# Introduction to Amazon Cloud & EC2 Overview

Milty Brizan (brizanm@amazon.com)

5/18/2021

# 3 - Day Agenda

- Day 1 (today) – Foundational Immersion Day
    - Virtual Servers on AWS (EC2)
    - Networking (VPC)
    - Security (IAM, etc.)
    - Storage and Databases on AWS
    - Instructor-led Labs (https://general-immersionday.workshop.aws/)
    - Survey! (https://survey.immersionday.com/00WohmqMg)
- Day 2 (tomorrow) – Application Development tools / Containerization
- Day 3 (Thursday) - Building a Document Management System (MVP)



aws

# Agenda

- Introduction to AWS Cloud
- Global Reach
- EC2 Overview
- EC2 Details



aws

# What is AWS?

AWS provides a highly reliable, scalable, low-cost infrastructure platform in the cloud that powers millions of businesses in over 190 countries around the world.

Benefits

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




aws

# What sets AWS apart?

| | |
|---|---|
| Security | Fine-grained control |
| Service Breadth & Depth; pace of innovation | 175+ services to support any cloud workload; rapid customer driven releases |
| Experience: 1M+ customers | Building and managing cloud since 2006 |
| Global Footprint | 77 Availability Zones within 24 geographic Regions, 1 Local Zone, 216 Points of Presence (205 Edge Locations and 11 Regional Edge Caches) in 84 cities across 42 countries. |
| Machine Learning | More machine learning happens on AWS than anywhere else. Machine learning in the hands of every developer and data scientist |
| Ecosystem | Tens of thousands of APN partners. The AWS Marketplace offers 50 categories, and more than 8,000 software listings |
| Enterprise leader | AWS positioned as a Leader in the Gartner Magic Quadrant for Cloud Infrastructure as a Service, Worldwide |



aws

# Experience with Operational Reliability

Our goal is to make our operational performance indistinguishable from perfect. We are driven to remove any all causes of failure.

- We have spent over a decade building the world's most reliable, secure, scalable, and cost-effective infrastructure.
- Service SLAs between 99.9% and 100% availability. Amazon S3 is designed for 99.999999999% durability.
- Availability Zones exist on isolated fault lines, flood plains, and electrical grids to substantially reduce the chance of simultaneous failure.
- The AWS Service Health Dashboard provides 24/7 visibility in the real-time operational status of all services around the globe.



aws

# Pricing Philosophy

## High volume / low margin businesses are in our core DNA

| Trade CapEX for variable expense | Our economies of scale provide us with lower costs | Pricing model choice to support variable and stable workloads | Save more money as you grow bigger |
| --- | --- | --- | --- |
| Pay for what you use | 85 price reductions since 2006 | On-demand<br>Reserved Instances<br>Spot | Tiered pricing<br>Volume discounts<br>Custom pricing |



aws

# Customer obsessed

90%

of roadmap originates with customer requests and are designed to meet specific needs

heroku

"Performance, reliability, and responsiveness are fundamental to our customer experience, and T3 instances help us to deliver on that customer promise while also controlling our costs."

—Heroku



aws

Figure 1. Magic Quadrant for Cloud Infrastructure and Platform Services

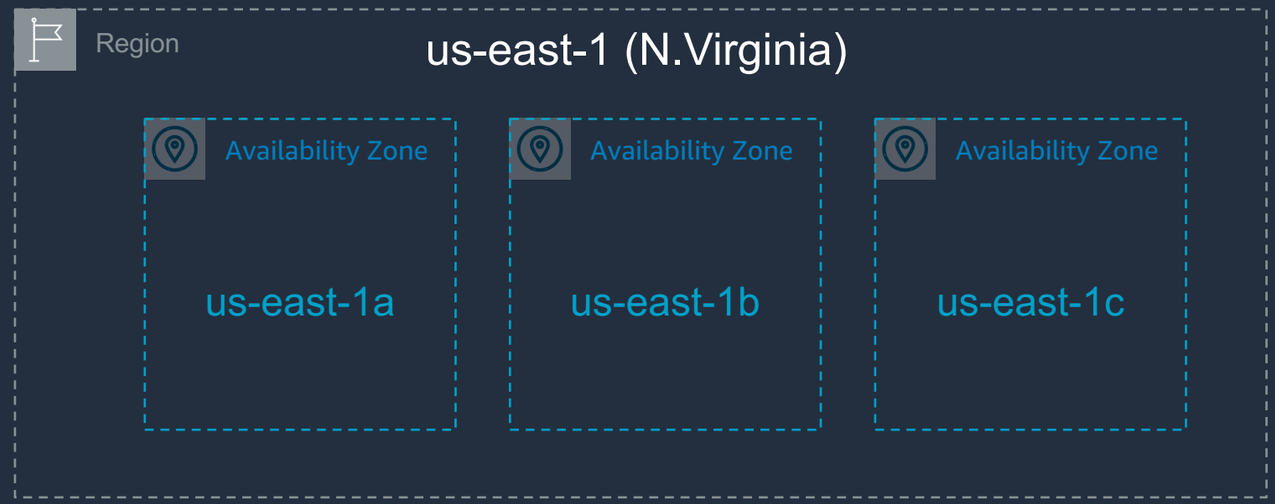

# AWS Recognized as a Cloud Leader for the 10th Consecutive Year

*Gartner, Magic Quadrant for Cloud Infrastructure & Platform Services, Raj Bala, Bob Gill, Dennis Smith, David Wright, Kevin Ji, 1 September 2020. 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.*


