AWS Elastic Load Balancer (ELB) & Auto Scaling (ASG) – Cheat Sheet

Elastic Load Balancer (ELB) – Key Points

- Distributes incoming traffic across multiple EC2 instances
- Improves High Availability
- Prevents single point of failure
- Works across multiple Availability Zones
- Performs health checks and stops routing to unhealthy instances
- Does NOT launch or replace instances

Types of Load Balancers (Basic Recognition)

- Application Load Balancer (ALB) – HTTP/HTTPS
- Network Load Balancer (NLB) – TCP/UDP
- Gateway Load Balancer (GLB) – Advanced networking

Auto Scaling Group (ASG) – Key Points

- Automatically adjusts number of EC2 instances
- Launches instances when demand increases
- Terminates instances when demand decreases
- Replaces unhealthy instances
- Maintains desired capacity

Important ASG Terms

- Minimum Capacity – Lowest number of instances allowed
- Maximum Capacity – Highest number of instances allowed
- Desired Capacity – Target number of running instances

Scaling Policies

- Scale based on CPU utilization
- Scale based on CloudWatch alarms
- Supports elasticity and cost optimization

ELB vs Auto Scaling Differences

- ELB distributes traffic
- Auto Scaling launches and terminates instances
- ELB performs health checks
- Auto Scaling replaces unhealthy instances

Common Exam Traps

- ELB does NOT replace instances
- ASG does NOT distribute traffic
- Single AZ deployment is still a failure risk
- Elasticity means scaling automatically
- High Availability requires Multi-AZ setup
