AWS re:Invent

CMP211-R

Amazon EC2 foundations

Chetan Kapoor

Principal Product Manager, EC2
Amazon Web Services

AWS re:Invent

Amazon EC2 foundations

Resources

Instances

Storage

Networking

Availability

Regions and AZs

Load Balancing

Auto Scaling

Management

Deployment

Monitoring

Administration

Purchase Options

On Demand

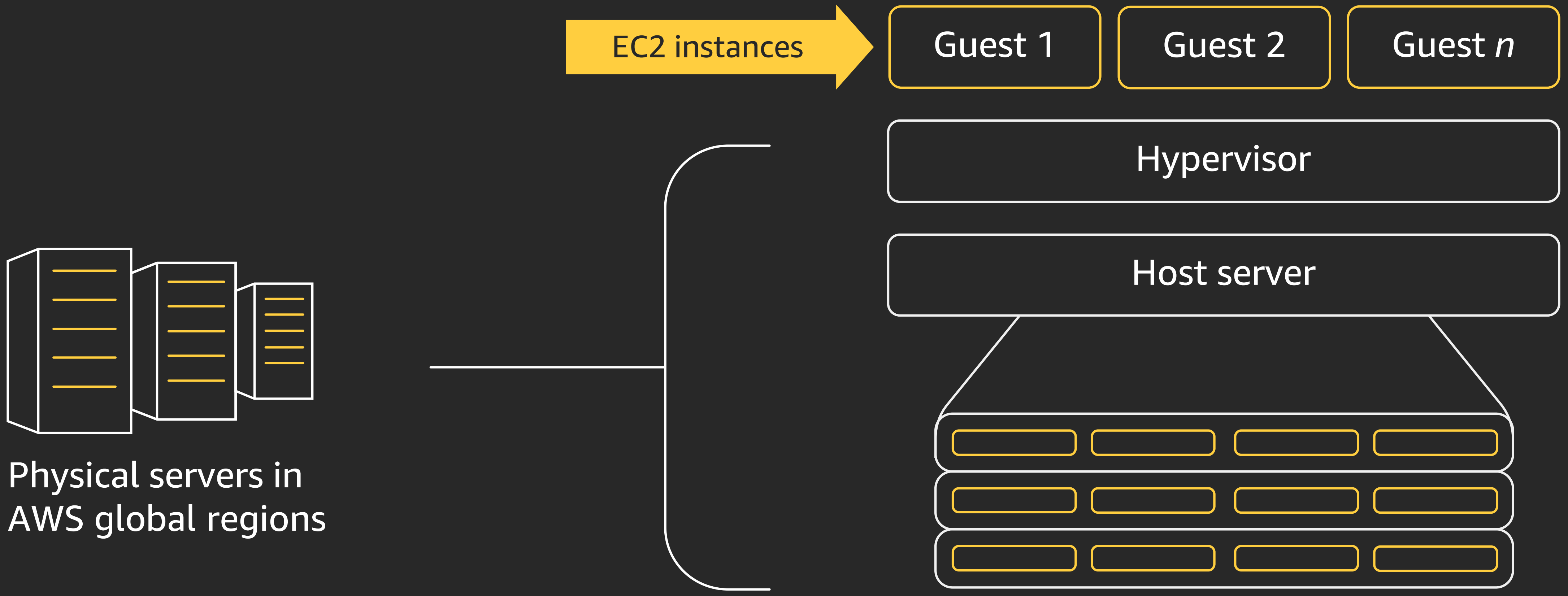
Reserved

Spot

Savings Plan

Amazon Elastic Compute Cloud (Amazon EC2)

Virtual servers in the cloud



Amazon EC2 13+ years ago…

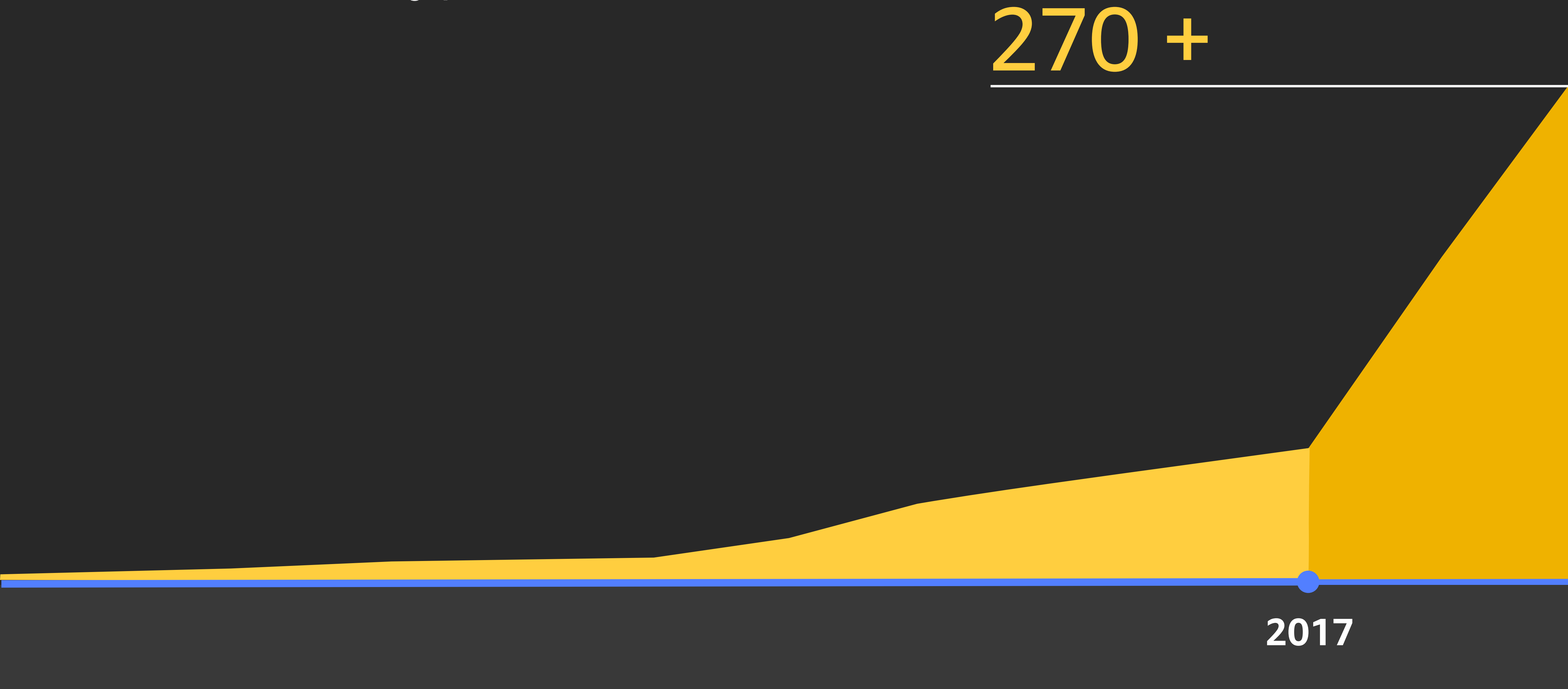
M1

“One size fits all”

Pay for what you use

Scale up or down quickly, as needed

270+ instances across 42 instance types



Journey from then to now

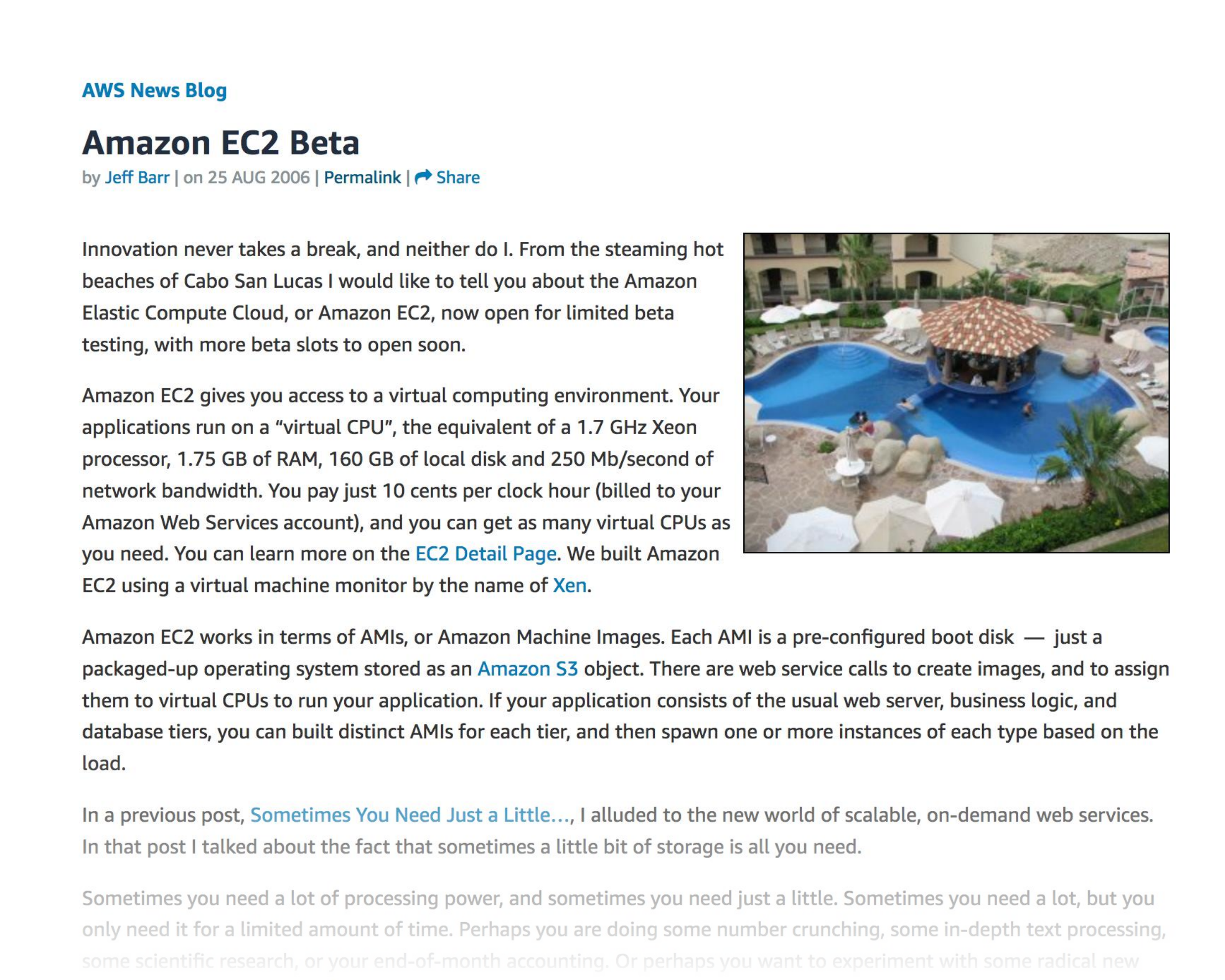
2006 "Instance"

1.7 GHz Xeon processor

1.75 GB of RAM

160 GB of local disk

250 Mbps network bandwidth



AWS News Blog

Amazon EC2 Beta

by Jeff Barr | on 25 AUG 2006 | Permalink | Share

Innovation never takes a break, and neither do I. From the steaming hot beaches of Cabo San Lucas I would like to tell you about the Amazon Elastic Compute Cloud, or Amazon EC2, now open for limited beta testing, with more beta slots to open soon.

Amazon EC2 gives you access to a virtual computing environment. Your applications run on a "virtual CPU", the equivalent of a 1.7 GHz Xeon processor, 1.75 GB of RAM, 160 GB of local disk and 250 Mb/second of network bandwidth. You pay just 10 cents per clock hour (billed to your Amazon Web Services account), and you can get as many virtual CPUs as you need. You can learn more on the EC2 Detail Page. We built Amazon EC2 using a virtual machine monitor by the name of Xen.

Amazon EC2 works in terms of AMIs, or Amazon Machine Images. Each AMI is a pre-configured boot disk — just a packaged-up operating system stored as an Amazon S3 object. There are web service calls to create images, and to assign them to virtual CPUs to run your application. If your application consists of the usual web server, business logic, and database tiers, you can built distinct AMIs for each tier, and then spawn one or more instances of each type based on the load.

In a previous post, Sometimes You Need Just a Little…, I alluded to the new world of scalable, on-demand web services. In that post I talked about the fact that sometimes a little bit of storage is all you need.

Sometimes you need a lot of processing power, and sometimes you need just a little. Sometimes you need a lot, but you only need it for a limited amount of time. Perhaps you are doing some number crunching, some in-depth text processing, some scientific research, or your end-of-month accounting. Or perhaps you want to experiment with some radical new

> Your applications run on a "virtual CPU", the equivalent of a 1.7 GHz Xeon processor, 1.75 GB of RAM, 160 GB of local disk and 250 Mbps of network bandwidth.

