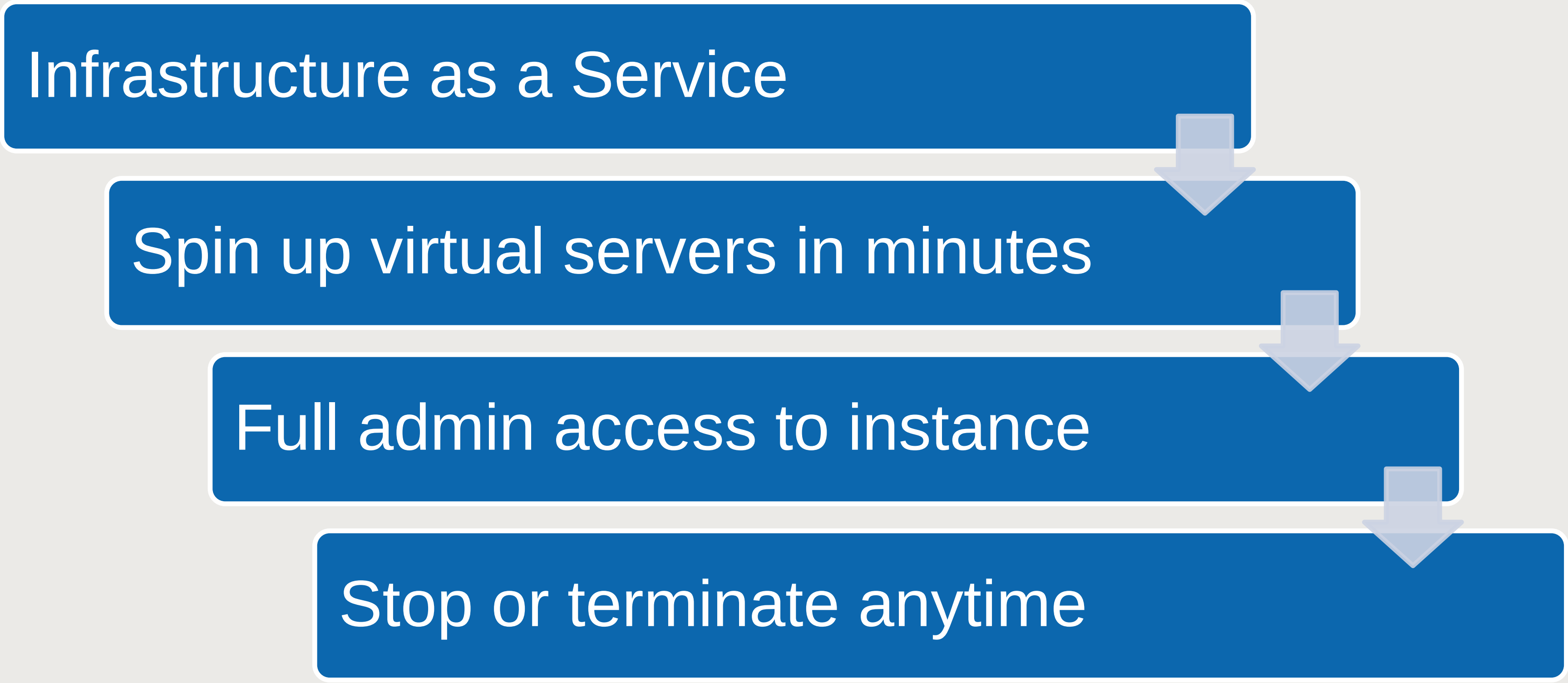
Elastic Compute Cloud (EC2)



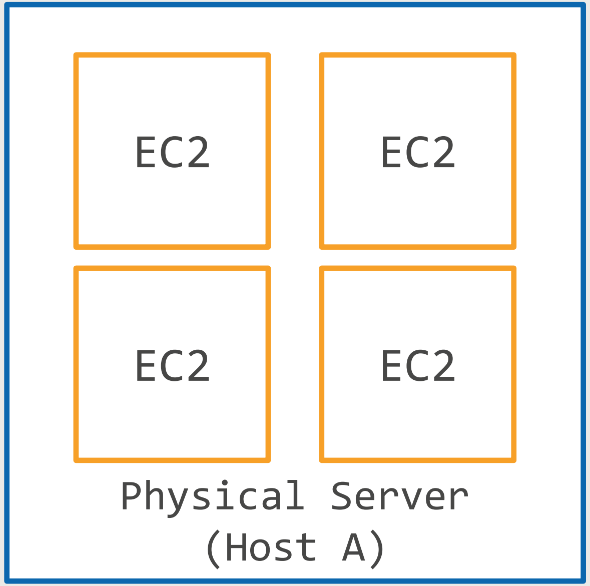
Infrastructure as a Service

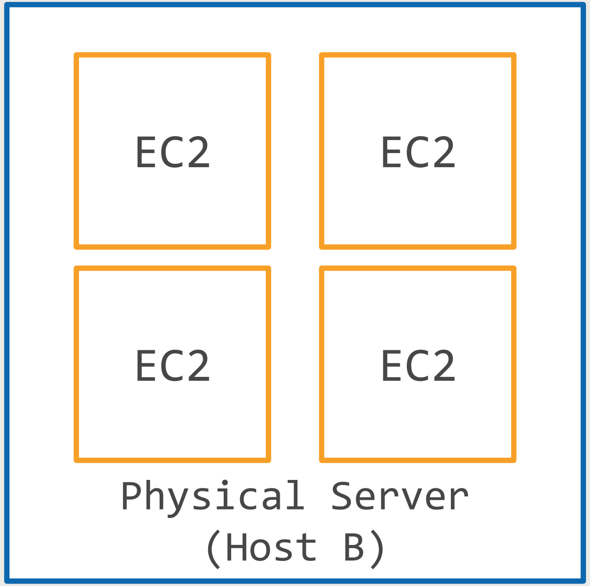
Spin up virtual servers in minutes

Full admin access to instance

Stop or terminate anytime

Multi-tenant





EC2 instances could belong to different customers

Shared Infrastructure

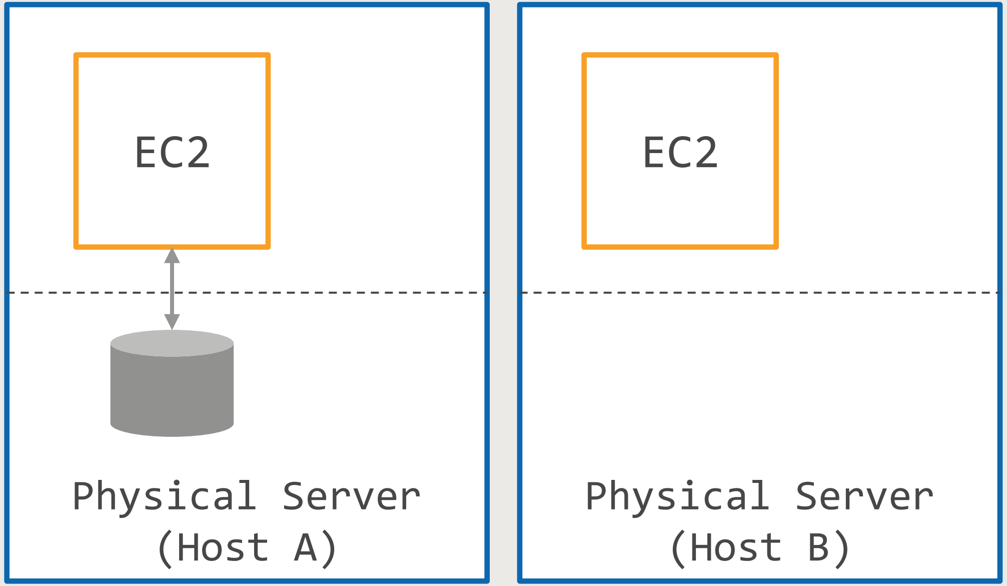
Physical Server Resources are freed up when not needed

