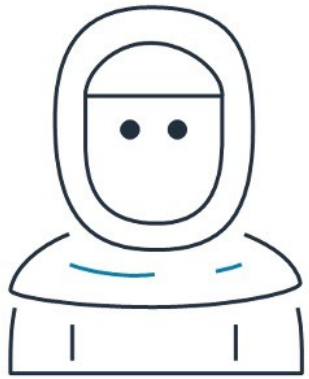
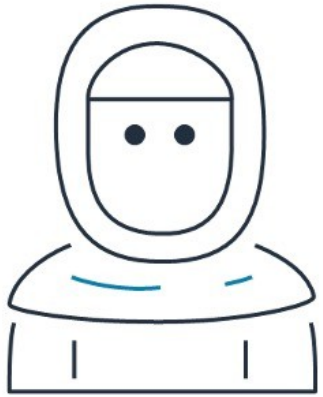
Agenda

- **Ec2 instance types**
- **Understanding Load Balancer**
- **Creating Load Balancer and attaching EC2 machines**
- **Experience the Load Balancer**

Coffee shop tasks

Employee 1	Employee 2	Employee 3
		
		

Coffee shop tasks specialization

Employee 1	Employee 2	Employee 3
		
		
Make coffee	Process transactions	Order supplies

Amazon EC2 instance types (1 of 2)

General purpose	Compute optimized	Memory optimized
• Balances compute, memory, and networking resources	• Offers high-performance processors	• Delivers fast performance for memory-intensive workloads
• Suitable for a broad range of workloads	• Ideal for compute-intensive applications and batch processing workloads	• Well suited for high-performance databases

Amazon EC2 instance types (2 of 2)

Accelerated computing

- Uses hardware accelerators to expedite data processing
- Ideal for application streaming and graphics workloads

Storage optimized

- Offers low latency and high input/output operations per second (IOPS)
- Suitable for workloads such as distributed file systems and data warehousing applications

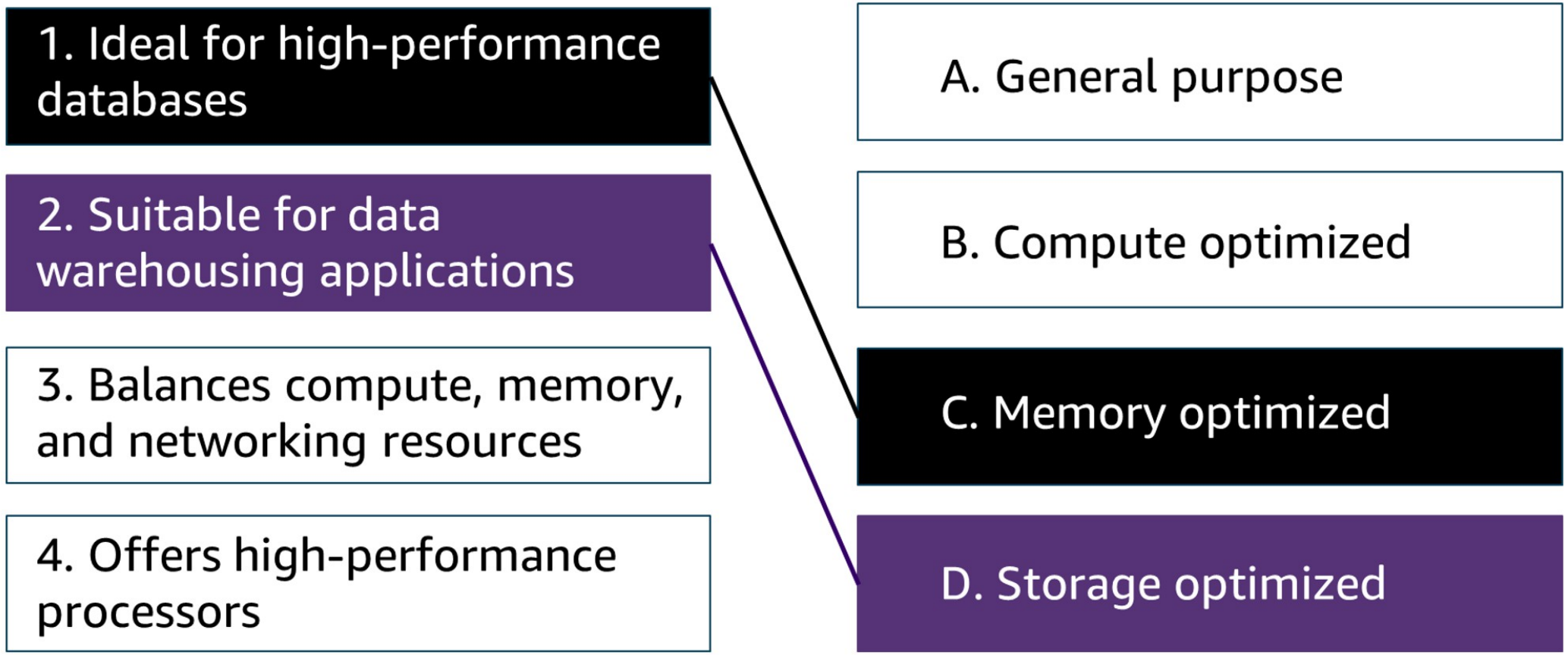
Match: Amazon EC2 instance types (1 of 5)