# AWS Global Reach

# Availability Zones

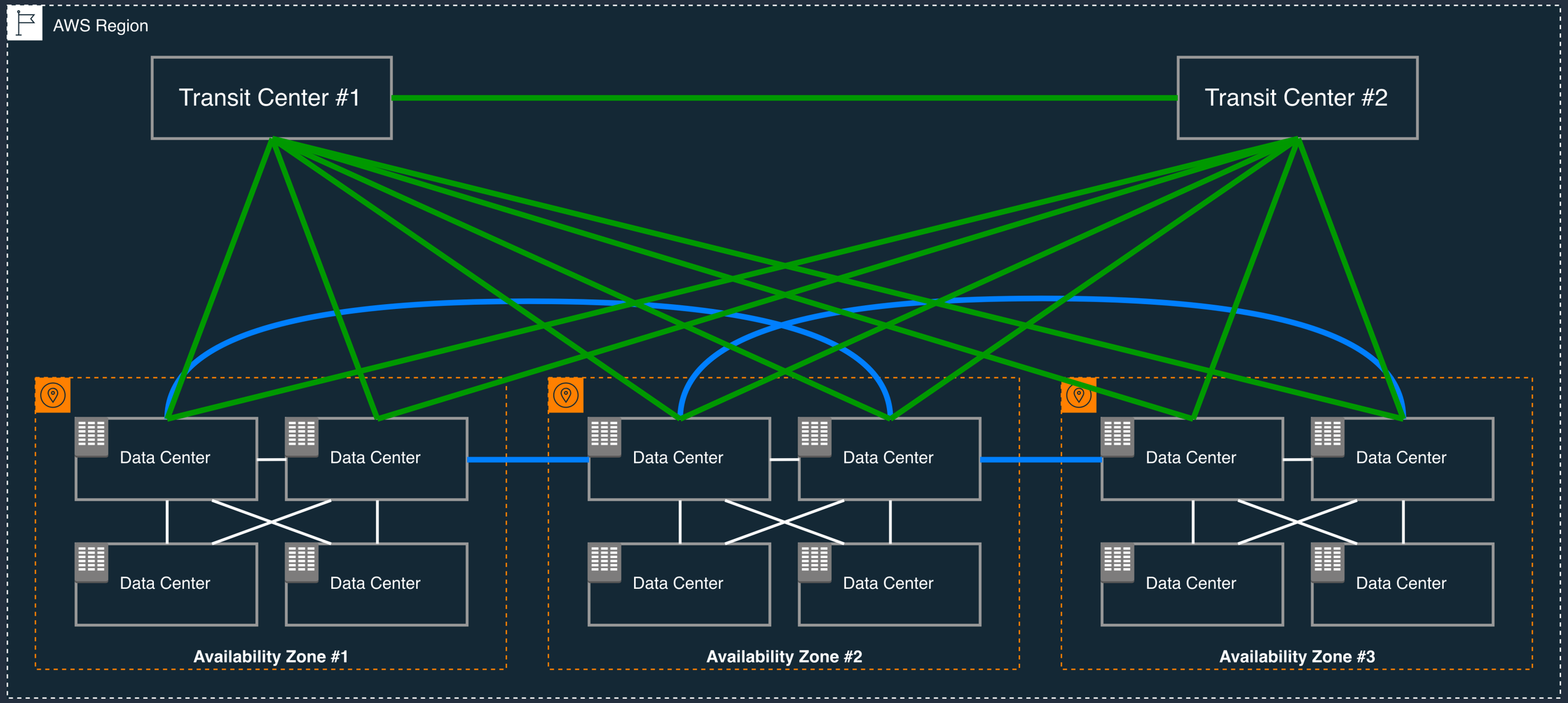
- A region is comprised of multiple Availability Zones (typically 3)
- An Availability Zone (AZ) is one or more discrete data centers with redundant power, networking, and connectivity in an AWS Region
- High throughput, low latency (<10mS) network  between Availability Zones
- All traffic between AZ’s is encrypted
- Physical Separation < 100km