Stop-Start would change public IP

Storage for EC2 instance OS

- Instance Storage
- Elastic Block Store (EBS)

Instance Storage



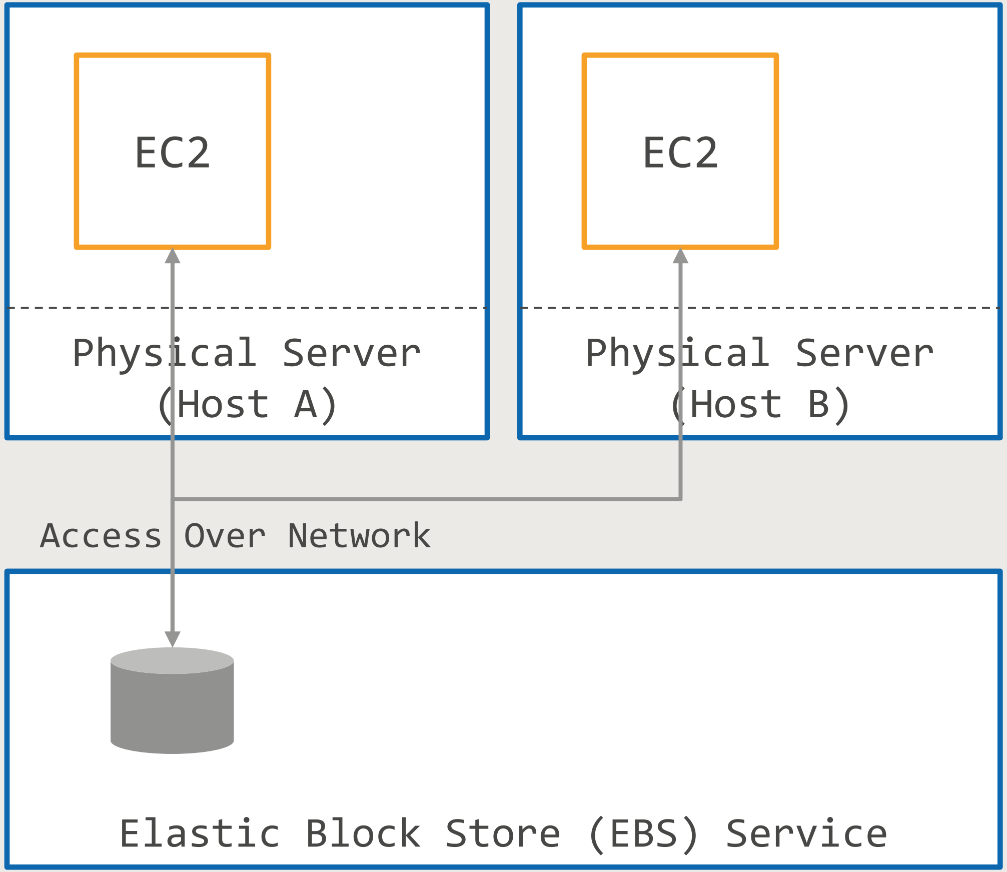
Stop or Terminate – you lose data

Ephemeral

Maintain a backup

Suitable for software that maintain redundancy like Hadoop File System

Elastic Block Store (EBS)