1. Ideal for high-performance databases	A. General purpose
2. Suitable for data warehousing applications	B. Compute optimized
3. Balances compute, memory, and networking resources	C. Memory optimized
4. Offers high-performance processors	D. Storage optimized

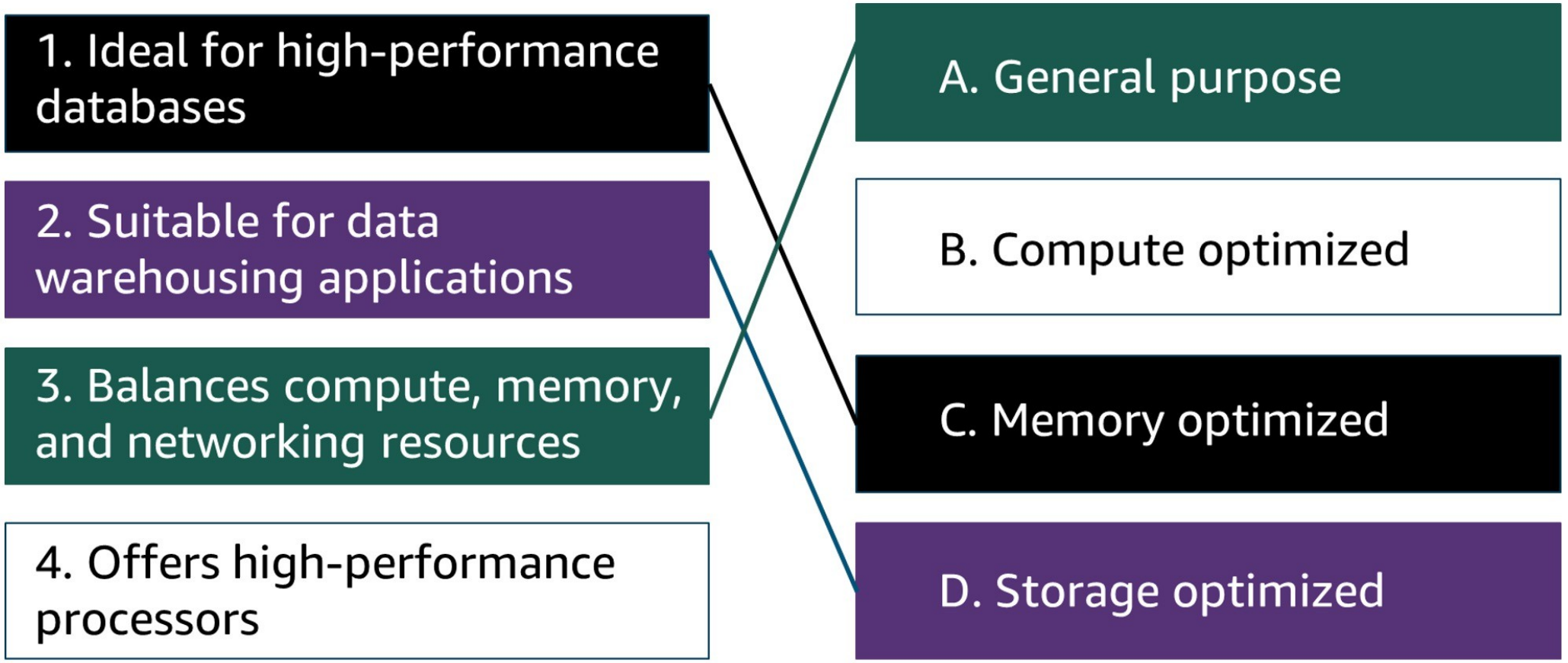
Match: Amazon EC2 instance types (2 of 5)