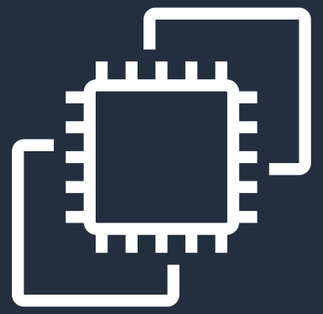

aws

# Availability Zones




# EC2 Overview

# Choices for Compute



## Amazon EC2

Virtual server instances
in the cloud

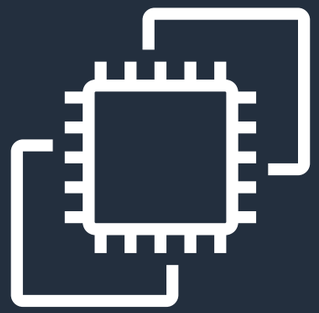
## Amazon ECS, EKS, and Fargate

Container management service
for running
Docker on a managed
cluster of EC2

## AWS Lambda

Serverless compute
for stateless code execution in
response to triggers

# Amazon EC2



Amazon EC2

Linux | Windows

Arm and x86 architectures

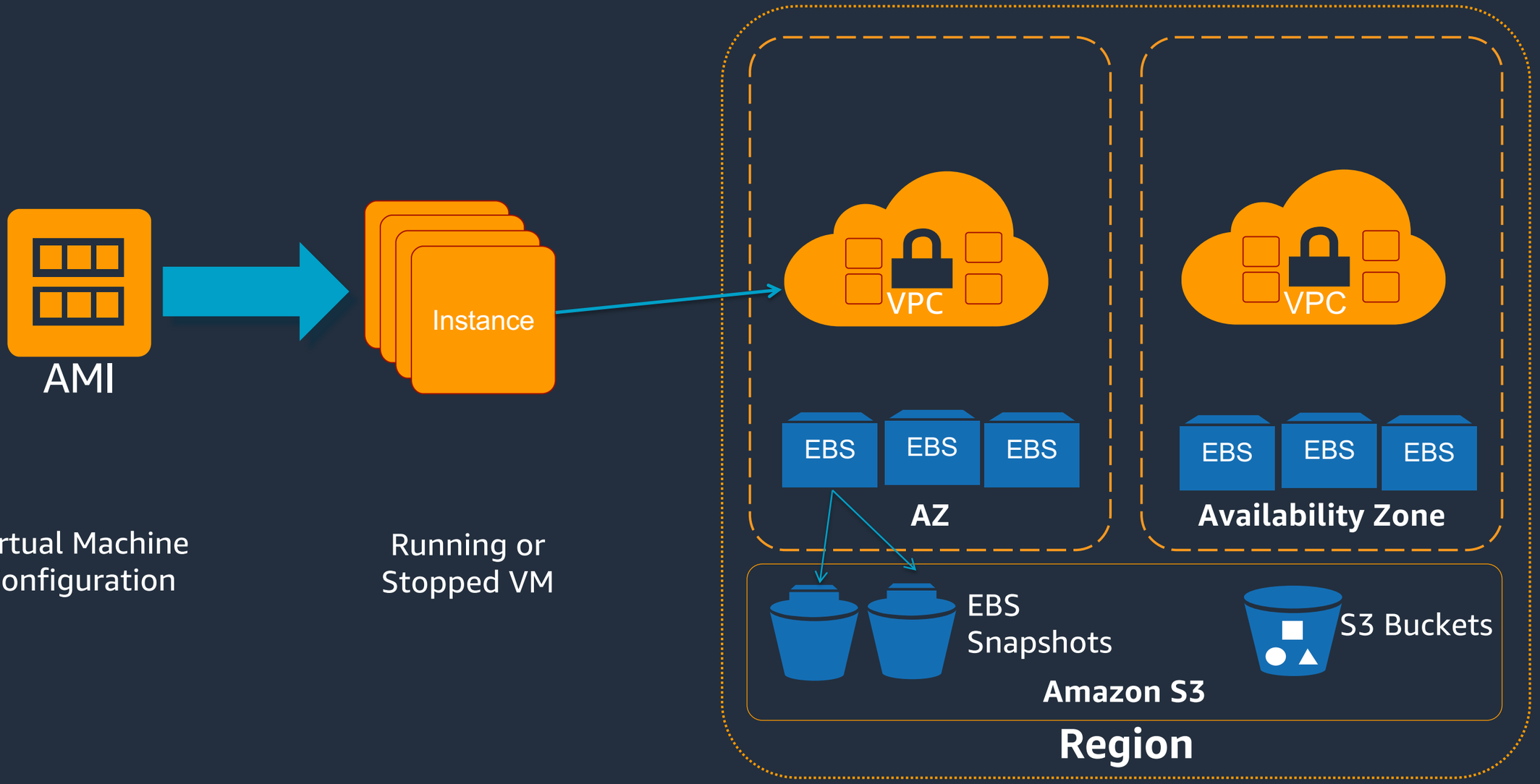
General purpose and workload optimized

Bare metal, disk, networking capabilities

Packaged | Custom | Community AMIs

Multiple purchase options: On-demand, RI, Spot

# EC2 Terminology




Virtual Machine Configuration

Running or Stopped VM

# 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/virtualcores/

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



aws

# Memory and Storage

**What's a GiB?**

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

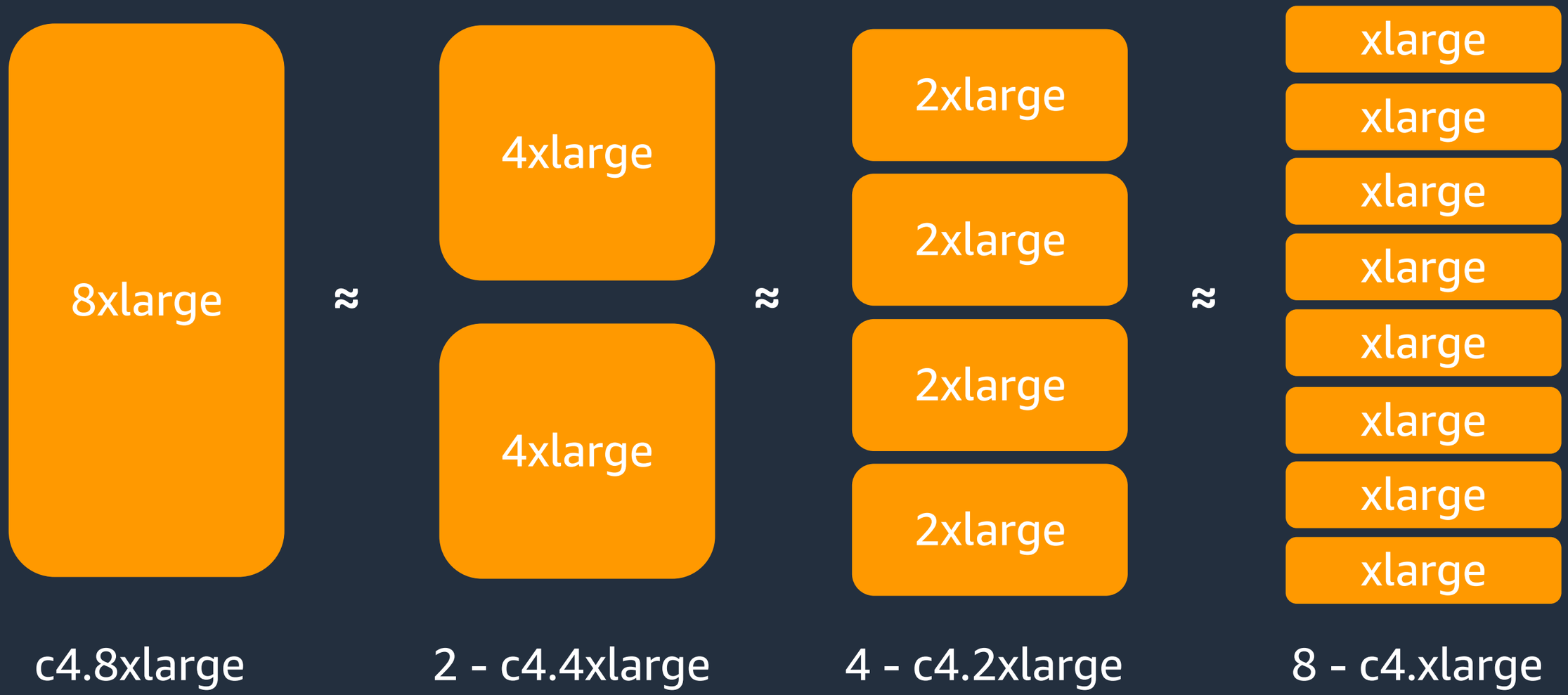
**What about storage?**

- Storage is independent of compute
- You allocate drives known as EBS volumes
- Max 16 TiB per volume
- Some instance types provide physically attached (ephemeral) storage



aws

# Instance sizing




c4.8xlarge | 2 - c4.4xlarge | 4 - c4.2xlarge | 8 - c4.xlarge



aws

# Resource allocation

- All resources assigned to you are dedicated to your instance with no over commitment*
  - All vCPUs are dedicated to you
  - Memory allocated is assigned only to your instance
  - Network resources are partitioned to avoid “noisy neighbors”
- Curious about the number of instances per host?
  - See “Dedicated Hosts Configuration Table” for a guide.