Stop and Start your instance

Persistent Storage

Suitable for long term retention like Databases

EC2 Storage

Choice of configurations with

- Instance Store
- Elastic Block Store

Physical Server - Resource Sharing

Allocated based on your EC2 instance configuration

- CPU
- Memory
- Instance Storage

Physical Server – Common Resource

NETWORK
I/O

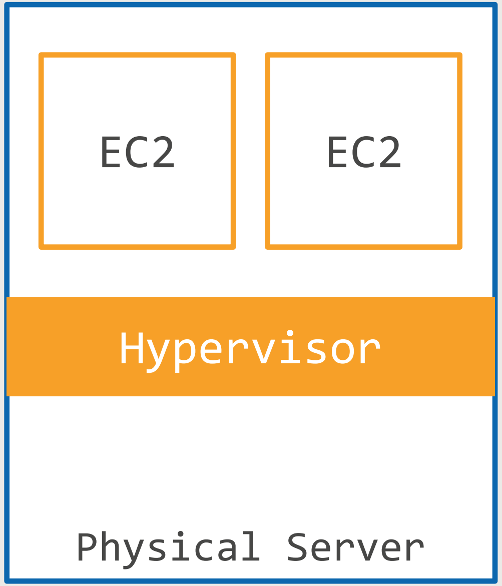
DISK
I/O

Shared by all instances

When underutilized, an instance can consume a larger portion

When in-demand, each instance is allowed to meet baseline performance

Virtualization



VMware

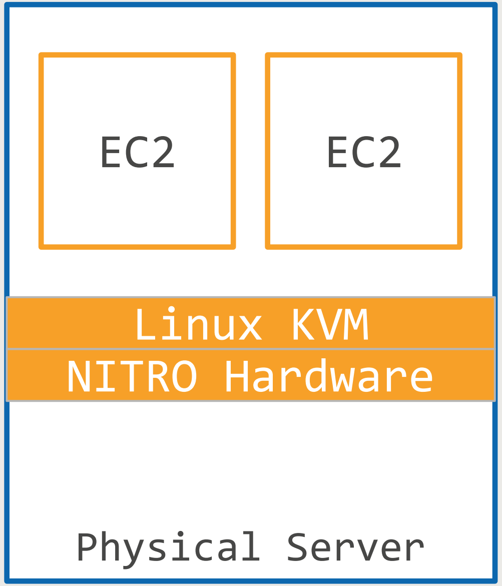
Xen

Linux KVM

Hyper-V

AWS uses NITRO

AWS NITRO Virtualization

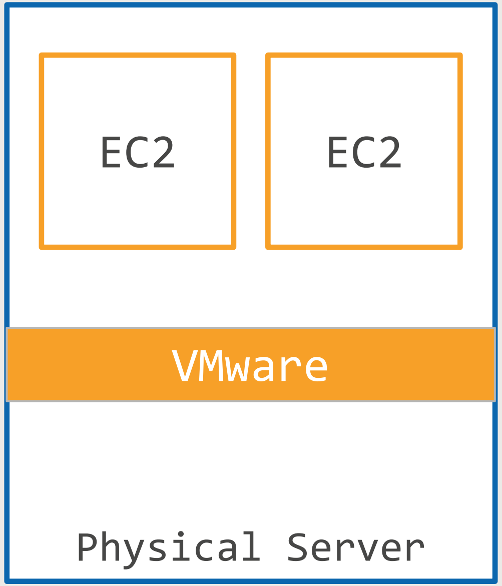


Custom Hardware Assisted Virtualization

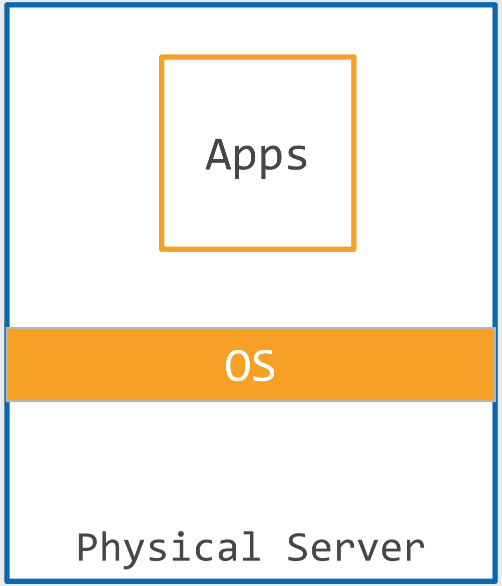
Consistent High-Performance Infrastructure

Uses light-weight Linux KVM

EC2 Bare Metal Instances



Use different virtualization environment like VMware



Run directly without hypervisor

Amazon Machine Image (AMI)

Quick Start AMIs

Popular Distributions

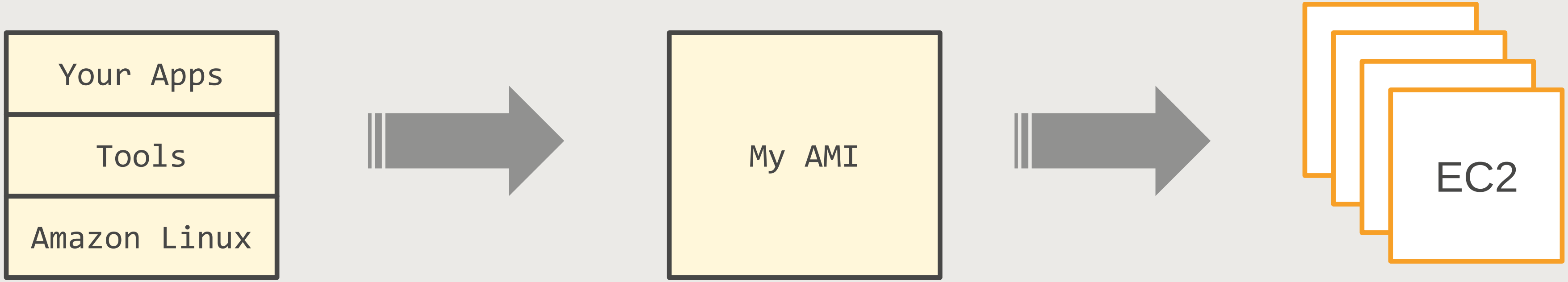
Amazon Linux, Red Hat, Suse, Ubuntu, Microsoft, macOS