Journey from then to now

2006 “Instance”

1.7 GHz Xeon processor

1.75 GB of RAM

160 GB of local disk

250 Mbps network bandwidth

2019

4.0 GHz Xeon processor

z1d instance

24 TiB of RAM

High Memory instances

60 TB of NVMe local storage

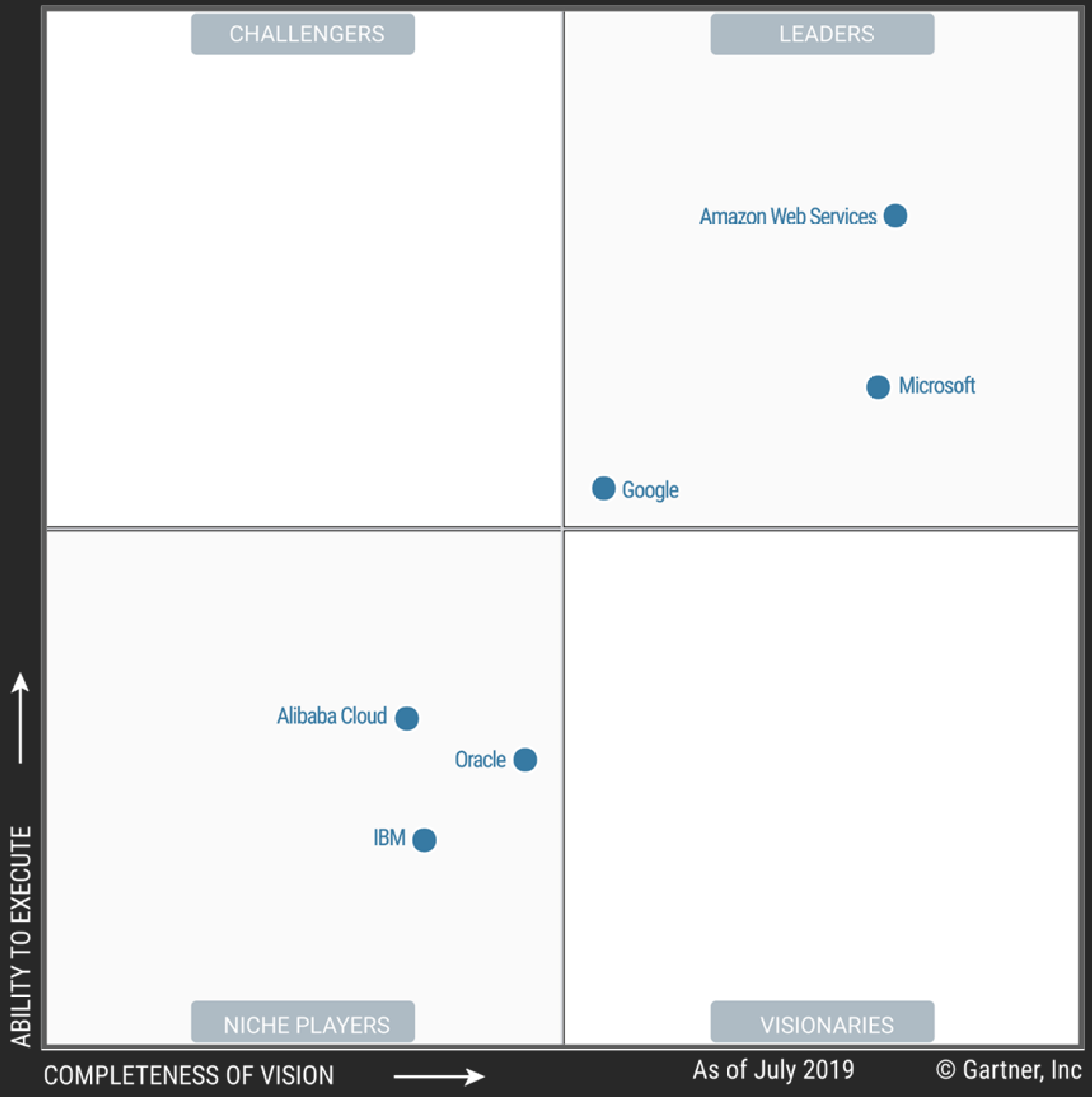
I3en.metal instances

48 TB of local disk

d2.8xlarge

100 Gbps network bandwidth

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

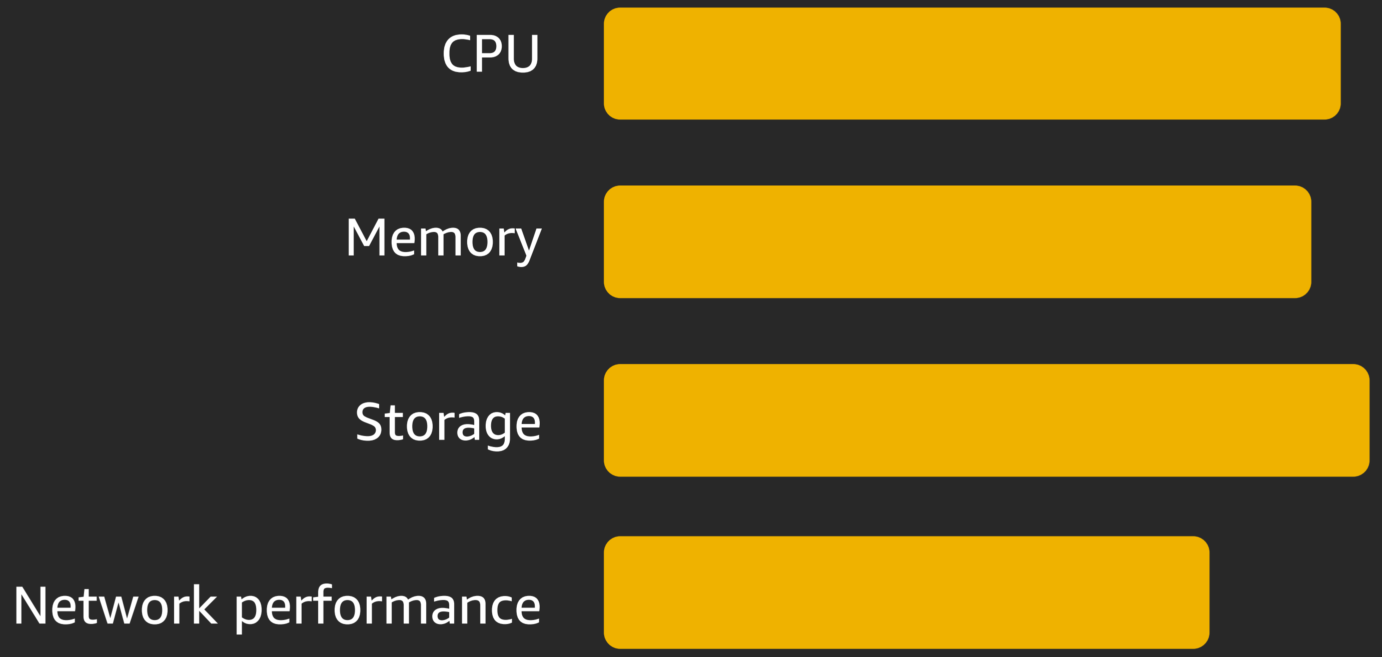


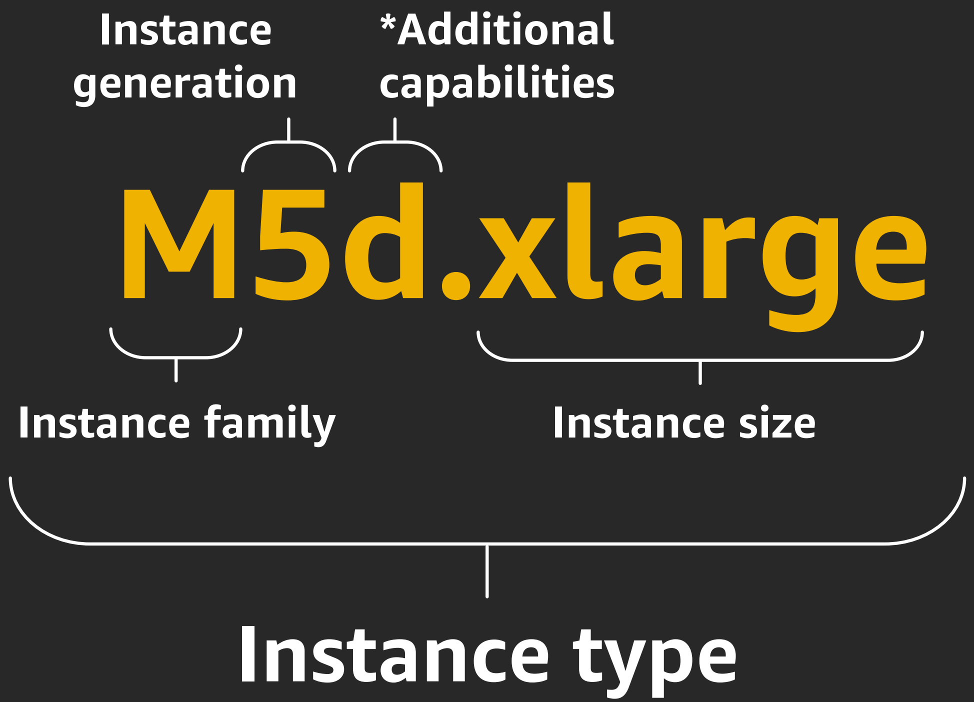
Source: Gartner (July 2019)

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. Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose. The Gartner logo is a trademark and service mark of Gartner, Inc., and/or its affiliates, and is used herein with permission. All rights reserved.

Amazon EC2 instance characteristics





Broadest choice of processors

Intel

Intel Xeon Scalable processors

AMD

AMD EPYC processors

aws

AWS Graviton processors

+ Choice of GPUs, FPGAs & Custom ASICs for compute acceleration

Right compute for the right application

Amazon Machine Images (AMIs)