*Again, the “T” family is special



aws

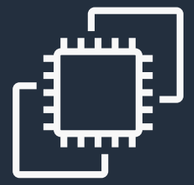
# Choose your processor and architecture


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



aws

# EC2 Naming Explained

Instance generation

c5n.xlarge

Instance family

Attribute

Instance size



aws

# Instance Types

| | General Purpose | | Compute Optimized | | Memory Optimized | | | | Accelerated Computing | | | Storage Optimized | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Burstable performance | General Purpose | Compute Intensive | Compute +memory 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 | X1e | | G3 | P3 | F1 | | D2 | H1 |
| Local storage (NVMe SSD) | | M5d | C5d | | R5d | | | Z1d | | | | I3 | | |
| AMD | T3a | M5a | | | R5a | | | | | | | | | |
| metal | | M5m | c5m | | R5m | | u-12tb1 | Z1dm | | | | I3m | | |
| others | A1 | M6g | C6g | | R6g | | | | | P3dn | | I3en | | |

arm



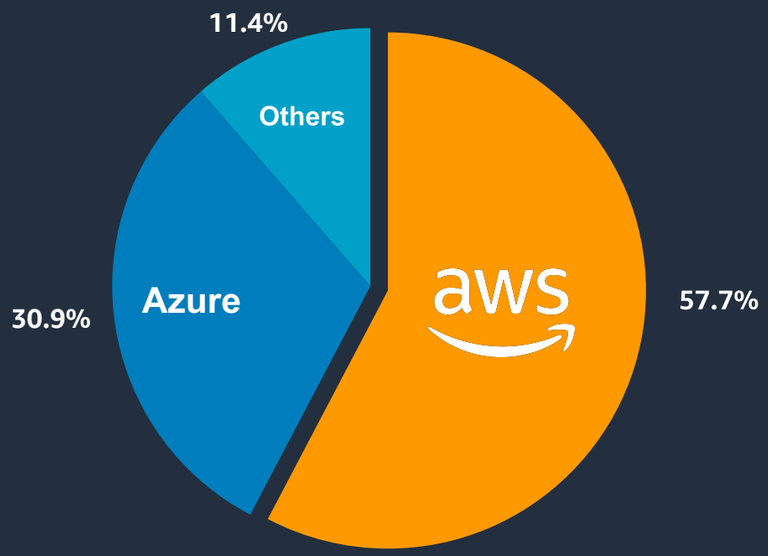
aws

# EC2 Operating Systems Supported

- Windows 2003R2*/2008*/2008R2*/2012/2012R2/2016/2019
- Amazon Linux
- Debian
- Suse
- CentOS
- Red Hat Enterprise Linux
- Ubuntu

for more OSes see: https://aws.amazon.com/marketplace/b/2649367011

# Windows Licenses by Cloud Provider




Note: Includes Windows instances deployed in the public cloud IaaS market during 2017 Source: IDC estimates, 2018

IDC, Windows Server Operating Environment Market Update, Doc # US44217118, Aug 2018

https://d1.awsstatic.com/analyst-reports/IDC_Slide_WindowsonAWS_JM181015.pdf

# What is an Amazon Machine Image (AMI)?


Provides the information required to launch an instance

Launch multiple instances from a single AMI

An AMI includes the following

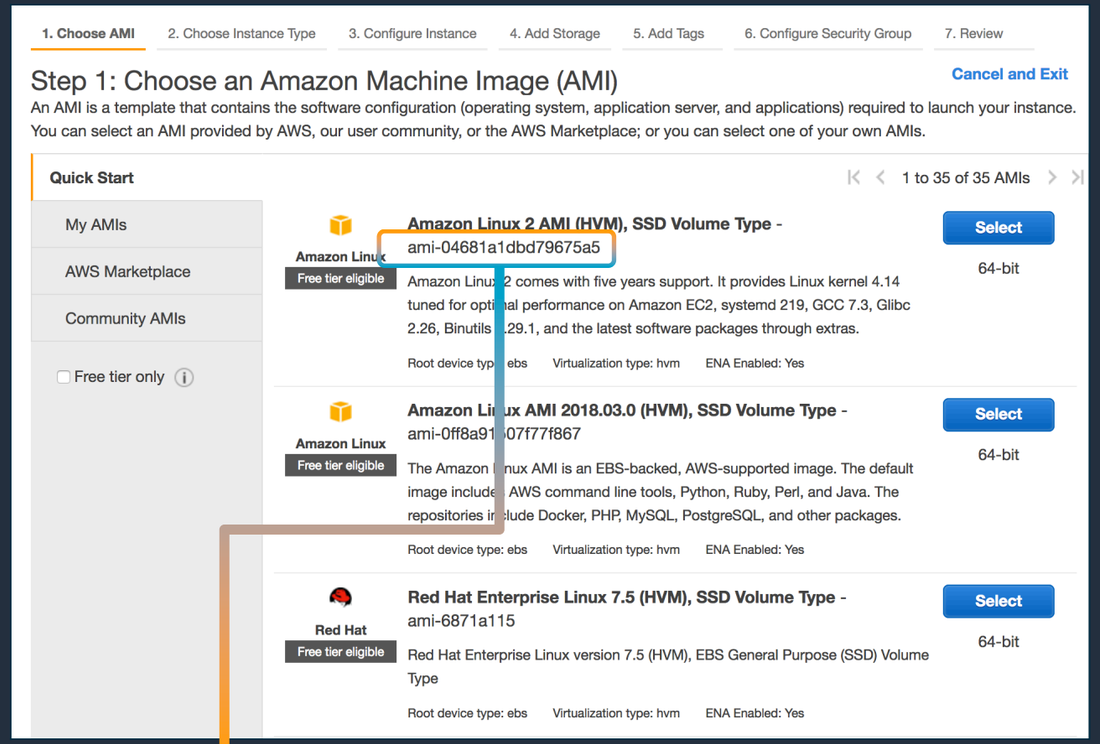
- A template for the root volume (for example, operating system, applications)
- Launch permissions that control which AWS accounts can use the AMI
- Block device mapping that specifies volumes to attach to the instance