Deep Learning AMI

Pre-installs commonly used tools

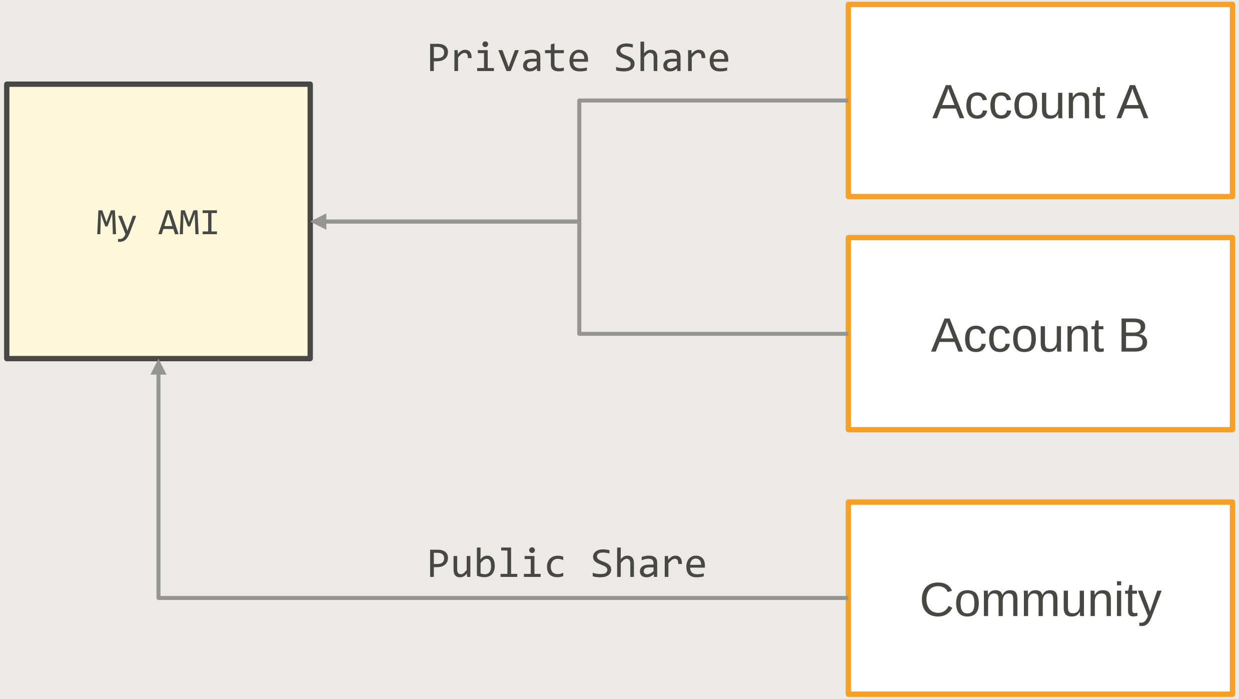
My AMI

Build your own AMI

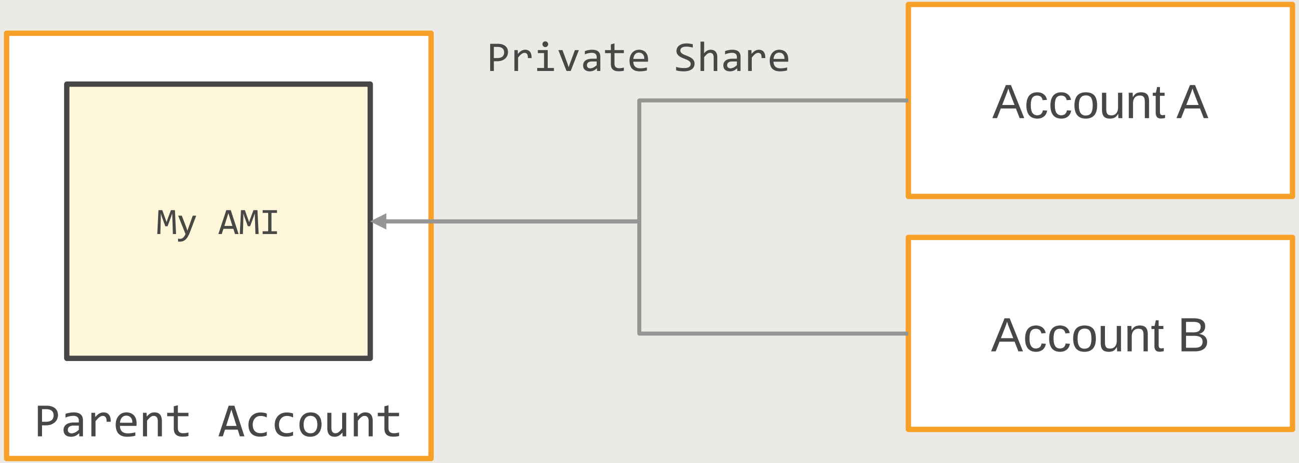


Reduce time needed to launch instances!

Share Your AMI



AMI Reuse Inside an Organization



Marketplace AMI

Ready to use AMI from Popular Vendors

Security Software, VPN, Business Apps, DevOps

EC2 Instance Families

Amazon Elastic Compute Cloud (Amazon EC2) offers … over 475 instances and choice of the latest processor, storage, networking, operating system, and purchase model to help you best match the needs of your workload.

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

Instance Configuration

CPU
[Intel, AMD, ARM]

Memory

Graphics

Storage

Network

Operating System
[Linux, Windows, macOS]

Organized by Instance Family

General Purpose Family

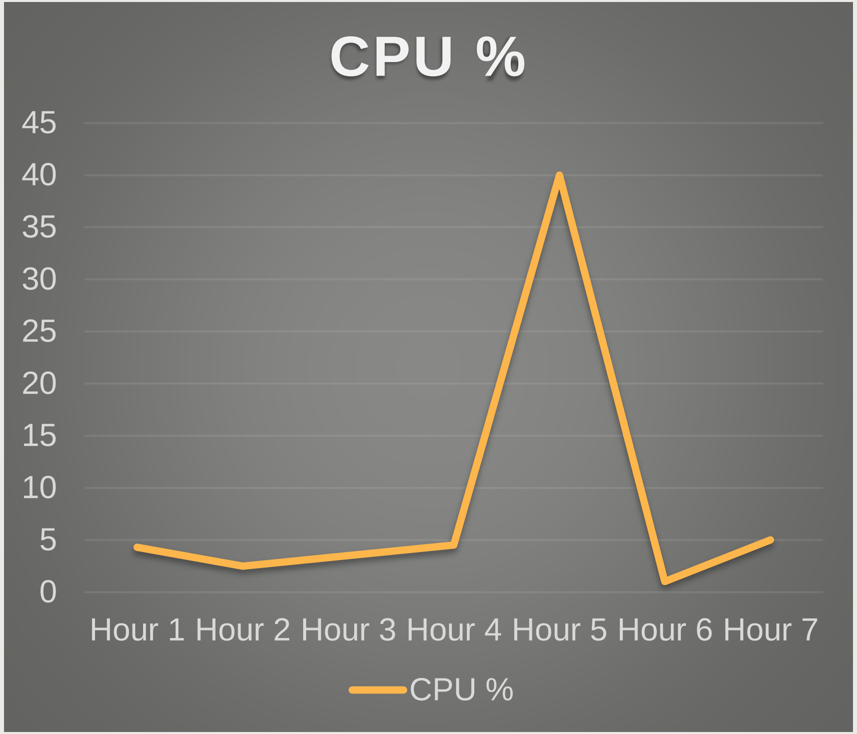
- Balanced Performance
- Suitable for many business applications
- Burstable and Fixed performance instances

Burstable Instances

T-type instances [T2,T3]

Business Apps – Low to Moderate CPU utilization with occasional increase

Burstable instances are designed for these workloads



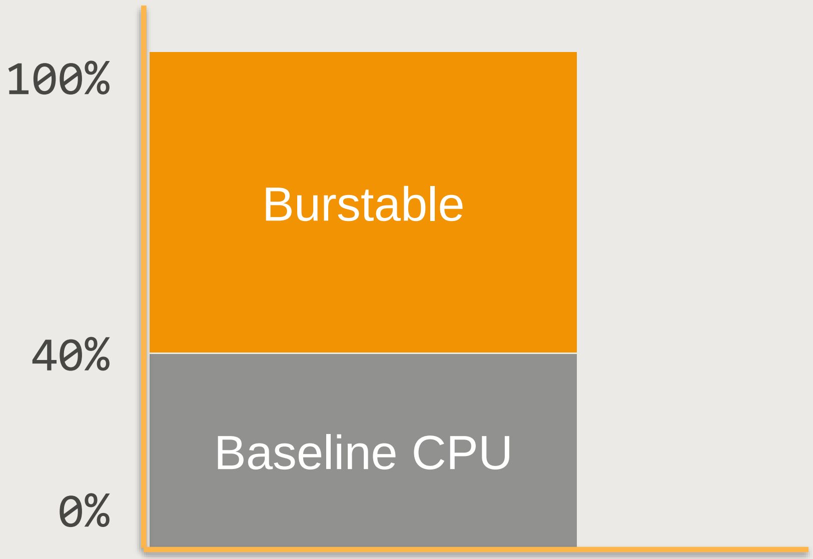
Burstable Instance - CPU

Guaranteed Baseline CPU Performance [5-40%]