Amazon maintained

Broad set of Linux and Windows images

Kept up to date by Amazon in each region

Amazon Linux 2 with five years of long-term support

Marketplace maintained

Managed and maintained by AWS Marketplace partners

Your machine images

AMIs you have created from Amazon EC2 instances

Can keep private, share with other accounts, or publish to the community

Demo:
EC2 instance launch & connect

General-purpose workloads

Web/App servers

Enterprise apps

Gaming servers

Caching fleets

Analytics applications

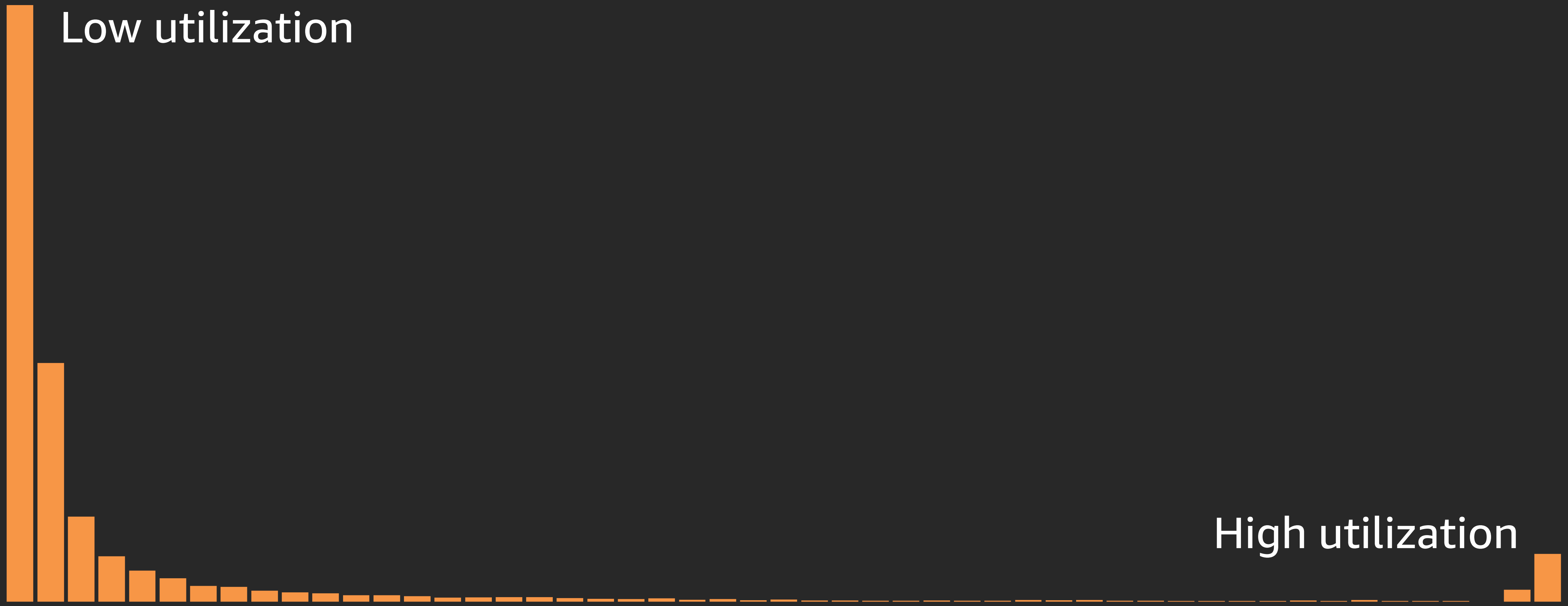
Dev/Test environments

Amazon EC2 general-purpose instances

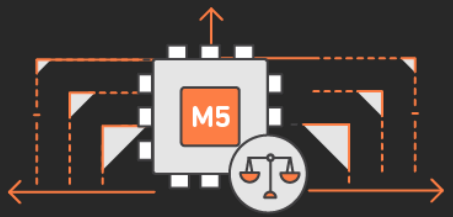
M5 instances

Balance of compute, memory, and network resources. 4:1 memory to vCPU ratio

Opportunity: Most instances aren't very busy

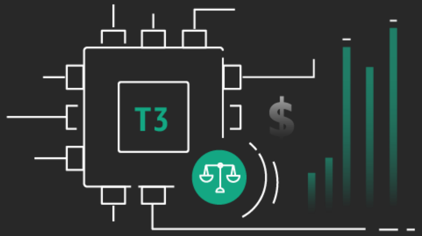


Amazon EC2 general-purpose instances



M5 instances

Balance of compute, memory, and network resources. 4:1 memory to vCPU ratio



T3 instances

Baseline level of CPU performance with the ability to burst above the baseline for workloads that don’t require sustained performance

A1 instances powered by AWS Graviton processors

AWS Graviton processor

Custom AWS silicon with 64-bit Arm Neoverse cores

Targeted workloads optimizations

Rapidly innovate, build, and iterate on behalf of customers

Amazon EC2 A1

Run scale-out and Arm-based applications in the cloud

Up to 45% cost savings

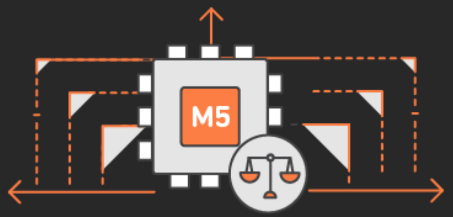
AWS Graviton Processor 64-bit Arm Neoverse cores and custom AWS silicon

Flexibility and choice for your workloads

Lower cost

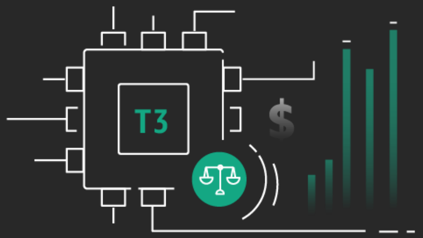
Maximize resource efficiency with AWS Nitro System

Amazon EC2 general-purpose instances



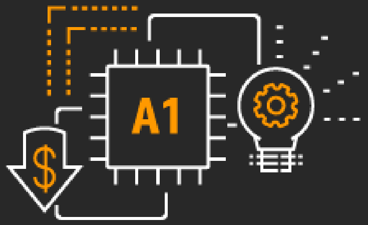
M5 instances

Balance of compute, memory, and network resources. 4:1 memory to vCPU ratio



T3 instances

Baseline level of CPU performance with the ability to burst above the baseline for workloads that don’t require sustained performance



A1 instances

Workloads that can scale out across multiple cores, fit within memory, run on ARM instructions

Announcing AWS Graviton2 processor

Graviton1 processor

First ARM-based processor in major cloud

Built on 64-bit ARM Neoverse cores with AWS designed 16 nm silicon

Up to 16 vCPUs,10 Gbps enhanced networking, 3.5 Gbps Amazon EBS bandwidth

Graviton2 processor

Built with 64-bit ARM Neoverse cores with AWS designed 7 nm silicon process

Up to 64 vCPUs, 20 Gbps enhanced networking, 14 Gbps Amazon EBS bandwidth

7x performance, 4x compute cores, and 5x faster memory

Announcing Graviton2 based instances

M6g

Available in preview

Instances with/without local instance storage

R6g **C6g**

Coming in 2020

Memory-intensive workloads

In-memory caches

High-performance databases

Big data analytics

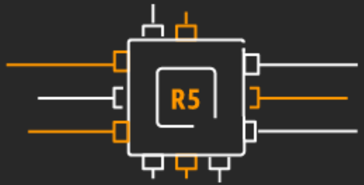
Amazon EC2 memory-optimized instances

R5 instances

Accelerate performance for workloads that process large data sets in memory

8:1 memory to vCPU ratio

Amazon EC2 memory-optimized instances



R5 instances

Accelerate performance for workloads that process large data sets in memory
8:1 memory to vCPU ratio

X1 / X1e instances

For memory-intensive workloads and very large in-memory workloads
16:1 and 32:1 memory to vCPU ratio

Amazon EC2 memory-optimized instances

R5 instances

Accelerate performance for workloads that process large data sets in memory
8:1 memory to vCPU ratio

X1 / X1e instances

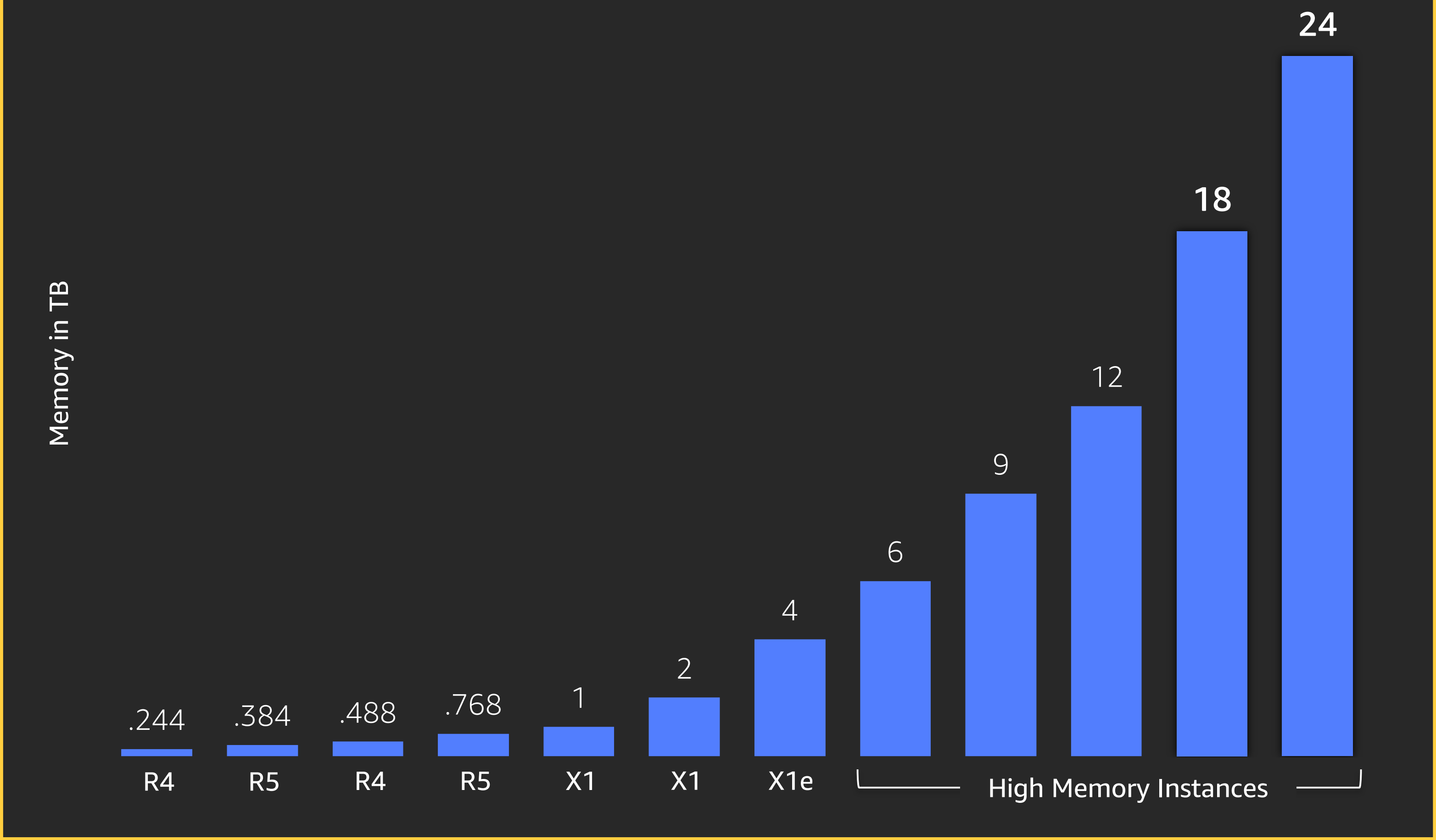
For memory-intensive workloads and very large in-memory workloads
16:1 and 32:1 memory to vCPU ratio

High memory instances

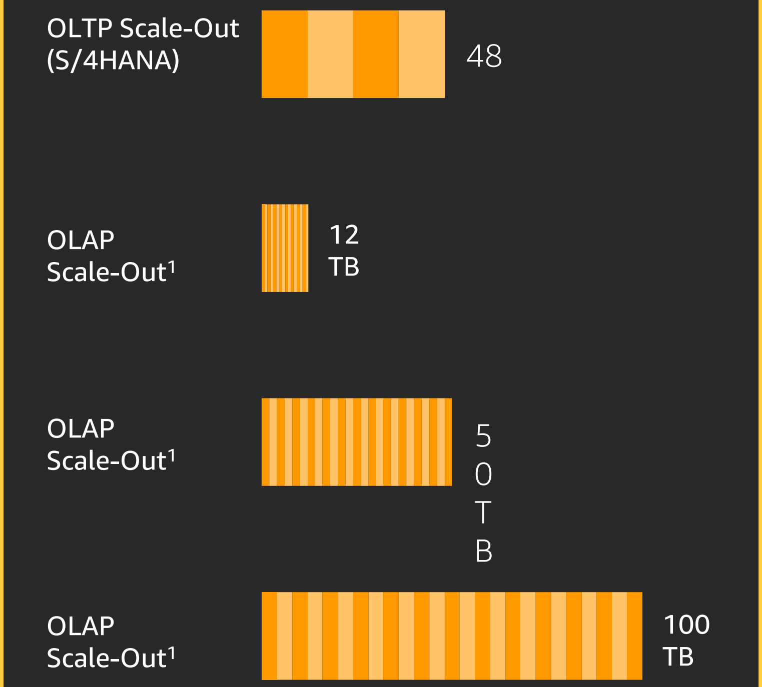
Extreme memory needs
Certified to run SAP HANA
From 6 to 24 TB of memory

Amazon EC2 instances for SAP HANA

Scale-up options



Scale-out options



[1] *BWoH, BW/4HANA and Datamart*

Compute-intensive workloads

Batch processing

Distributed analytics

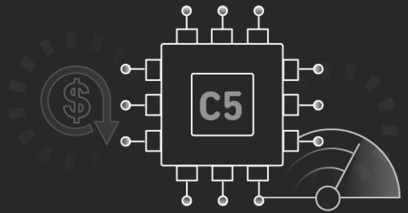
High-perf computing (HPC)

Ad serving

Multiplayer gaming

Video encoding

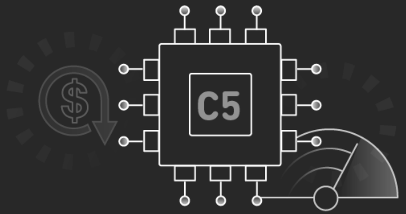
Amazon EC2 compute-optimized instances



C5 instances

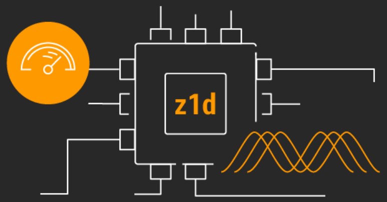
High performance at a low price per vCPU ratio
2:1 memory to vCPU ratio