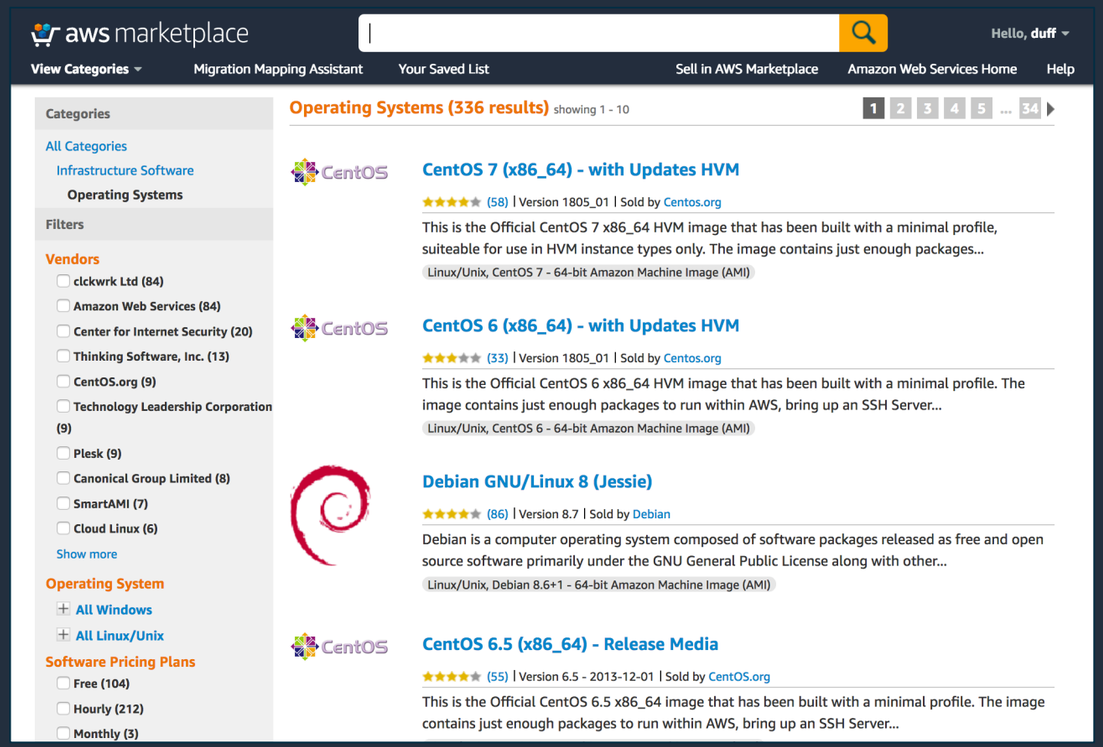
aws

# Choosing an AMI

## AWS Console



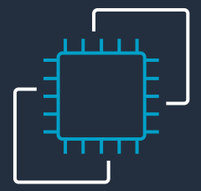

## AWS Marketplace




Use the AMI ID to launch through the API or AWS Command Line Interface (AWS CLI)

```
aws ec2 run-instances --image-id ami-04681a1dbd79675a5 --instance-type c4.8xlarge --count 10 --key-name MyKey
```

# Choice of accelerators for specialized workloads



## Elastic Graphics

Easily add graphics acceleration to your EC2 instance

Configure right amount of graphics acceleration for your workload

Accelerate application for fraction of cost of standalone graphics instances

## Elastic Inference

Reduce deep learning inference costs by up to 75%

Easily attach fractional sizes of a full GPU instance to EC2 or SageMaker instances

Scale inference acceleration up or down as needed with EC2 Auto Scaling

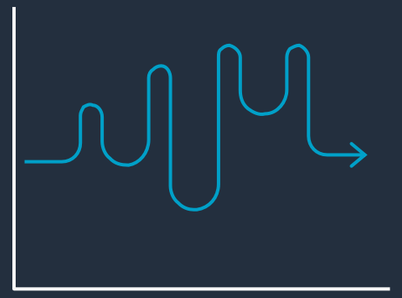


aws

# Amazon EC2 purchase options

## On-Demand

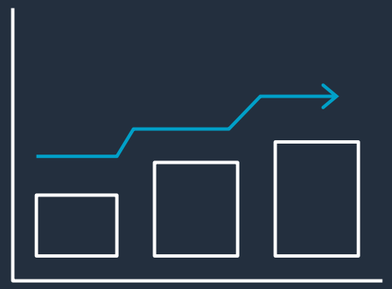
Pay for compute capacity by **the second** with no long-term commitments



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 Amazon EC2 RIs with **more flexibility**

Committed flexible access to compute

## Spot Instances

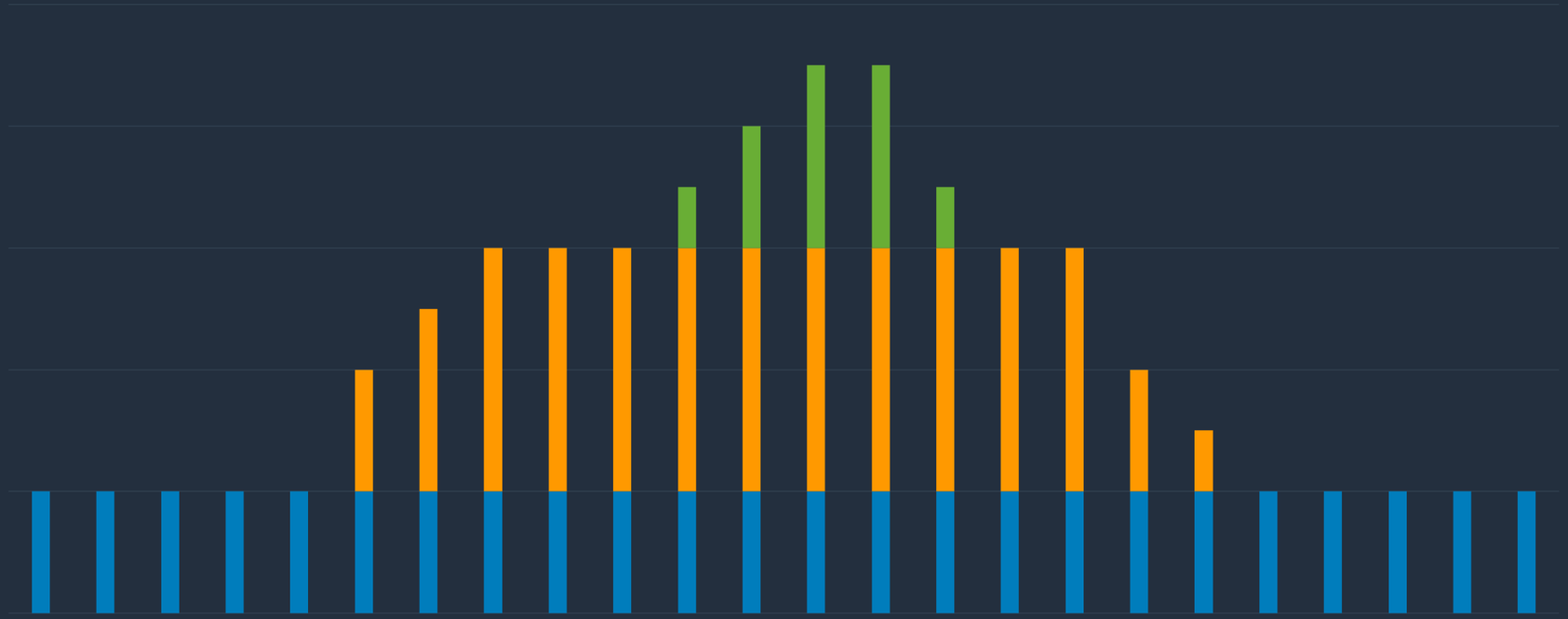
Spare Amazon EC2 capacity at **savings of up to 90%** off On-Demand prices