Instance earns a CPU Credit when usage is less than baseline

Instance performance can burst up to 100% using CPU credits

CPU Utilization

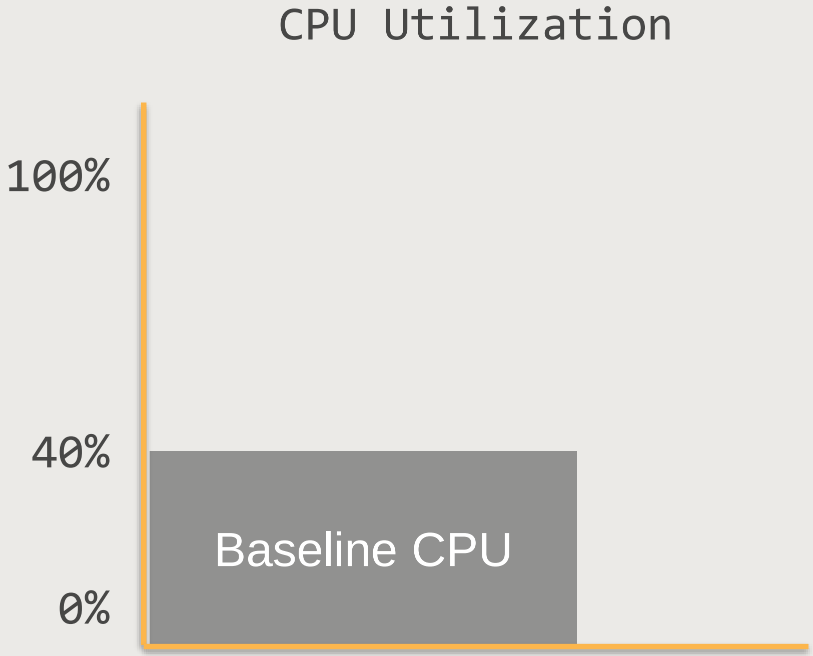


No CPU Credit

CPU is throttled to baseline performance

Throttling is not desirable

Monitor CPUCreditBalance metric in CloudWatch



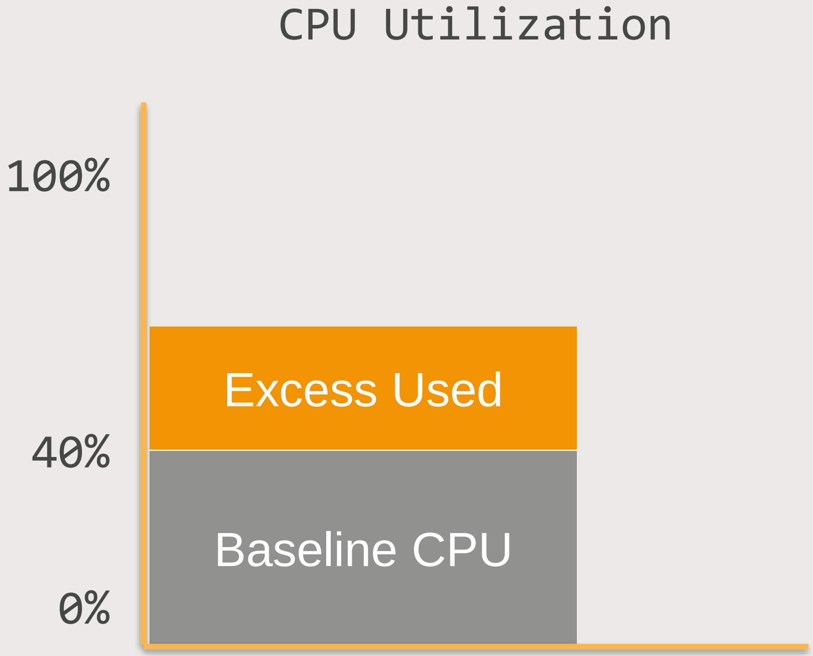
Unlimited Mode [Burstable Instance]

Unlimited mode – no need to worry about throttling

Pay for excess capacity consumed

Recommendation: Enable unlimited mode

- Auto enabled in T3
- You need to enable in T2



Burstable Instance Usage

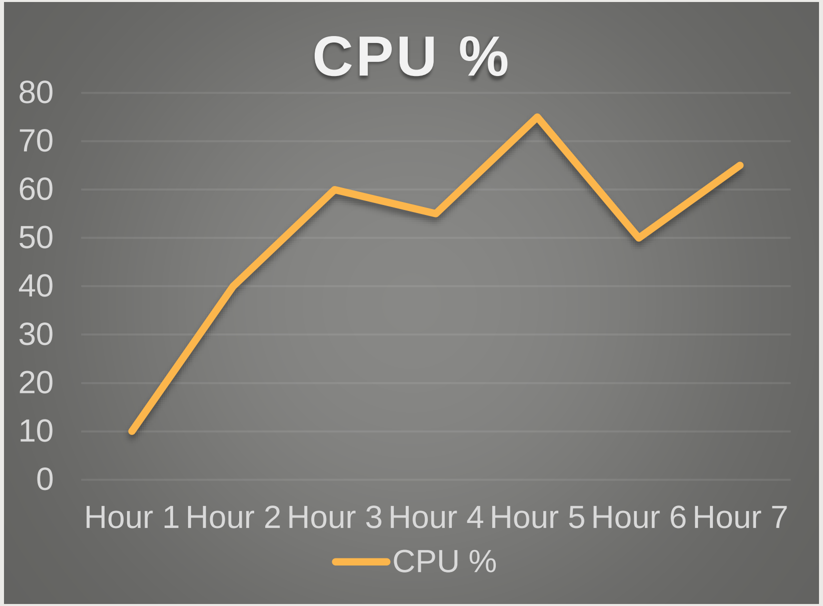
Burstable instances are suitable for micro-services, small and medium databases, virtual desktops, and business-critical applications

https://aws.amazon.com/ec2/instance-types/t3/

Fixed Performance Instances

M-type instances [M4,M5]

Suitable for apps that consistently use high CPU



Fixed Performance Instance Usage

M5 instances are suitable for web and application servers, small and mid-sized databases, cluster computing, gaming servers, caching fleets, and app development environments

https://aws.amazon.com/ec2/instance-types/m5/

Compute Optimized Family

CPU INTENSIVE WORKLOAD

LATEST GENERATION CPU

C-TYPE INSTANCES

C5,C6,C7

Compute Optimized Instance Usage

Batch processing workloads, media transcoding, high-performance web servers, high-performance computing (HPC), scientific modeling, gaming servers, ad server engines, machine learning, and other compute-intensive applications.