Amazon EC2 compute-optimized instances



C5 instances

High performance at a low price per vCPU ratio
2:1 memory to vCPU ratio



z1d instances

High single thread performance
Fastest processor in the cloud at 4.0 GHz
8:1 memory to vCPU ratio

Storage-intensive workloads

High IO

- High-perf databases
- Real-time analytics
- Transactional workloads
- No SQL databases

Dense storage

- Big data
- Data warehousing
- Kafka
- HDFS
- MapReduce
- Log processing

Amazon EC2 storage-optimized instances

I3 / I3en instances

I/O optimized for high transaction workloads, low latency workloads

Amazon EC2 storage-optimized instances

I3 / I3en instances	I/O optimized for high transaction workloads, low latency workloads
D2 instances	Lowest cost per storage ($/GB) Supports high sequential disk throughput

Amazon EC2 storage-optimized instances

I3 / I3en instances

I/O optimized for high transaction workloads, low latency workloads

D2 instances

Lowest cost per storage ($/GB)

Supports high sequential disk throughput

H1 instances

Designed for applications that require low cost, high disk throughput and high sequential disk I/O access to very large data sets

More vCPUs and memory per TB of disk than D2

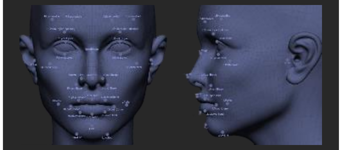
Accelerated computing workloads

Applications that benefit from hardware acceleration

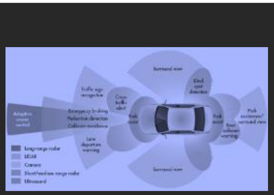
Machine learning/AI

Image and Video Recognition

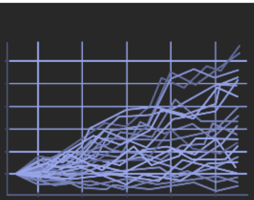
Natural Language Processing



Autonomous Vehicle Systems

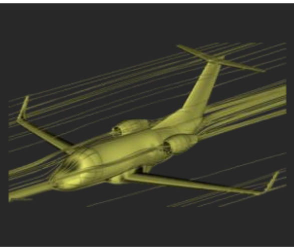


Personalization & Recommendation



High-performance computing

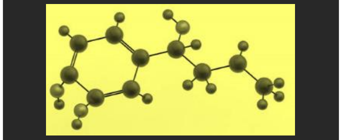
Computational Fluid Dynamics



Financial and Data Analytics

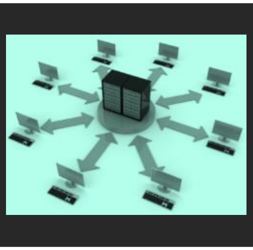
Genomics

Computational Chemistry



Graphics

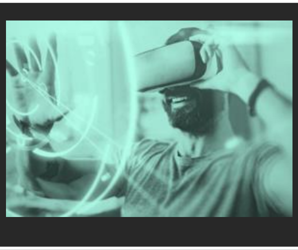
Virtual Graphic Workstation



3D Modeling & Rendering

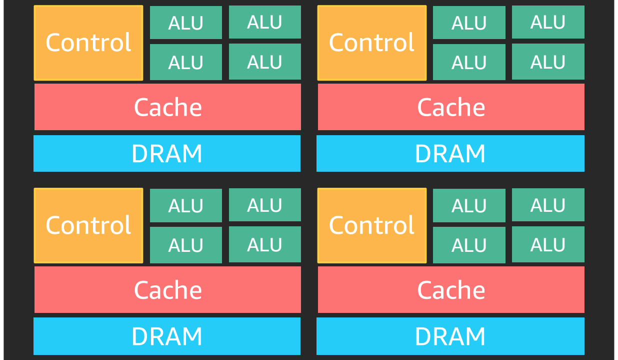
Video Encoding

AR/VR



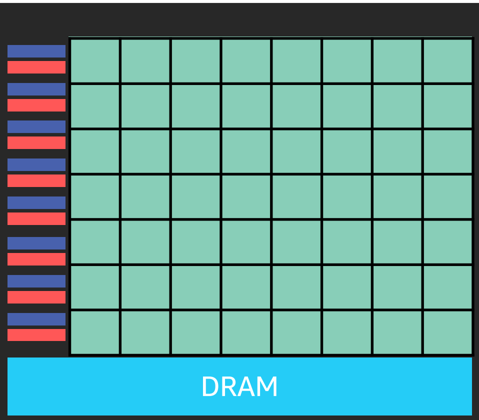
CPUs vs GPUs vs FPGA vs ASICs for compute acceleration

CPU



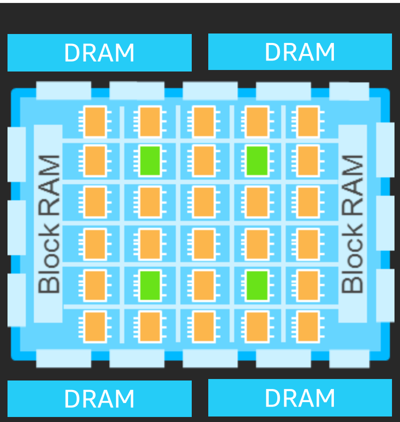
- 10s-100s of processing cores
- Pre-defined instruction set & datapath widths
- Optimized for general-purpose computing

GPU



- 1,000s of processing cores
- Pre-defined instruction set and datapath widths
- Highly effective at parallel execution

FPGA



- Millions of programmable digital logic cells
- No predefined instruction set or datapath widths
- Hardware timed execution

ASICs

- Optimized & custom design for particular use/function
- Predefined software experience exposed through API

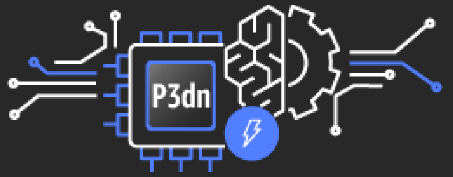
Amazon EC2 accelerated computing instances

P-Series
P2/P3 instances

GPU compute instance for use cases including deep learning training, HPC simulations, financial computing, and batch rendering

Feature latest NVIDIA high-end GPUs, including Volta V100

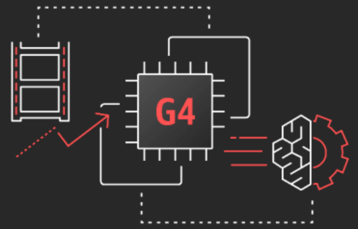
Amazon EC2 accelerated computing instances



P-Series
P2/P3 instances

GPU compute instance for use cases including deep learning training, HPC simulations, financial computing, and batch rendering

Feature latest NVIDIA high-end GPUs including Volta V100

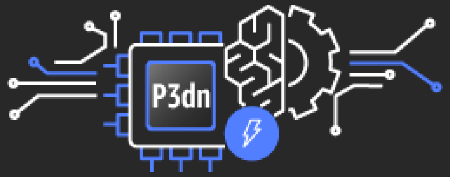


G-Series
G3/G4 instances

GPU graphics instance designed for workloads such as 3D rendering, remote graphics workstations, video encoding, and AR/VR

Feature NVIDIA mid-range GPUs such as Turing T4 GPUs, with GRID Virtual Workstation features and license

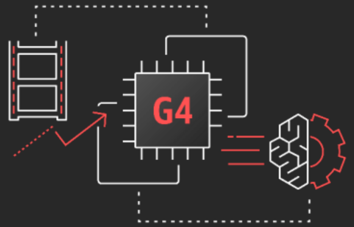
Amazon EC2 accelerated computing instances



P-Series
P2/P3 instances

GPU compute instance for use cases including deep learning training, HPC simulations, financial computing, and batch rendering

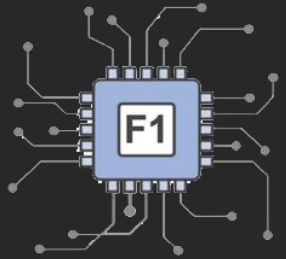
Feature latest NVIDIA high-end GPUs including Volta V100



G-Series
G3/G4 instances

GPU graphics instance designed for workloads such as 3D rendering, remote graphics workstations, video encoding, and AR/VR

Feature NVIDIA mid-range GPUs such as Turing T4 GPUs, with GRID Virtual Workstation features and license



FPGA instances
F1 instances

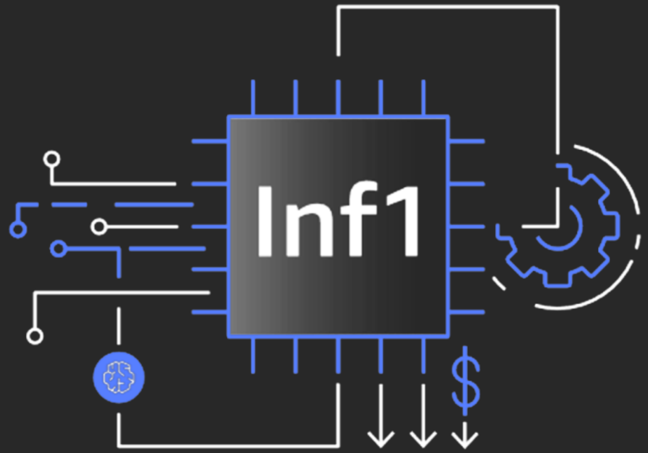
Customer programmable FPGAs that provide dramatic performance improvements for applications such as financial computing, genomics, accelerated search, and image processing

Feature Xilinx Virtex UltraScale+ VU9P FPGAs in a single instance

Programmable via VHDL, Verilog, or OpenCL

Announcing Inf1 instances

Announcing Inf1 instances



High performance and the lowest cost machine learning inference in the cloud

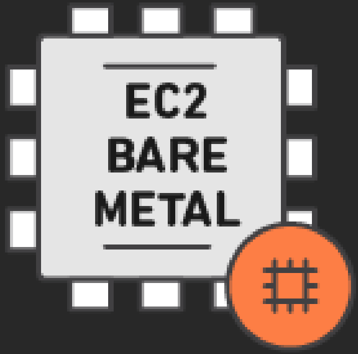
40% lower cost-per-inference than any Amazon EC2 GPU instance

2x higher inference throughput with up to 2,000 TOPS at sub-millisecond latency

Integration with popular ML frameworks TensorFlow, PyTorch, and MXNet

EC2 Bare Metal

*Run bare metal workloads on EC2
with all the elasticity, security, scale,
and services of AWS*



Designed for workloads that are not virtualized, require specific types of hypervisors, or have licensing models that restrict virtualization

It all starts with our investments in the Nitro platform

Nitro Card

Local NVMe storage
Amazon Elastic Block Storage
Networking, monitoring, and security

Nitro Security Chip

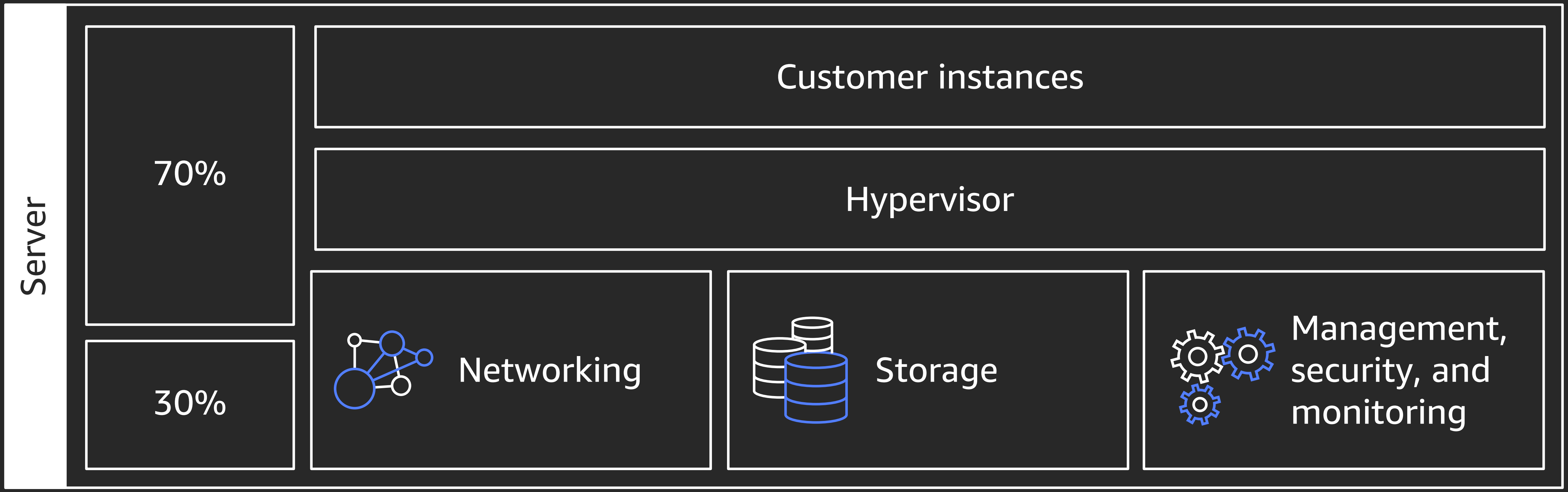
Integrated into motherboard
Protects hardware resources