Fault-tolerant, flexible, stateless workloads



aws

# Simplify capacity and cost optimization



Now with hibernate

Scale using
**Spot,**
**On-Demand,**
or both

Use **Reserved Instances**
for known/steady-state
workloads

## AWS services make this easy and efficient

- Amazon EC2 Auto Scaling
- EC2 Fleet
- Amazon Elastic Container Service
- Amazon Elastic Container Service for Kubernetes
- AWS Thinkbox
- Amazon EMR
- AWS CloudFormation
- AWS Batch

# Hibernate Amazon EC2 Instances

## Maintain a fleet of pre-warmed instances to quickly get to a productive state

Available with Amazon EBS-backed instances

Use familiar Stop and Start APIs

Memory data saved in EBS root volume

RAM contents are encrypted on EBS

Its just like closing and opening your laptop!

Applications can pick up right where it left off



aws

# 270+ instances across 42 instance Families




aws

# Broadest and deepest platform choice

| Categories | Capabilities | Options |
| --- | --- | --- |
| General purpose | Choice of processor (AWS, Intel, AMD) | Elastic Block Store |
| Burstable | Fast processors (up to 4.0 GHz) | Elastic Inference |
| Compute intensive | High memory footprint (up to 12 TiB) | Elastic Graphics |
| Memory intensive | Instance storage (HDD and NVMe) | |
| Storage (High I/O) | Accelerated computing (GPUs and FPGA) | |
| Dense storage | Networking (up to 100 Gbps) | |
| GPU compute | Bare Metal | |
| Graphics intensive | Size (Nano to 32xlarge) | |

**270 +**

**instance types**

for virtually every workload and business need



aws

# Broadest choice of processors


Second generation of Intel® Xeon processor


AMD Rome


Graviton



aws

# AWS Graviton2 Processor

## Enabling the best price/performance for your cloud workloads

### Graviton Processor

First Arm-based processor available in major cloud

Built on 64-bit Arm Neoverse cores with AWS-designed silicon using 16 nm manufacturing technology

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

### Graviton2 Processor

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

Built with 64-bit Arm Neoverse cores with AWS-designed silicon using 7 nm manufacturing technology

Up to 64 vCPUs, 25 Gbps enhanced networking, 18 Gbps EBS bandwidth



aws

# AWS Graviton2 based instances

Up to 40% better price-performance for general purpose, compute intensive, and memory intensive workloads.

## M6g

Built for: General-purpose workloads such as application servers, mid-size data stores, and microservices.

## C6g

Built for: Compute intensive applications such as HPC, video encoding, gaming, and simulation workloads.

## R6g

Built for: Memory intensive workloads such as open-source databases, or in-memory caches.

---

Launched **in 2020**

Local NVMe-based SSD storage options also available in general purpose (M6gd), compute-optimized (C6gd), and memory-optimized (R6gd) instances



aws


# EC2 Design

# Which hypervisor do we use?

**Original host architecture: Xen-based**

- Hypervisor consumed resources from the underlying host
- Limited optimization

**AWS Nitro Hypervisor: Custom KVM based hypervisor**

- AWS Nitro System (launched on Nov 2017)
- Less server resources used, more resources for the customer
- AWS optimized

Bare metal: **Direct access to processor and memory resources**

- Built on the AWS Nitro system
- Enables custom hypervisors and micro-VM runtimes