https://aws.amazon.com/ec2/instance-types/

Memory Optimized Family

Designed for workloads that process large datasets in memory

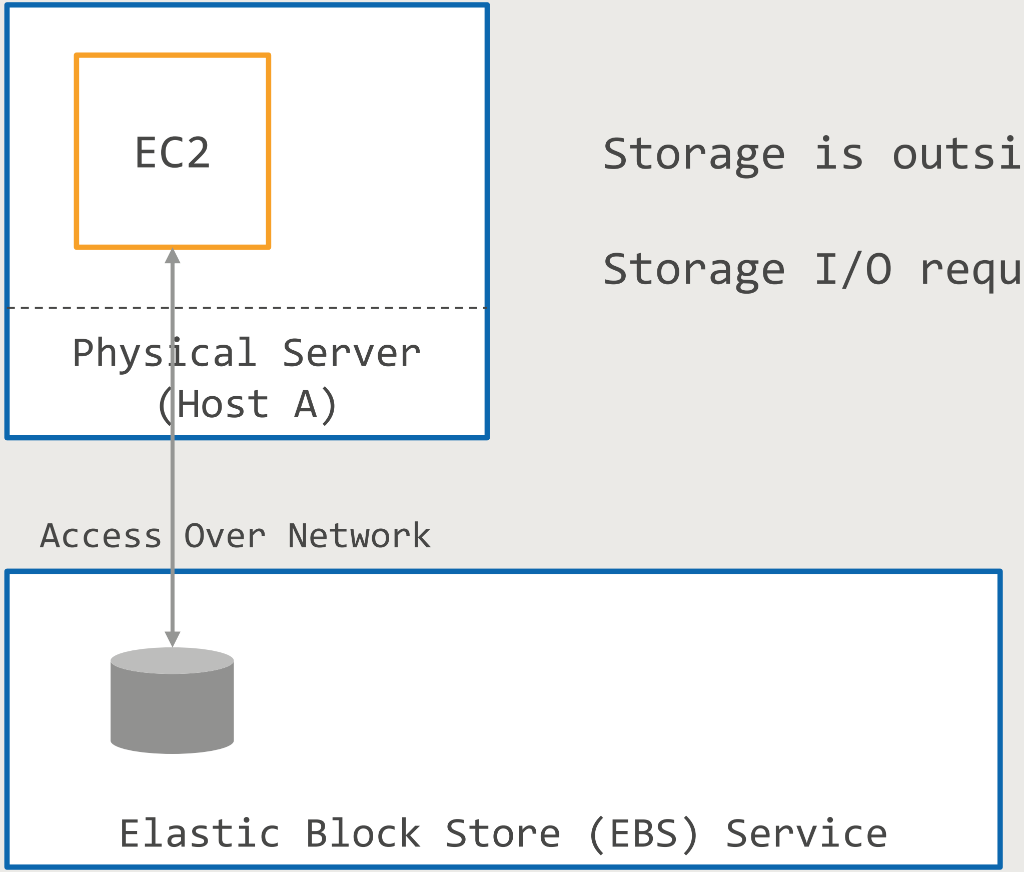
R-type instances (R5,R6) and more

Ideal for in-memory databases, caches, and big data analytics

Storage Optimized Family

Instances come with high-performance instance storage

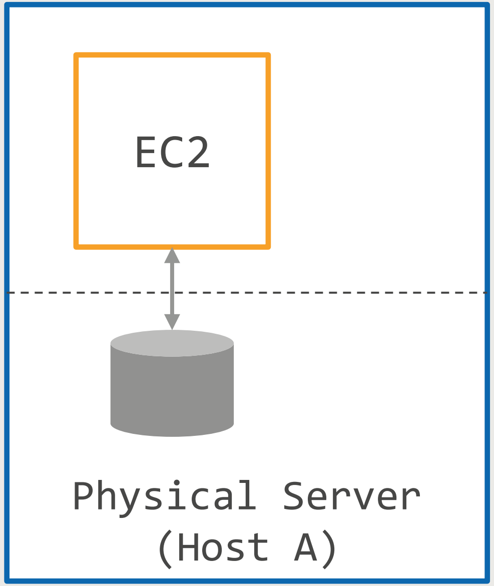
Elastic Block Store (EBS)



Storage is outside of host server

Storage I/O requests go over the network

Instance Storage



Storage is part of the host server

Direct access to storage

I-type instances [I3,I4]

SSD Storage

Very high random I/O and sequential reads

Ideal for NoSQL databases, in-memory databases, data warehousing, Elasticsearch, and analytics workloads

https://aws.amazon.com/ec2/instance-types/i3/

D-type instances [D2,D3]

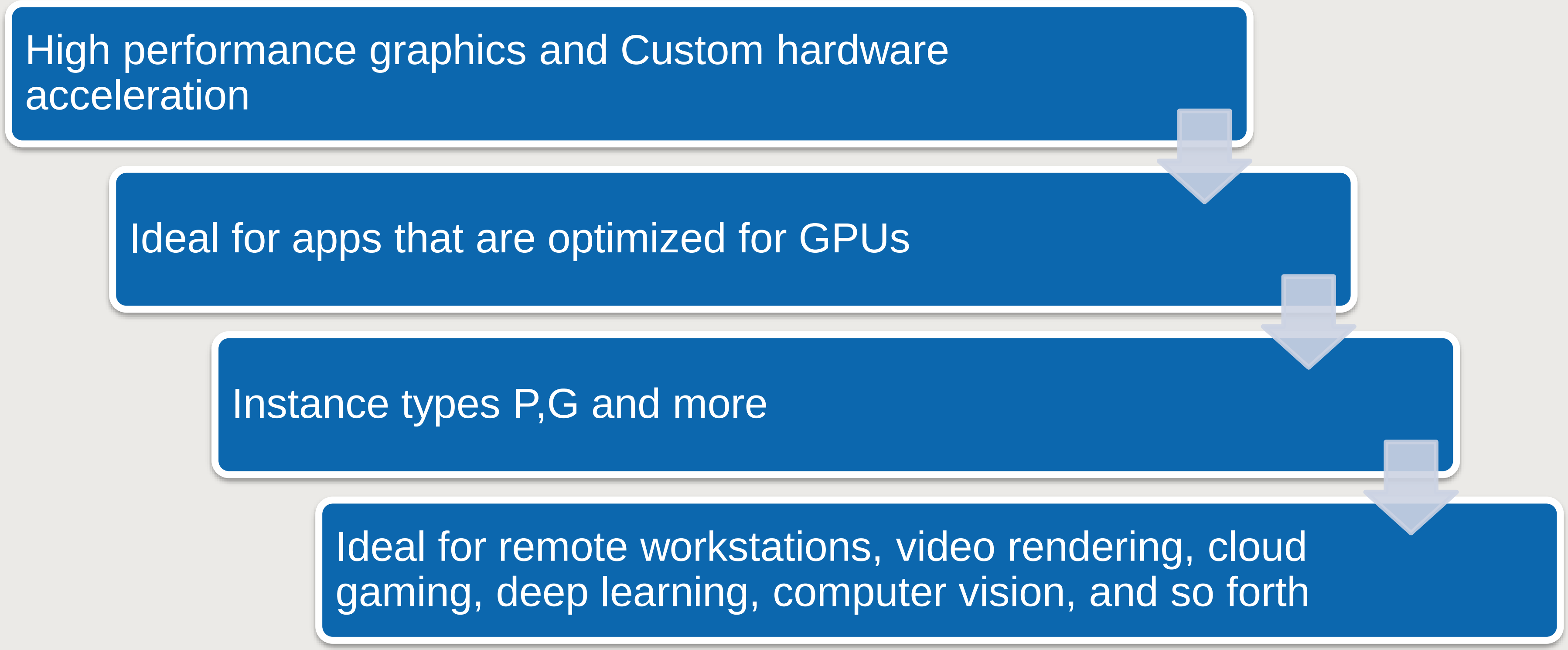
High-capacity HDD Storage [Magnetic]

High sequential I/O

Ideal for big data and analytics, data warehousing, and distributed file systems

https://aws.amazon.com/ec2/instance-types/d3/

Accelerated Computing Family



High performance graphics and Custom hardware acceleration

Ideal for apps that are optimized for GPUs

Instance types P,G and more

Ideal for remote workstations, video rendering, cloud gaming, deep learning, computer vision, and so forth

EBS Optimized Instances

Instance has dedicated bandwidth for EBS Storage I/O

Consistent throughput and high performance

Enabled by default on latest generation

Previous generation - You need to enable for supported instance types for additional hourly cost