Nitro Hypervisor

Lightweight hypervisor
Memory and CPU allocation
Bare Metal-like performance

Modular building blocks for rapid design and delivery of EC2 instances

EC2 “instance” host architecture



Server

70%

30%

Customer instances

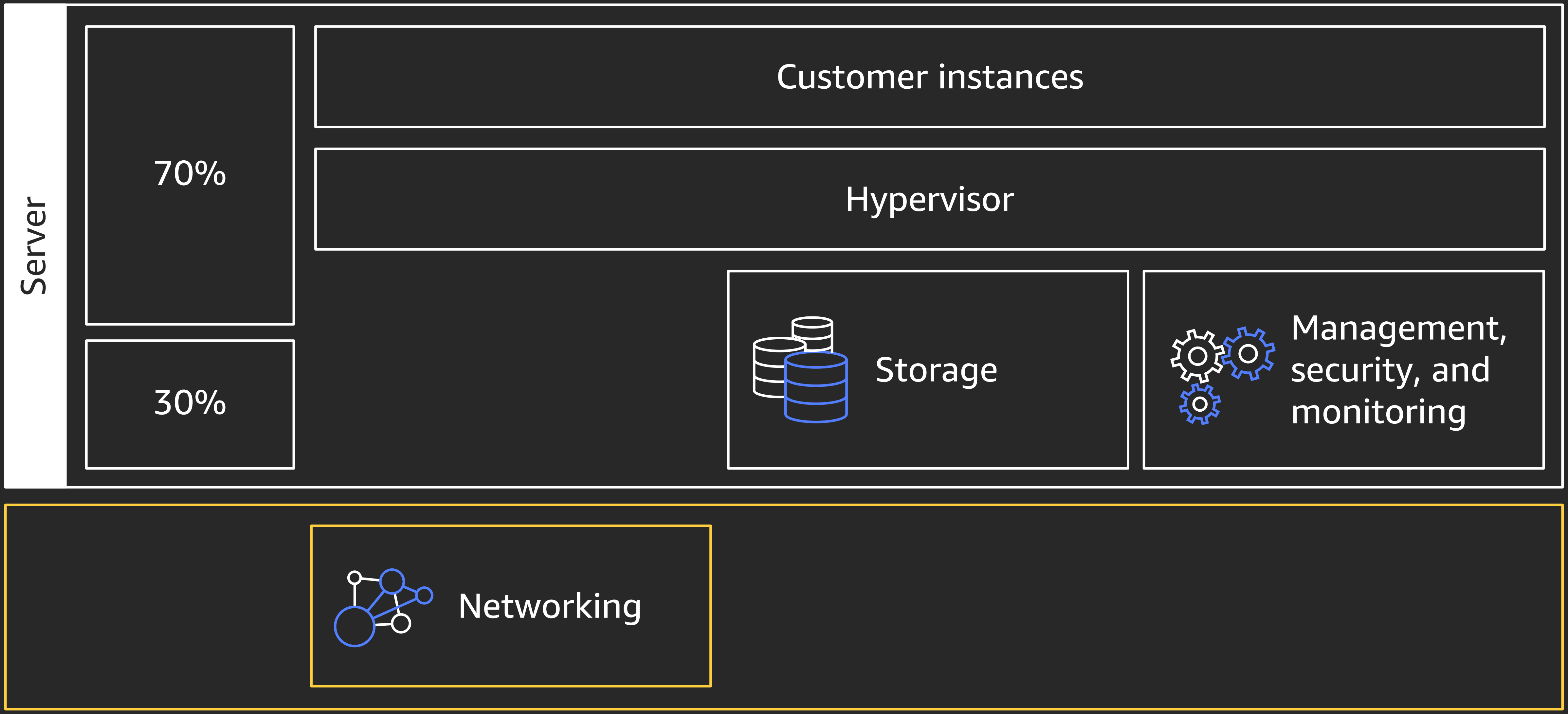
Hypervisor

Networking

Storage

Management, security, and monitoring

2012: EC2 “instance” host architecture

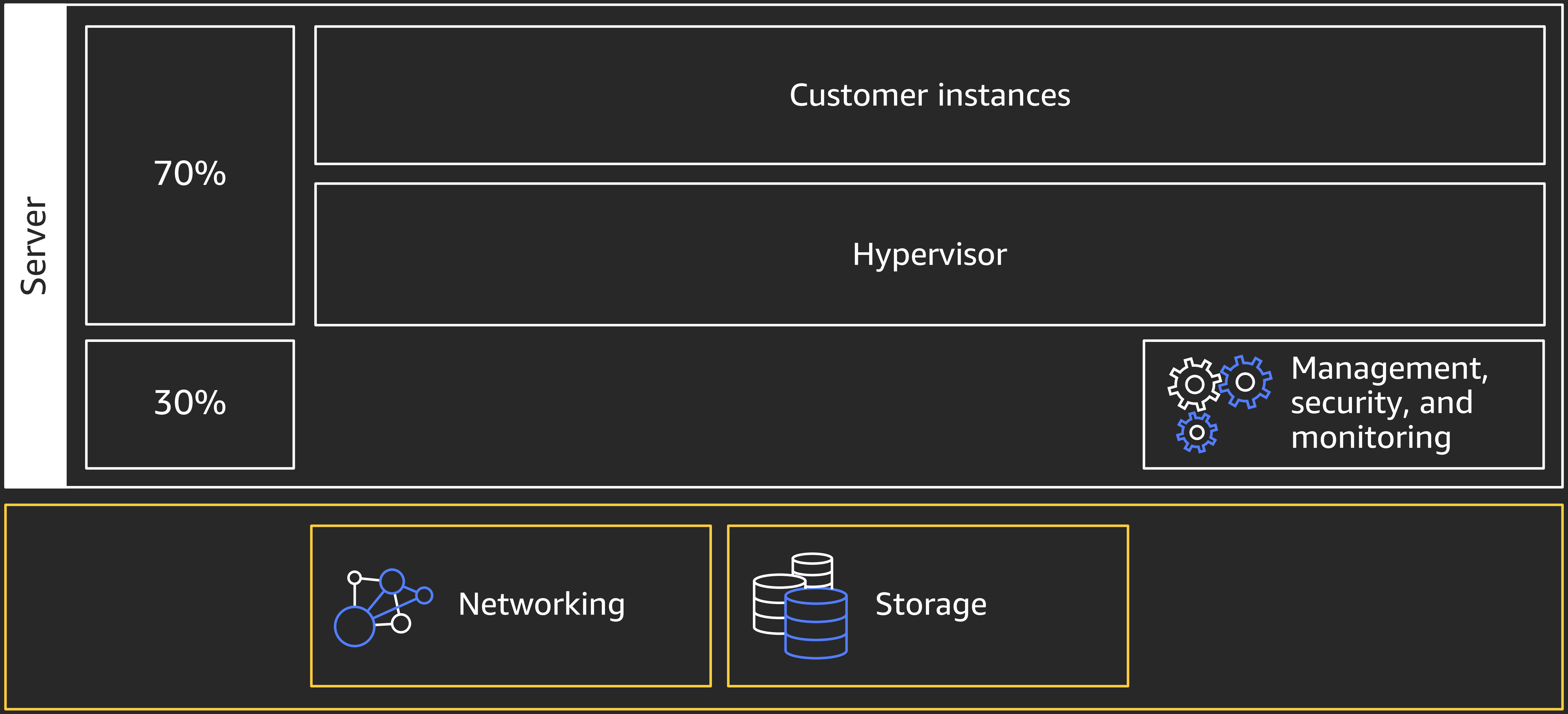


powered by

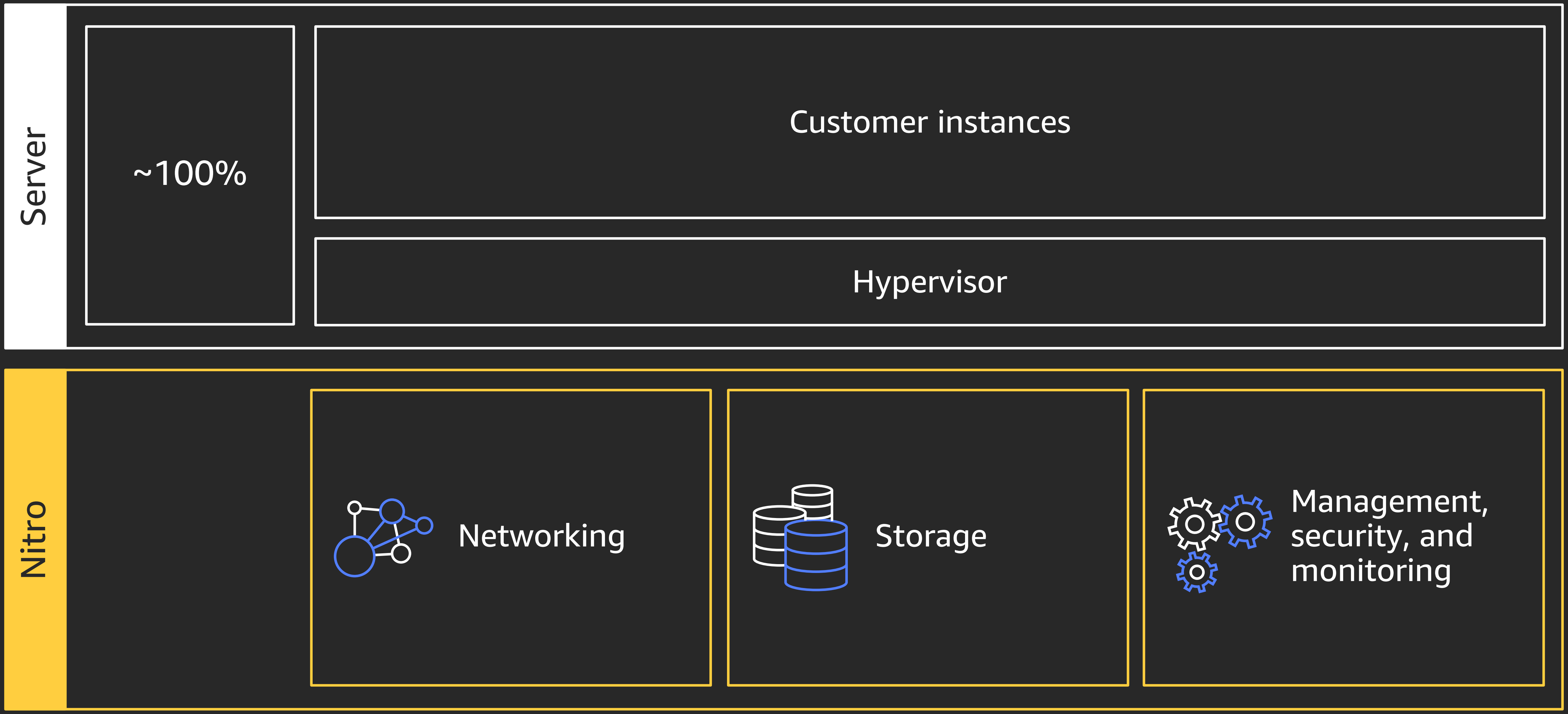
annapurnalabs

an amazon company

2013: EC2 "instance" host architecture



2017: Introducing Nitro architecture



Server

~100%

Customer instances

Hypervisor

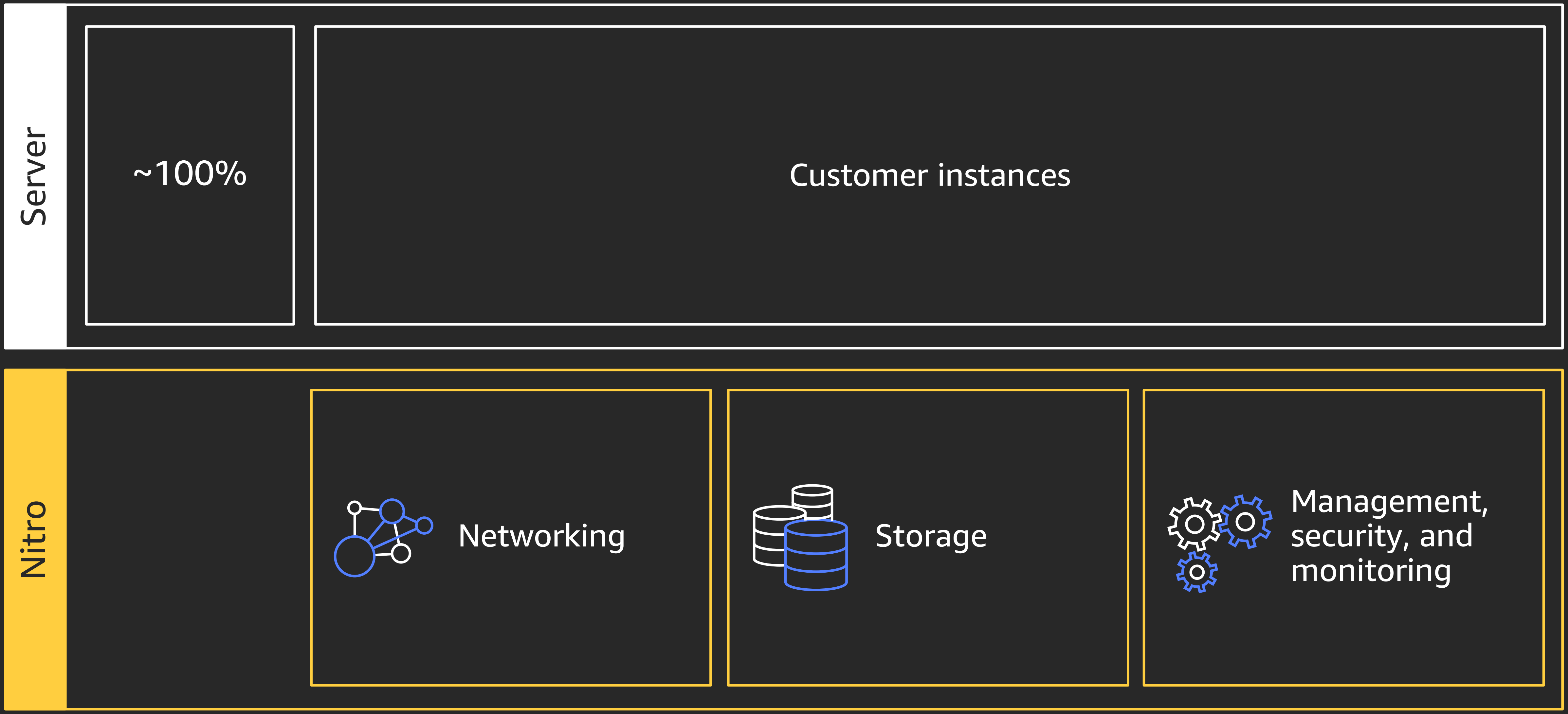
Nitro

Networking

Storage

Management, security, and monitoring

2018: Nitro enabling Bare Metal instances



Server

~100%

Customer instances

Nitro

Networking

Storage

Management, security, and monitoring

Nitro delivers

Performance

Security

Pace of innovation

Broadest and deepest 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 TB)

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

Amazon Elastic Inference

=

270 +

instance types

for virtually every workload and business need

How do you select the right instance to launch and optimize?

Announcing

Instance Discovery

New search and discovery experience to easily find EC2 instance types

Quicker and easier for you to find and compare different instance types and project costs

AWS Compute Optimizer

Machine learning based service that recommends optimal AWS resources

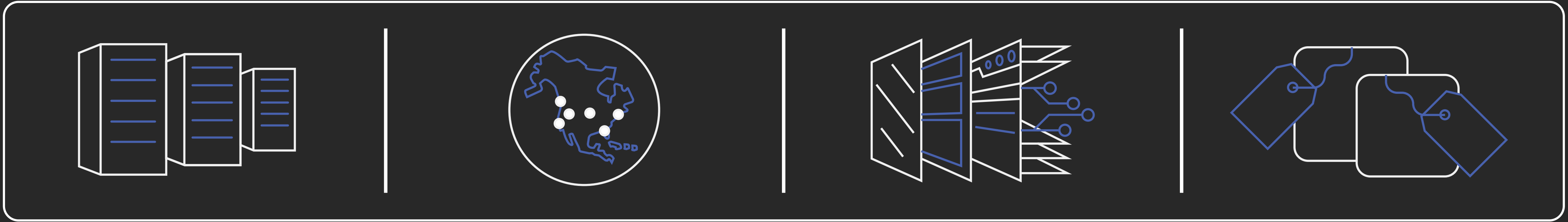
Recommends optimal EC2 instances and Amazon EC2 Auto Scaling group config

Lower costs

Optimize performance

Get started quickly