aws

# AWS Nitro System

## Nitro Card

Local NVMe storage

Elastic Block Storage

Networking, monitoring, and security

## Nitro Security Chip

Integrated into motherboard

Protects hardware resources

## Nitro Hypervisor

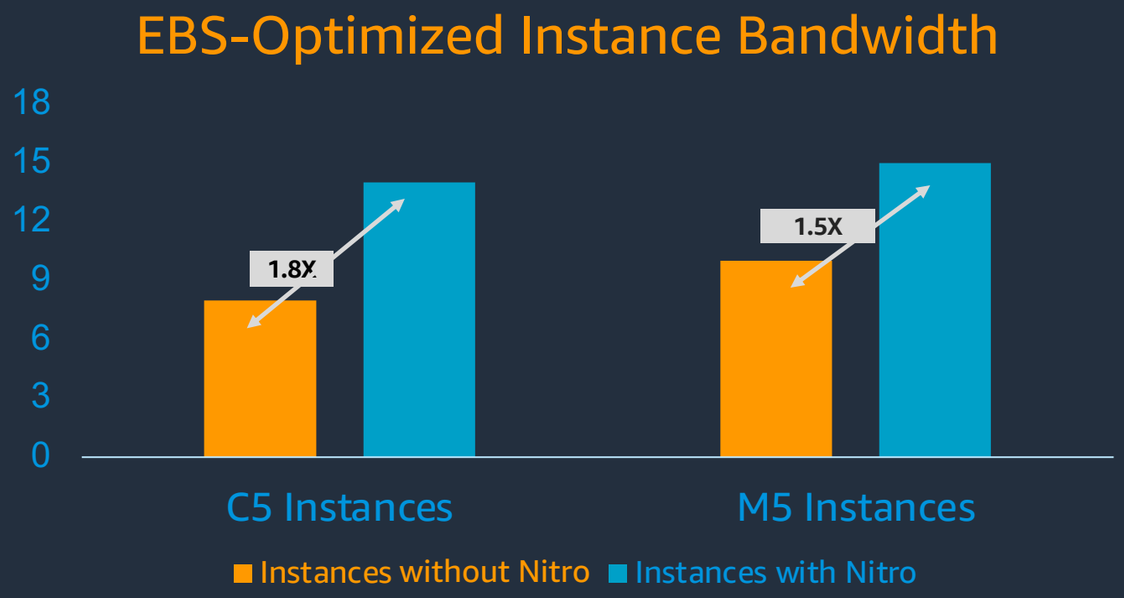
Lightweight hypervisor

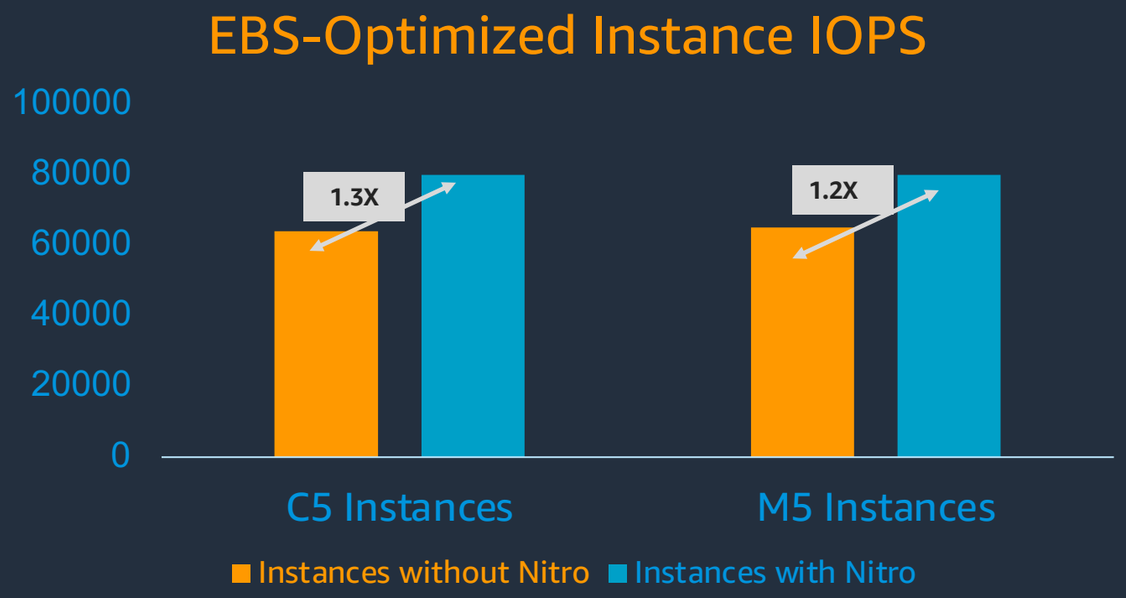
Memory and CPU allocation

Bare metal-like performance

**Modular building blocks for rapid design and delivery of EC2 instances**

# AWS Nitro System




Nitro instances provide **bandwidth, performance, and price improvements** over previous instance generations

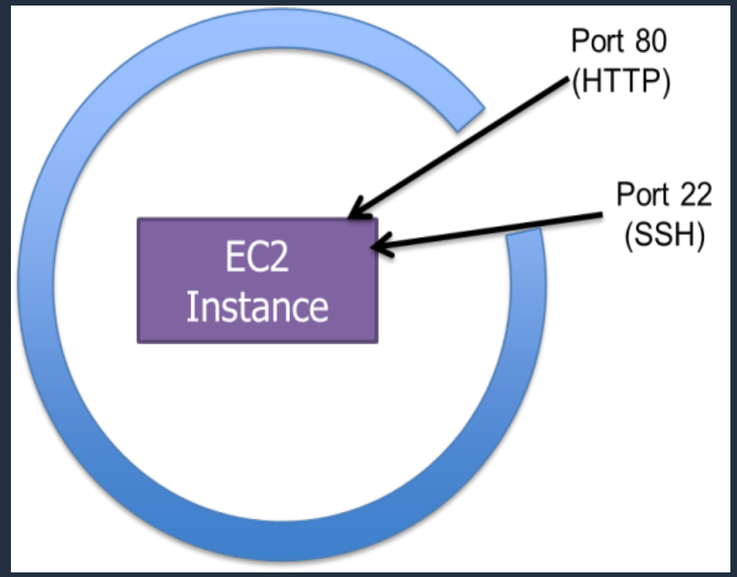


aws

# EC2 Security Groups

## Security Group Rules

- Name
- Description
- Protocol
- Port range
- IP address, IP range, Security Group name



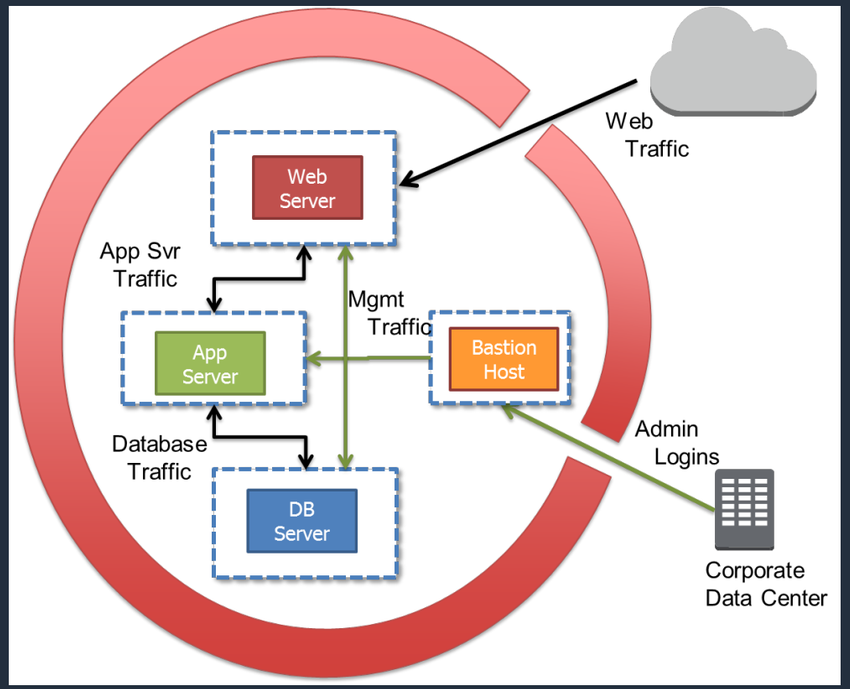



aws

# Tiered EC2 Security Groups

## Hierarchical Security Group Rules

- Dynamically created rules
- Based on Security Group membership
- Create tiered network architectures