Enhanced Networking

Higher bandwidth

Higher packets per second

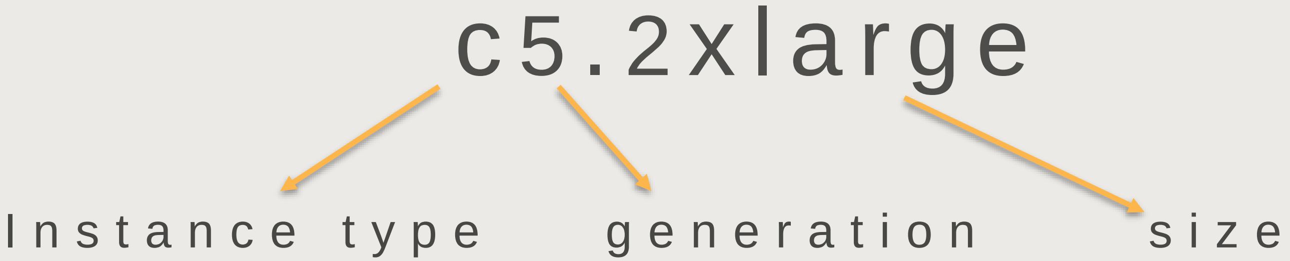
Reduced jitter

No-cost option on supported instance types

NVMe

Low-latency and High-performance Interface for SSD Instance Storage

Instance Naming Convention



c5.2xlarge

Instance type

generation

size

c5.2xlarge = Compute Optimized, 5th generation, 2xlarge (8 vCPUs, 16 GB Memory)

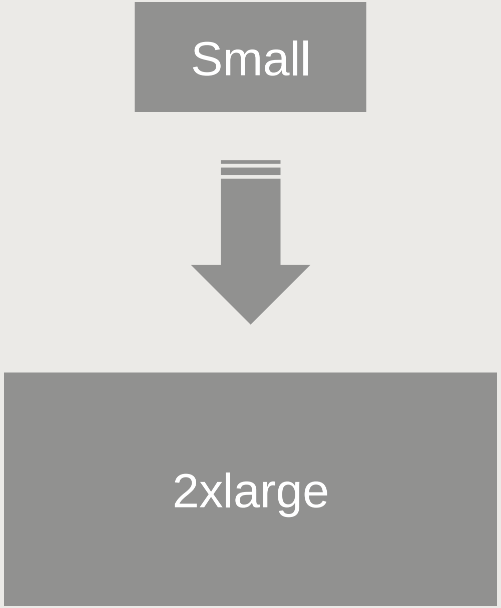
Resize Instances

Easily resize instances

Stop-Change-Start

Check compatibility

[OS, Storage, and so forth]



Single Tenant Options

Dedicated host

- Useful for BYOL (bring your own license) tied to physical sockets/cores

Bare Metal

- Direct access to hardware
- Use a different hypervisor

Where to start?

Map app to instance family

Pick an appropriate size

[small, large, 2xlarge and so forth]

Run performance tests to right-size

Placement Groups

Cloud Best Practice

Distribute instances across multiple availability zones

Protects from hardware and AZ failures

But there is always an exception!

HPC and ML

Use a cluster of instances [in 1000s]

Requires very high network I/O performance

Exchange of data and messages among instances

Supercomputer on AWS

*“Descartes Labs Achieves **#40 in TOP500** with Cloud-based Supercomputing Demonstration Powered by AWS, Signaling New Era for Geospatial Data Analysis at Scale”*

https://blog.descarteslabs.com/achieves-number-41-in-top500-cloud-based-supercomputing

https://aws.amazon.com/blogs/aws/planetary-scale-computing-9-95-pflops-position-41-on-the-top500-list/

Placement Groups

Minimize network latency and enable very high network throughput

Three types

- Cluster
- Partition
- Spread

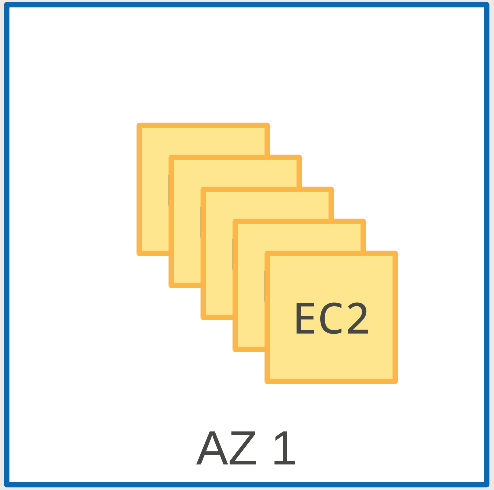
Cluster Placement Group

Instances are packed closely together in a single Availability Zone

Instances share rack and network infrastructure

Low network latency

Enhanced Networking recommended



Partition Placement Group

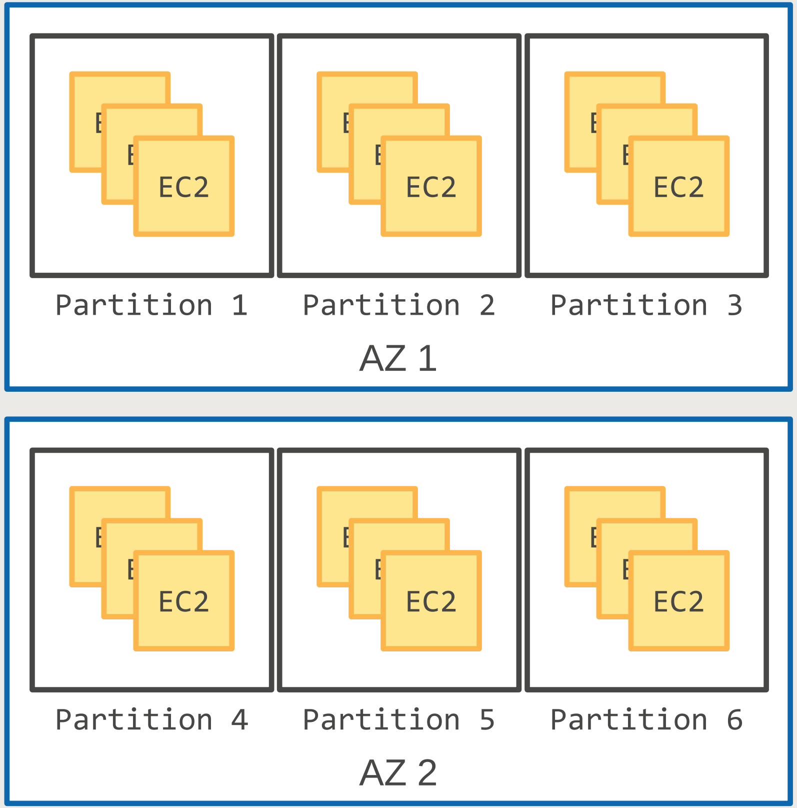
Minimizes impact to due to hardware failure