1. Ideal for high-performance databases	A. General purpose
2. Suitable for data warehousing applications	B. Compute optimized
3. Balances compute, memory, and networking resources	C. Memory optimized
4. Offers high-performance processors	D. Storage optimized

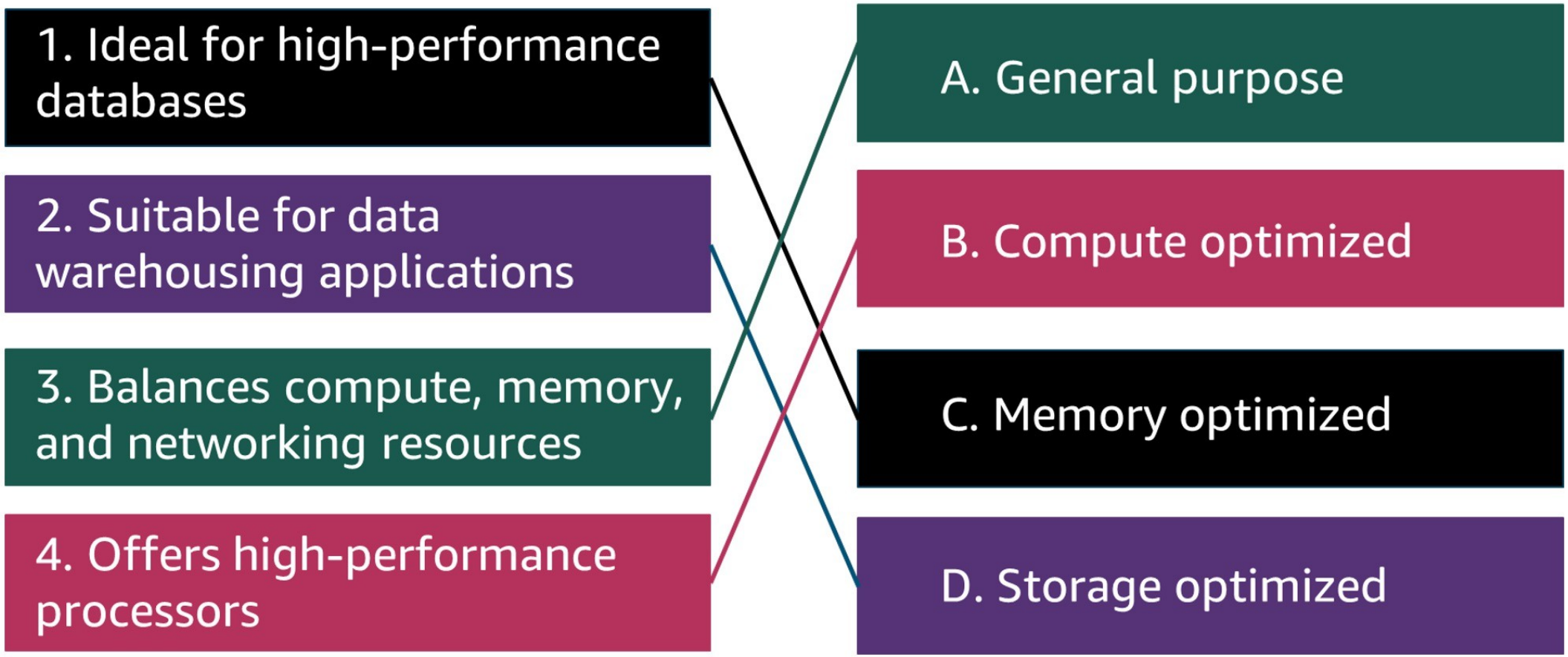
Match: Amazon EC2 instance types (3 of 5)