Amazon EC2 foundations



Resources

Instances

Storage

Networking

Availability

Regions and AZs

Load Balancing

Auto Scaling

Management

Deployment

Monitoring

Administration

Purchase Options

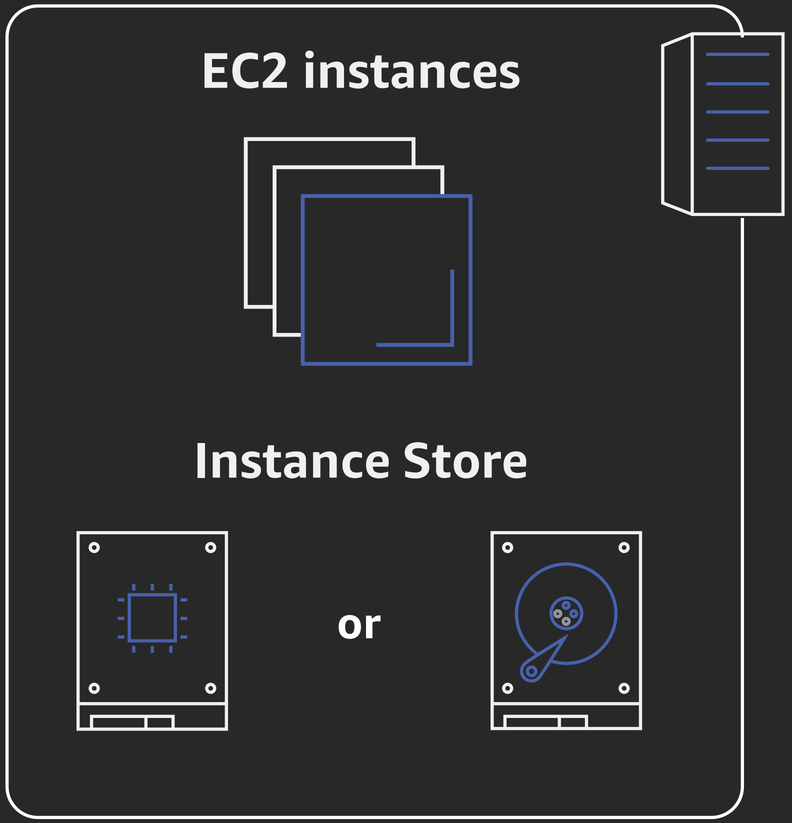
On Demand

Reserved

Spot

Savings Plan

Amazon EC2 instance store



Physical host machine

Local to instance

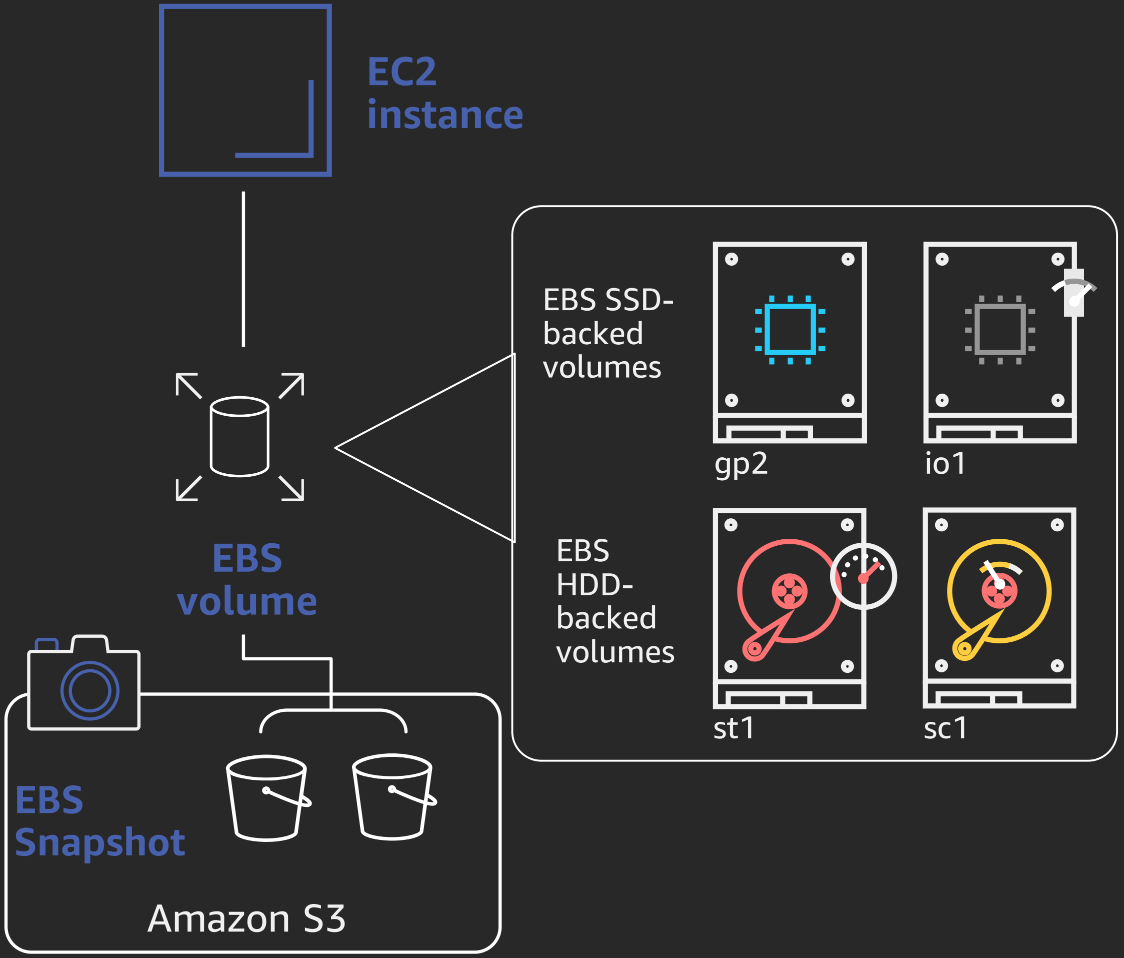
Non-persistent data store

Data not replicated (by default)

No snapshot support

SSD or HDD

Amazon EBS



Block storage as a service

Create, attach, modify through an API

Select storage and compute based on your workload

Detach and attach between instances

Choice of magnetic and SSD-based volume types

Supports snapshots: Point-in-time backup of modified volume blocks

New EBS performance and security improvements

Encryption by default for EBS volumes with opt-in setting

Encrypt all newly created EBS volumes for an account in a region

Easy to ensure compliance without change to workflows

Fast Snapshot Restore (FSR)

6x lower recovery time objective (RTO)

Skip pre-warming: Instant access to data in snapshot and full performance upon volume creation

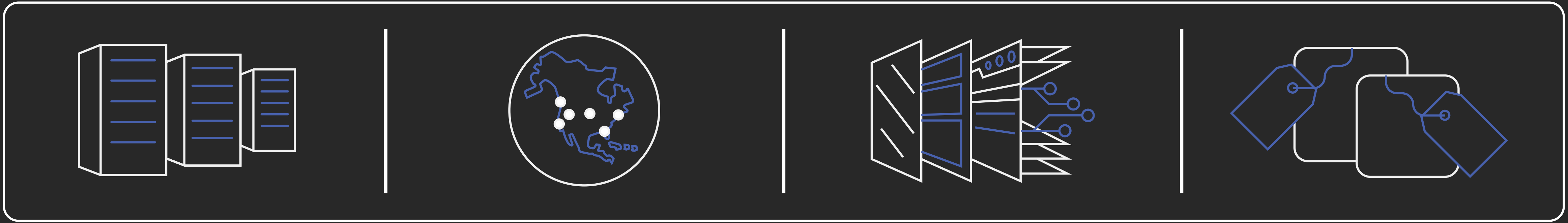
Restore up to 10 volumes simultaneously

36% higher EBS-optimized bandwidth on C5/C5d, M5/M5d, R5/R5d instance types

Dedicated bandwidth to Amazon EBS

19 Gbps maximum bandwidth, the highest across EC2 instances

Amazon EC2 foundations



Resources

Instances

Storage

Networking

Availability

Regions and AZs

Load Balancing

Auto Scaling

Management

Deployment

Monitoring

Administration

Purchase Options

On Demand

Reserved

Spot

Savings Plan

Amazon Virtual Private Cloud (Amazon VPC)