Instances are distributed across specified number of partitions

Each partition has a separate rack, power source and network

Place partitions in multiple Azs

Recommended for HDFS, HBase, Cassandra

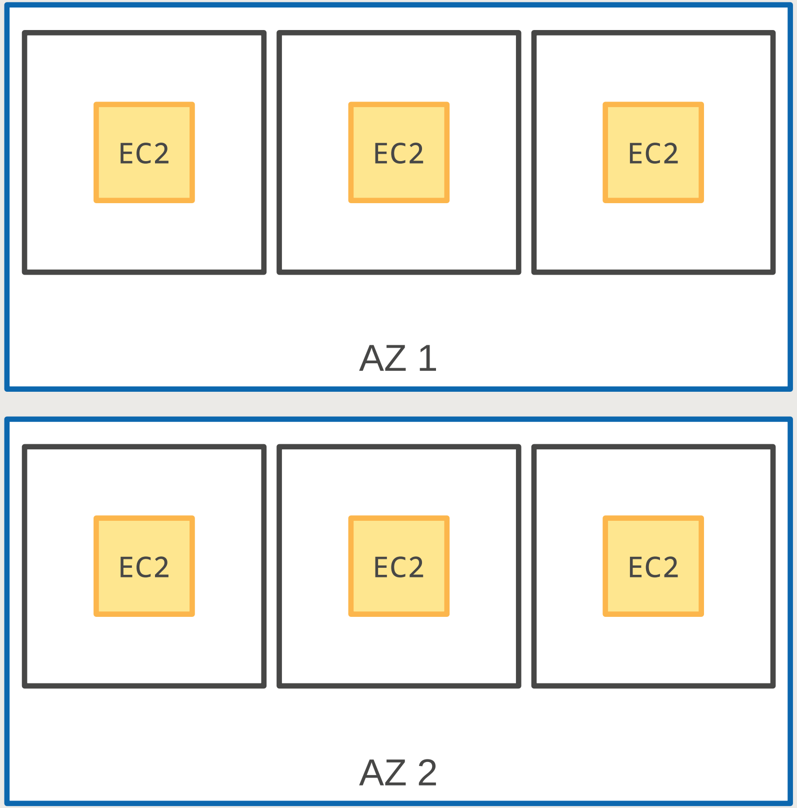


Spread Placement Group

Each instance in a separate rack, power source and network

Small number of critical instances that are kept separate from each other

Span multiple Availability Zones



Check Account Quota Limits

Is quota sufficient to launch the required instances?

Contact AWS Support to increase

Handling Capacity Issues

Use EC2 On-Demand Capacity Reservation

No long-term commitment

Specify required Number of instances, AZ, Instance attributes

Billing starts when reservation state is Active with a guaranteed access to capacity

EC2 On-Demand Capacity Reservation

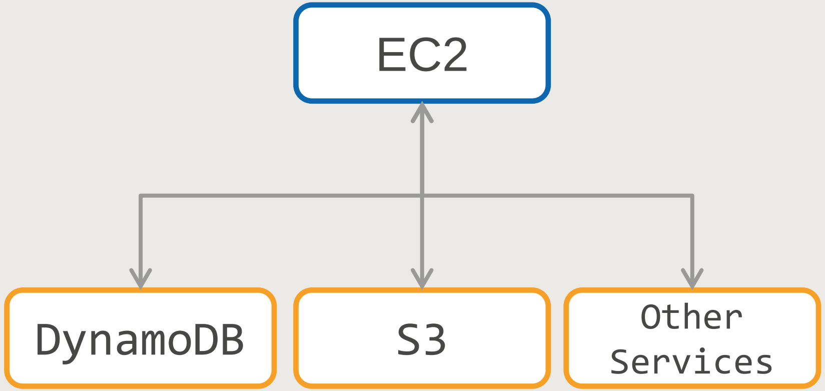
When reservation is Active

- You are charged on-demand rates
- Launch instance to match the reservation attributes
- Unused reservation shows up in the bill

Cancel reservation when you no longer need it

EC2 – IAM Roles

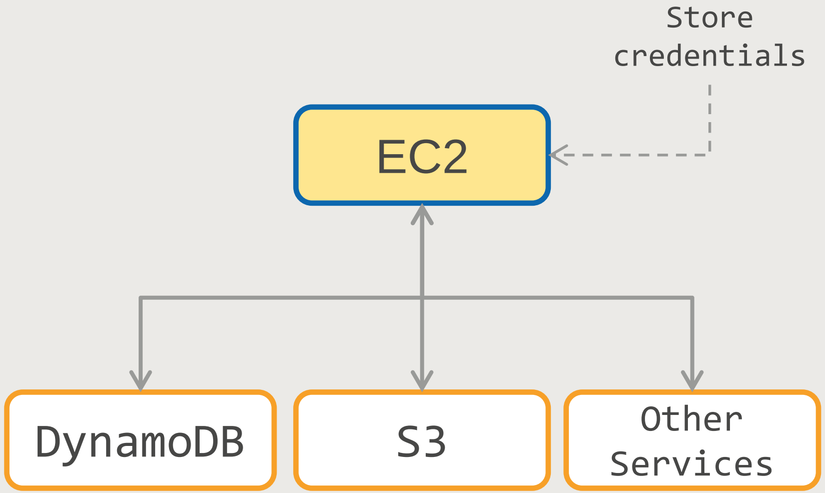
Application Access to AWS Resources



How to grant access

API call to AWS services need to be signed

Treat app as another IAM user

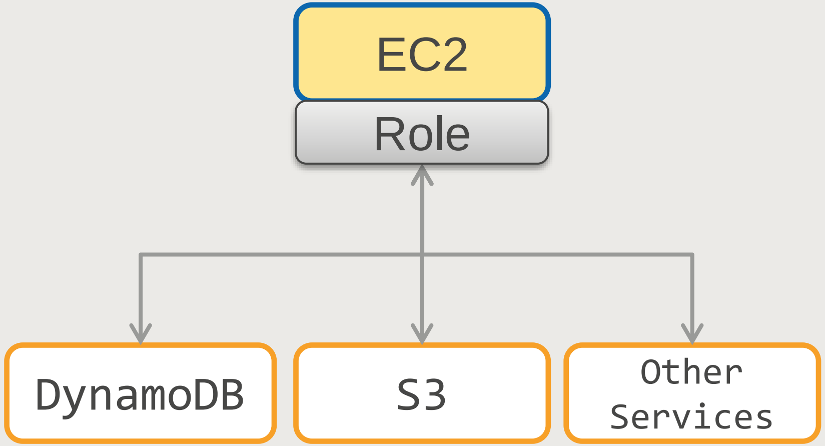


Create Access Key Credentials and distribute with app

Issues:

1. How to securely distribute
2. How to protect from misuse
3. How to rotate credentials

Use IAM Roles



Create IAM Role and attach to instance

Benefits

1. App can get temporary credentials when needed [using EC2 metadata service]
2. Credentials are automatically rotated
3. No need to maintain credentials in the app
4. AWS SDK/CLI has built-in support

Chandra Lingam

75,000+ Students

aws certified Solutions Architect Associate
aws certified Developer Associate
aws certified SysOps Administrator Associate
aws certified Cloud Practitioner
aws certified Solutions Architect Professional
aws certified Machine Learning Specialty
aws certified Security Specialty

Instructor, Course Developer

7X AWS Certified

For a list of courses, visit
https://www.cloudwavetraining.com/

Connect with me on LinkedIn
https://www.linkedin.com/in/chandralingam/

Cloud Wave LLC