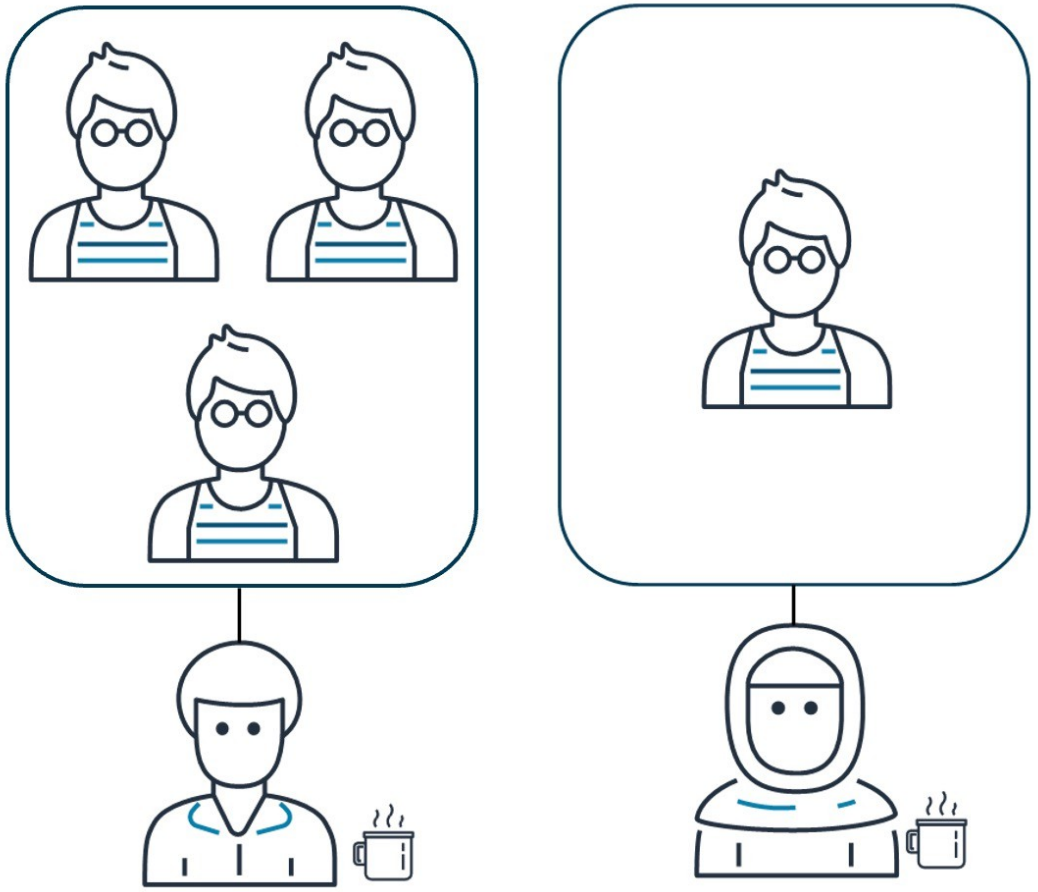
Match: Amazon EC2 instance types (4 of 5)