Virtual Private Cloud

Provision a logically isolated cloud where you can launch AWS resources into a virtual network

Security groups & ACLs

NAT gateway

Flow logs

VPC endpoints

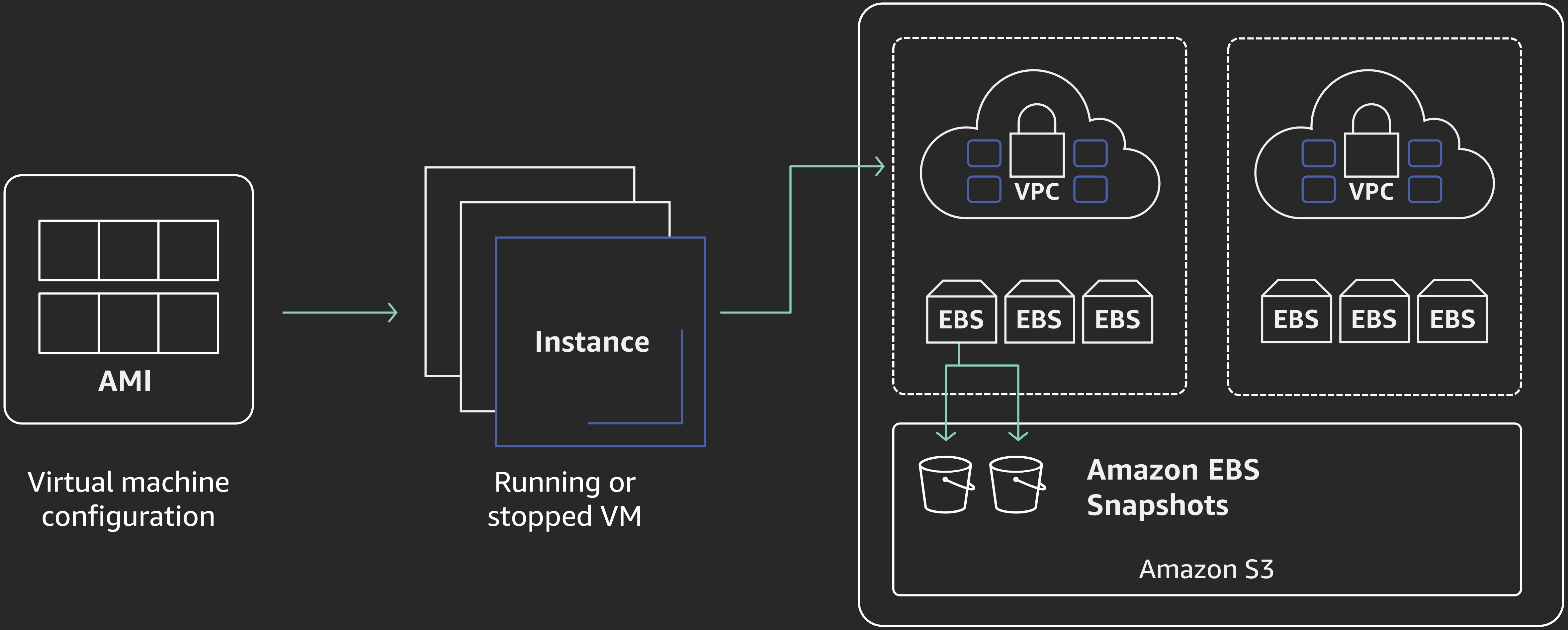
Private and secure connectivity to Amazon S3 and Amazon DynamoDB

Amazon S3

Amazon DynamoDB

Shared VPC allows multiple accounts to launch their applications into a VPC

Amazon EC2 resources recap



Virtual machine configuration

Running or stopped VM

Amazon EC2 foundations

Resources

Instances

Storage

Networking

Availability

Regions and AZs

Load Balancing

Auto Scaling

Management

Deployment

Monitoring

Administration

Purchase Options

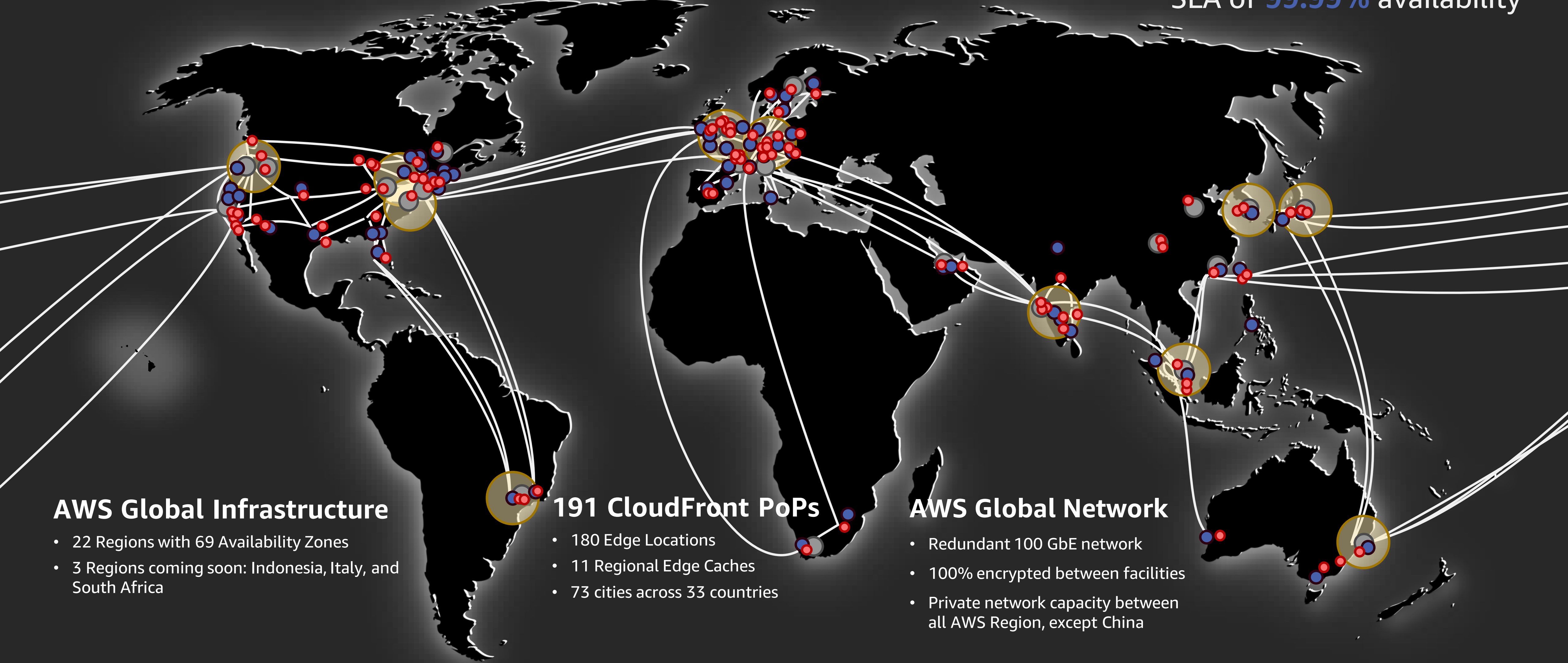
On Demand

Reserved

Spot

AWS global platform

SLA of 99.99% availability



AWS Global Infrastructure

- 22 Regions with 69 Availability Zones
- 3 Regions coming soon: Indonesia, Italy, and South Africa

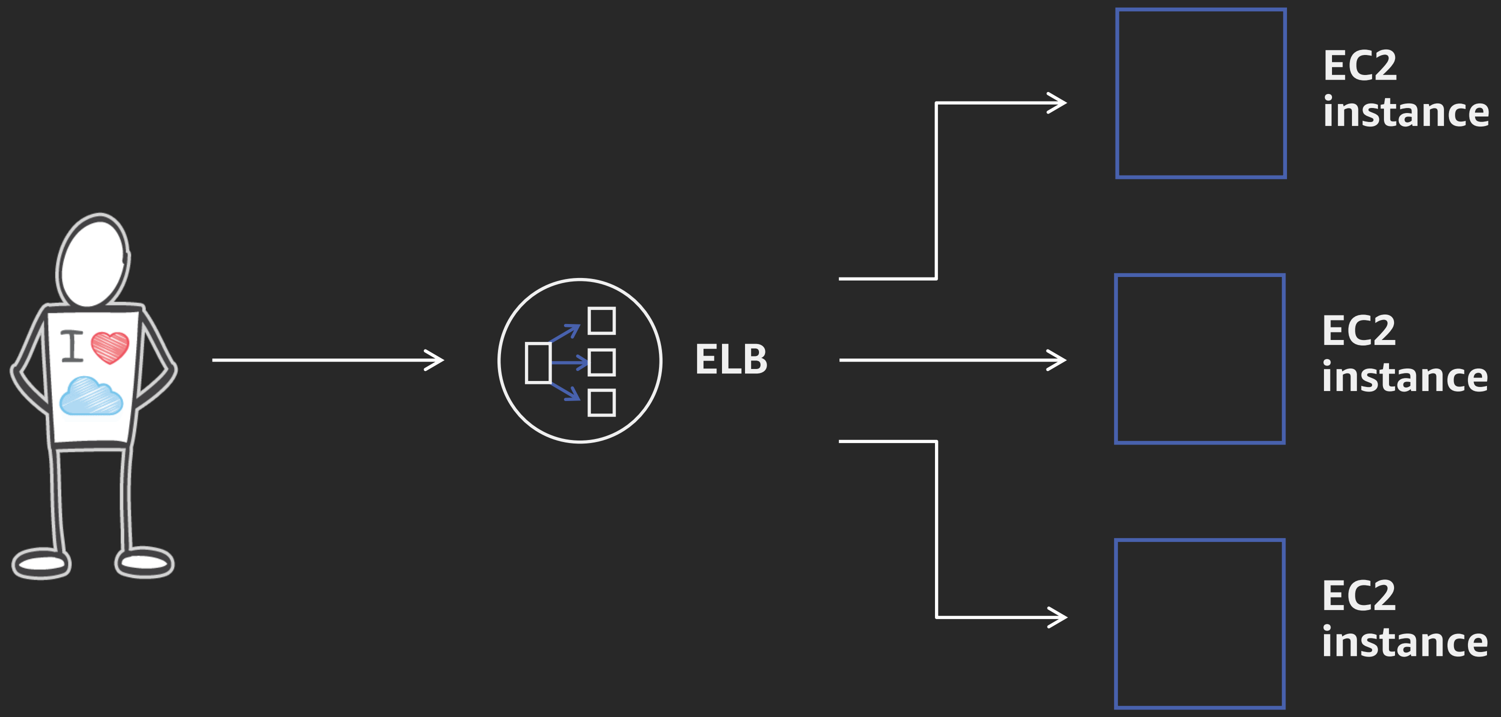
191 CloudFront PoPs

- 180 Edge Locations
- 11 Regional Edge Caches
- 73 cities across 33 countries

AWS Global Network

- Redundant 100 GbE network
- 100% encrypted between facilities
- Private network capacity between all AWS Region, except China

Elastic Load Balancing



Load balancer used to route incoming requests to multiple Amazon EC2 instances, containers, or IP addresses in your VPC

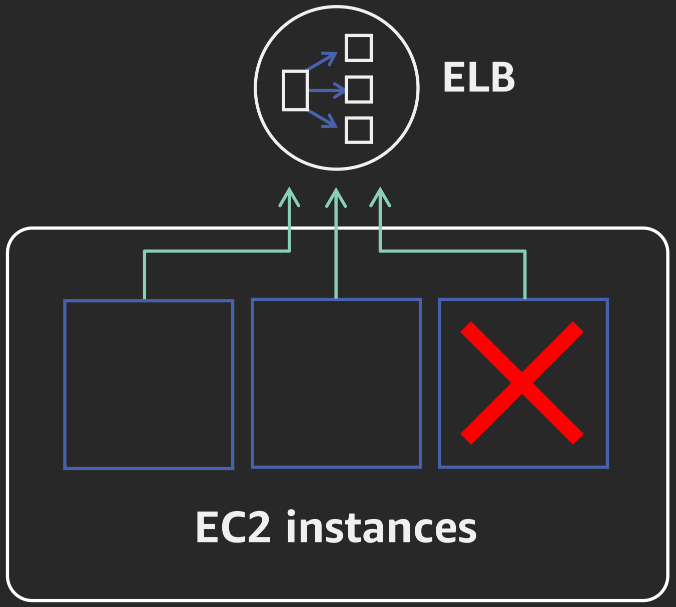
Elastic Load Balancing provides **high-availability** by utilizing multiple Availability Zones