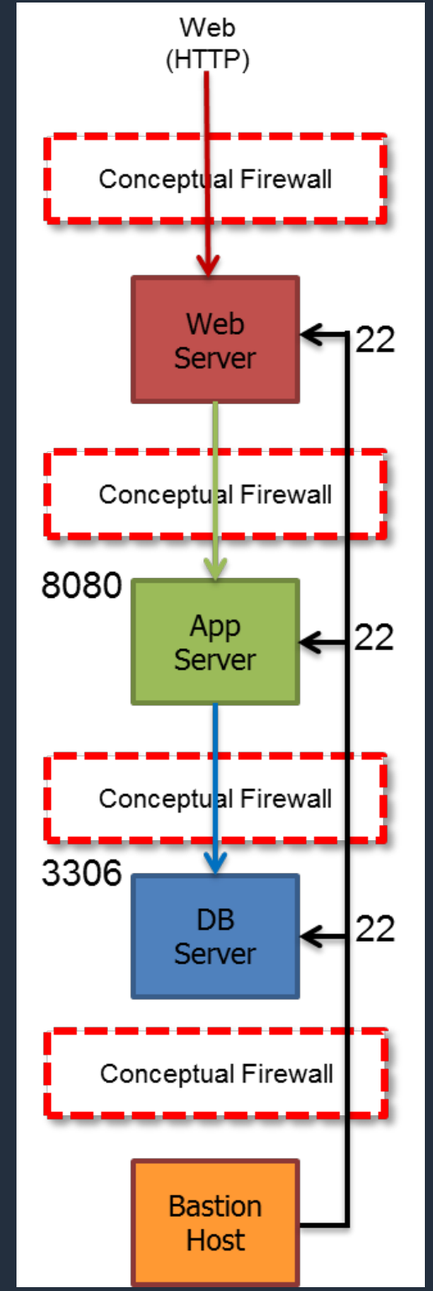



"Web" Security Group:
TCP 80 0.0.0.0/0
TCP 22 "Mgmt"

"App" Security Group:
TCP 8080 "Web"
TCP 22 "Mgmt"

"DB" Security Group:
TCP 3306 "App"
TCP 22 "Mgmt"

"Mgmt" Security Group:
TCP 22 163.128.25.32/32






aws

# EC2 IP Addressing

| Default VPC | Virtual Private Cloud |
| --- | --- |
| Dynamic Private IP | Dynamic or Static Private IP Address |
| Dynamic Public IP | None by default (can be created with publicIP=true) |
| Optional Static Public IP (EIP) | Optional Static Public IP (EIP), BYOIP |
| AWS-provided DNS names<br>• Private DNS name<br>• Public DNS name | AWS-provided public DNS lookup<br>AWS-provided private DNS names<br>Customer-controlled DNS options |



aws

# EC2-Specific Credentials

## EC2 key pairs

- Linux – SSH key pair for first-time host login
- Windows – Retrieve Administrator password

## Standard SSH RSA key pair

- Public/Private Keys
- Private keys are not stored by AWS

## AWS approach for providing **initial** access to a generic OS

- Secure
- Personalized
- Non-generic (NIST, PCI DSS)

"Public Half" inserted by Amazon into each EC2 instance that you launch

"Private Half" downloaded to your desktop

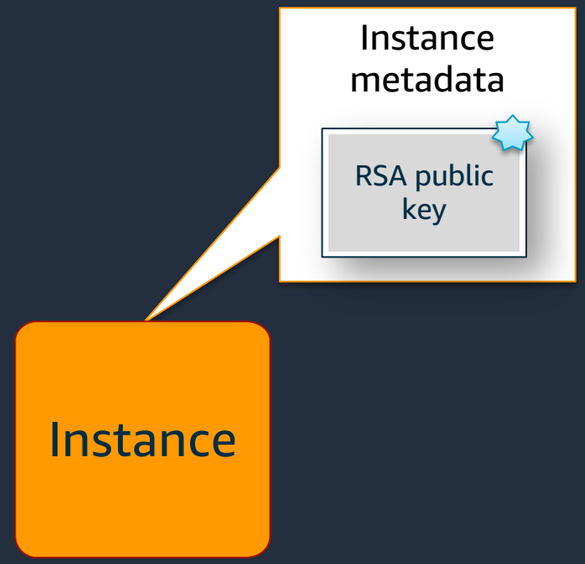


aws

# EC2 Instance access and Key Pairs

Linux launch (first boot)

- **Public key** made available through metadata
- Public key inserted into ~/.ssh/authorized_keys
- User connects with SSH using their **private key**






aws

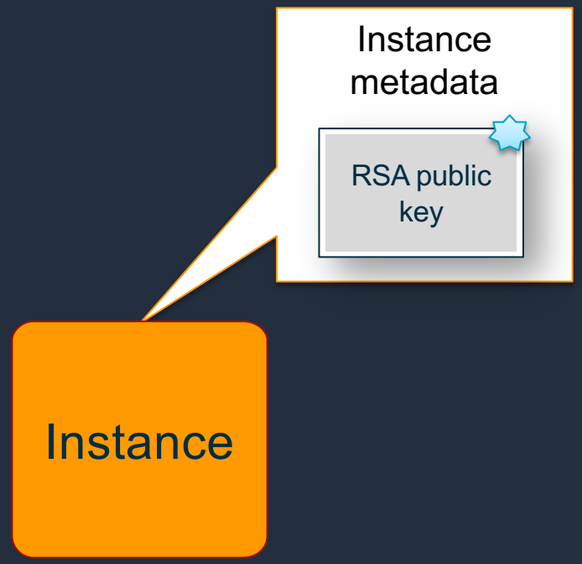
# EC2 Instance access and Key Pairs

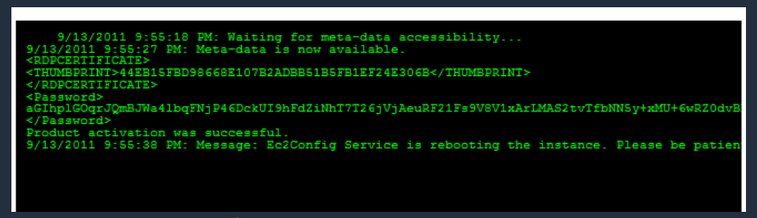
## Linux launch (first boot)

- **Public key** made available through metadata
- Public key inserted into ~/.ssh/authorized_keys
- User connects with SSH using their **private key**

## Windows launch (first boot sequence)

- **Public key** made available through metadata
- Sysprep
- Random Administrator password
- Password encrypted with public key
- User decrypts password with their **private key**




**System log**
<Password>
aGIhplGOqrJQmBJW
…
K9gTD31Q==
</Password>

# Instance Metadata

http://169.254.169.254/latest/meta-data/ contains a wealth of info

- ami-id
- ami-launch-index
- ami-manifest-path
- block-device-mapping/
- hostname
- instance-action
- ★ **instance-id**
- instance-type
- kernel-id
- local-hostname
- local-ipv4
- mac
- network/
- ★ **placement/availability-zone**
- profile
- public-hostname
- public-ipv4
- public-keys/



aws

# Any Questions?



aws