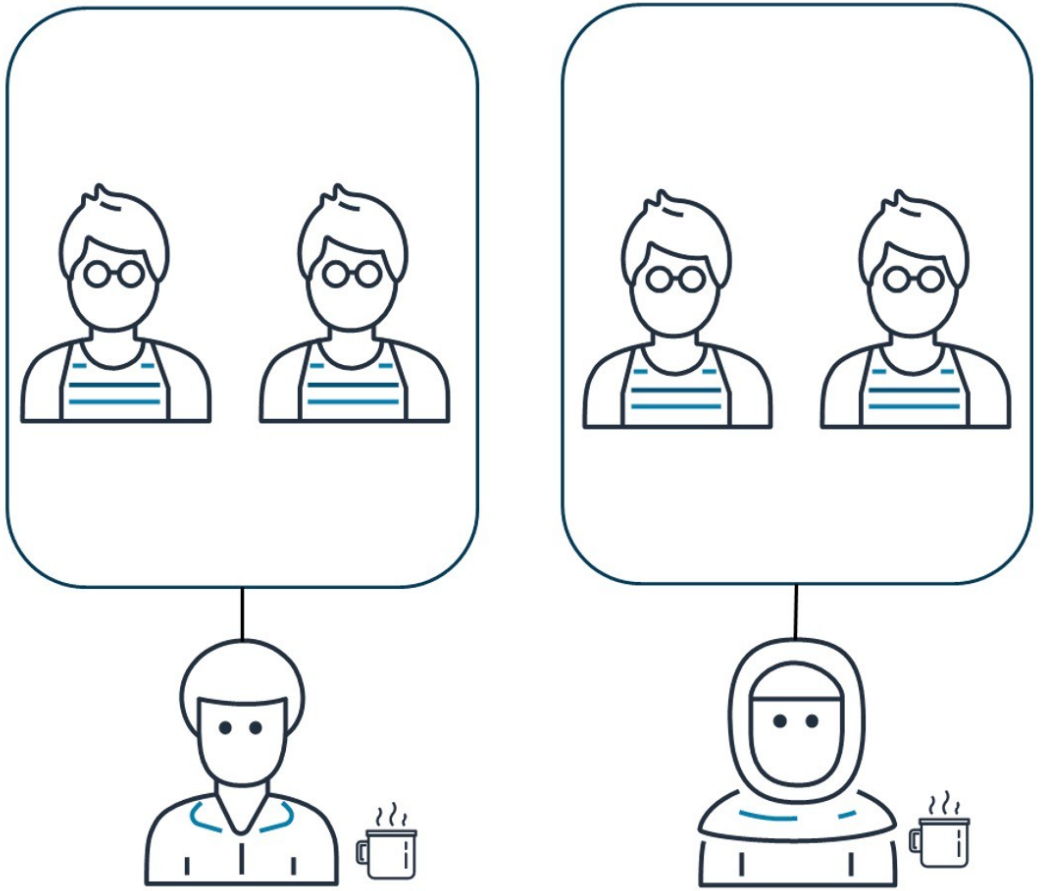
Match: Amazon EC2 instance types (5 of 5)



Unbalanced workload



Balanced workload



Elastic Load Balancing

Elastic Load Balancing

- Automatically distributes traffic across multiple resources
- Provides a single point of contact for your Auto Scaling group

Load Balancer

- A load balancer accepts incoming traffic from clients and routes requests to EC2 instances (Targets), Such as EC2 instances, containers, and IP addresses, in one or more Availability Zones.

- The load balancer also monitors the health of its registered targets and ensures that it routes traffic only to healthy targets.

- When the load balancer detects an unhealthy target, it stops routing traffic to that target. It then resumes routing traffic to that target when it detects that the target is healthy again.

Use Cases

Secure

Secure access through a single point

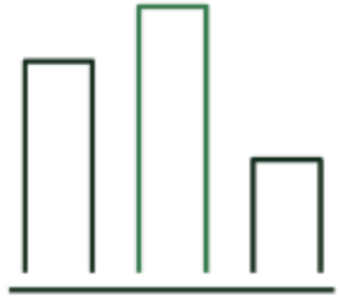
Decoupled

Decouple your application environment

Fault tolerant

Provide high availability and fault tolerance

Expansive



Increase elasticity and scalability

Features

- High availability (HA)
- Health checks
- Security features
- TLS termination
- Layer 4 or layer 7 load balancing
- Operational monitoring

Types of Load Balancer

- Application Load Balancer – Application Load Balancer Operates at layer-7 of the OSI (Open Systems Interconnection) model. ALB can distribute incoming traffic to multiple targets based on the application-level details such as HTTP and HTTPS traffic, Content of the message.
- Network Load Balancer – Network Load Balancer Operates at layer-4 of the OSI (Open Systems Interconnection) model. It is useful for load balancing based on TCP (Transmission Control Protocol) and UDP (User Datagram Protocol). NLB is capable of handling millions of requests per second while maintaining high throughput and ultra-low latencies. NLB is very well optimized for handling sudden and volatile traffic patterns.
- Gateway Load Balance – Gateway Load Balancer Operates at layer-3 of the OSI (Open Systems Interconnection) model. It allows you to deploy, scale, and manage virtual appliances, such as firewalls, intrusion detection and prevention systems, and deep packet inspection systems.