Amazon EC2 Auto Scaling

Dynamically react to changing demand, optimize cost

Fleet management

Replace unhealthy instances



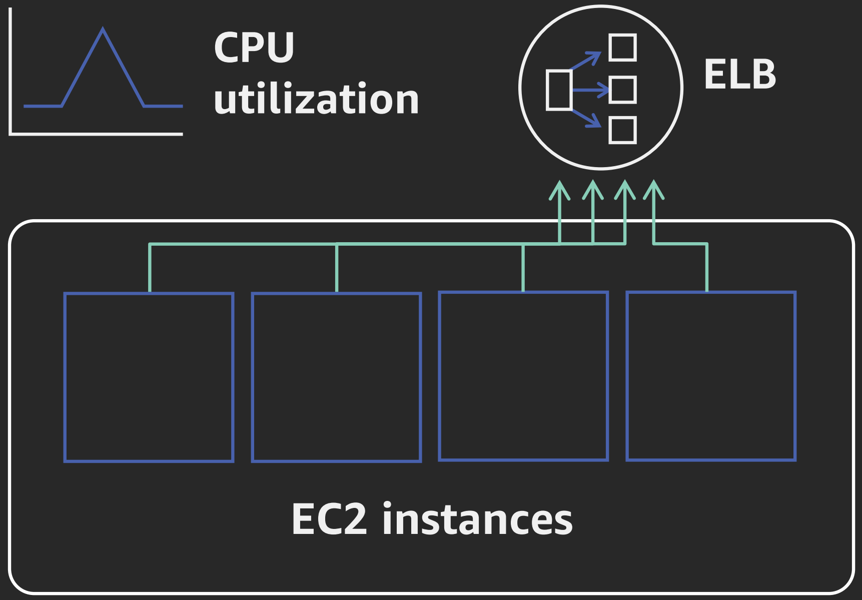
ELB

EC2 instances

Auto Scaling group

Dynamic scaling

Scale to demand



CPU utilization

ELB

EC2 instances

Auto Scaling group

Amazon EC2 foundations

Resources

Instances
Storage
Networking

Availability

Regions and AZs
Load Balancing
Auto Scaling

Management

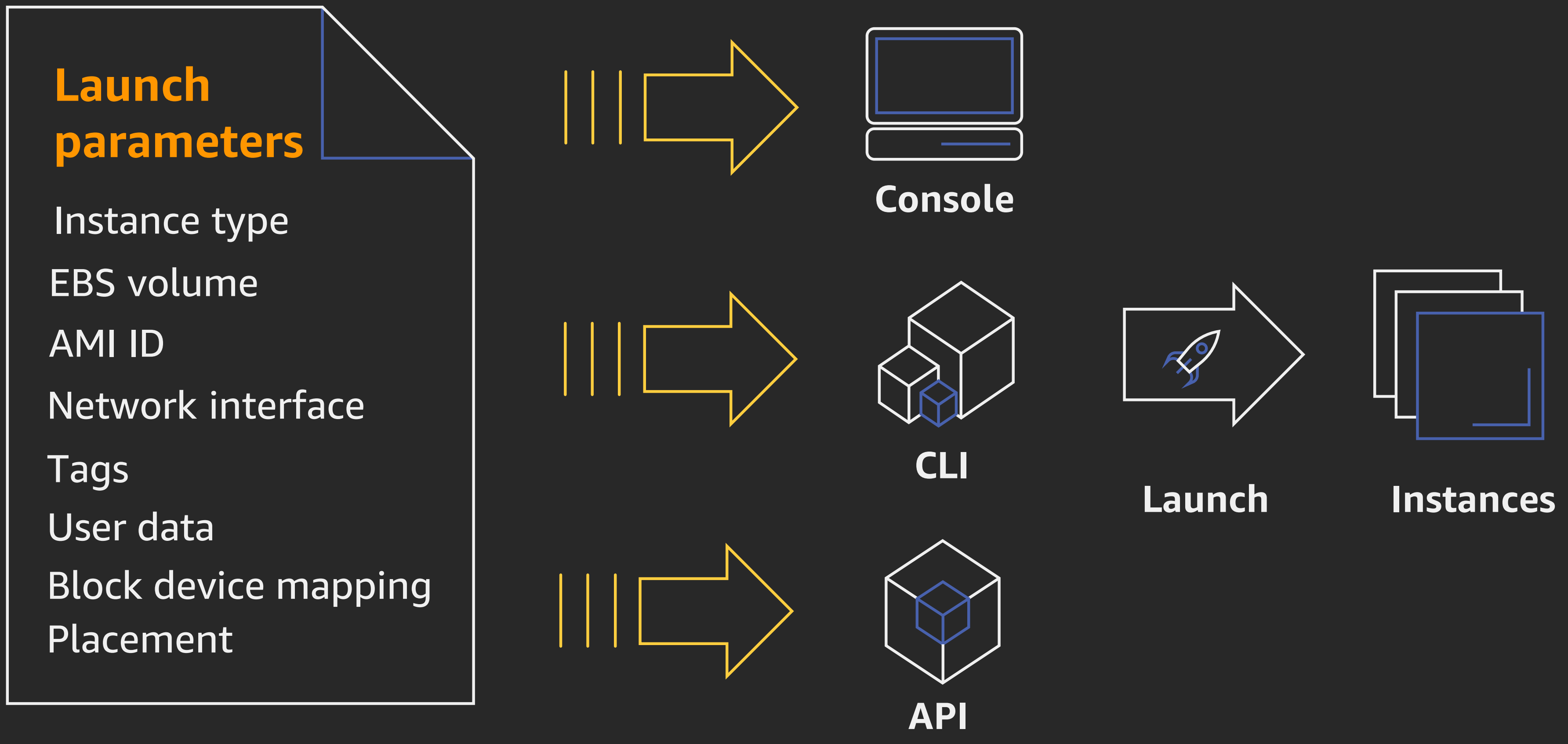
Deployment
Monitoring
Administration

Purchase Options

On Demand
Reserved
Spot
Savings Plan

Launching instances with Launch Templates

Templatize launch requests in order to streamline and simplify future launches



Launch parameters

Instance type

EBS volume

AMI ID

Network interface

Tags

User data

Block device mapping

Placement

Console

CLI

API

Launch

Instances

Consistent experience

Simple permissions

Governance and best practices

Increased productivity

AWS Systems Manager: Operate safely at scale

Cloud
and
on-premises

Linux
and
Windows

Stay patch and configuration compliant

Automate across accounts and regions

Connect to Amazon EC2 instances via browser and CLI

Track software inventory across accounts

Install agents safely across instances with rate control

AWS License Manager

Simplified license management for on-premises and cloud

More easily manage licenses from software vendors

Define licensing rules, discover usage, manage access

Gain single view of license across AWS and on-premises

Discover non-compliant software and help prevent misuse

Seamless integration with AWS Systems Manager and AWS Organizations

Free service for all customers

SAP®

Microsoft Windows

Microsoft SQL Server

Oracle

EC2 foundations

Resources

Instances

Storage

Networking

Availability

Regions and AZs

Load Balancing

Auto Scaling

Management

Deployment

Monitoring

Administration

Purchase Options

On Demand

Reserved

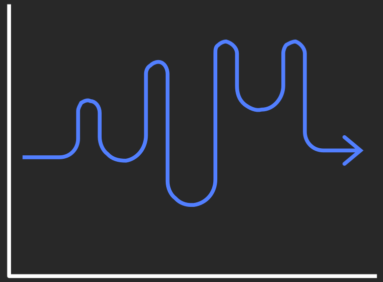
Spot

Savings Plan

Amazon EC2 purchase options

On-Demand

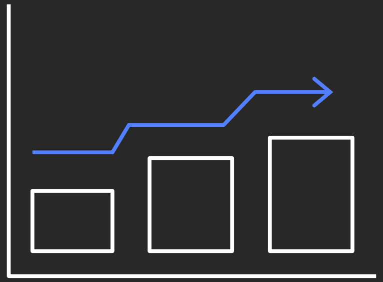
Pay for compute capacity by **the second** with no long-term commitments



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

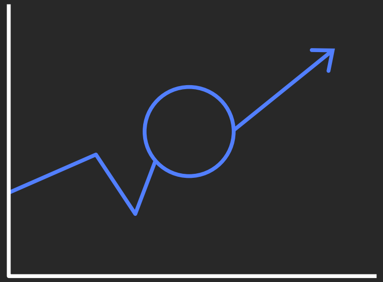
Savings Plan

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

Flexibility to access compute across EC2 and AWS Fargate

Spot Instances

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



Fault-tolerant, flexible, stateless workloads

NEW! Savings Plan

Amazon EC2 Reserved Instances pricing

Discount up to 75% off the On-Demand price

Steady state and committed usage

1- and 3-year terms

Payment flexibility with 3 upfront payment options (all, partial, none)

Convertible RIs
Change instance family, OS, tenancy, and payment

Reserve capacity or opt for flexibility across AZs and instance sizes

On-Demand capacity reservations: Manage capacity and RI decisions independently

1-Year Convertible RIs

Simplifying purchasing with Savings Plans

Flexible purchase option that offers savings of up to 72% on Amazon EC2 and AWS Fargate usage

Easy
to use

Significant
savings

Flexible

Same great prices as EC2 RIs with more flexibility

Amazon EC2 Spot pricing

Spare Amazon EC2 capacity at 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 (ex. Amazon ECS, Amazon EKS, AWS Batch, Amazon EMR) or integrated third-parties

Lean on Spot for these workloads!

Big data

CI/CD

Web services

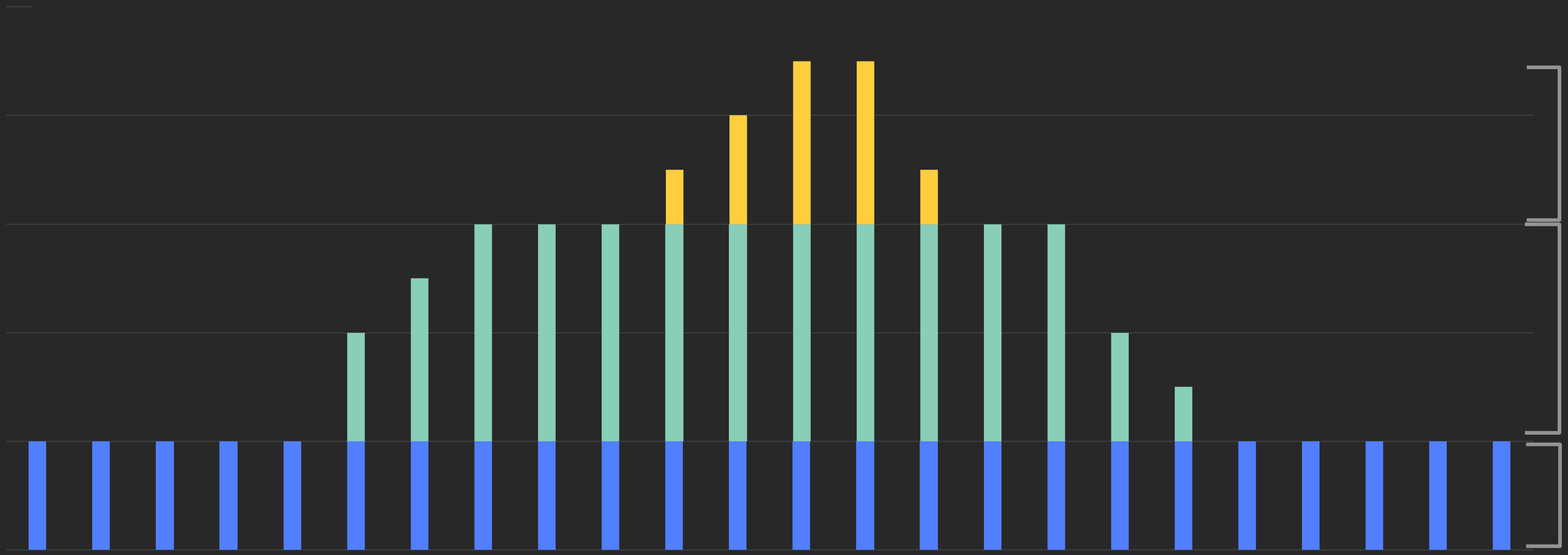
HPC

Or **containerized** workloads

Spot is ideal for:

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

To optimize Amazon EC2, combine purchase options



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

On-Demand, for new or stateful spiky workloads

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

Amazon EC2 foundations

Resources

Instances

Storage

Networking

Availability

Regions and AZs

Load Balancing

Auto Scaling

Management

Deployment

Monitoring

Administration

Purchase Options

On Demand

Reserved

Spot

Thank you!

AWS re:Invent

aws

Please complete the session survey in the mobile app.

AWS re:Invent

aws