services:

- Amazon EC2
- Amazon EC2 Auto Scaling
- AWS Systems Manager
- AWS Inferentia and Amazon EC2 Inf1 instances

Compute is also covered in the classroom offering, Architecting on AWS, which features AWS expert instructors and hands-on activities

Visit the learning library at https://aws.training

Thank you!

aws SUMMIT ONLINE