AWS Certified Solutions Architect: Associate - 5.0 ELB, CloudWatch and Auto Scaling

filename: amazon-acsaa-5-1-elb_and_auto_scaling
Title: ELB and Auto Scaling
Subtitle: AWS Certified Solutions Architect: Associate

5.1 ELB and Auto Scaling

- Elastic Load Balancing
    - Distributes traffic load across multiple instances
    - Managed by AWS
        - Highly available
        - Scales in/out automatically
    - Types
        - Internet facing
            - Uses public DNS/IP for routing
            - You should always use the DNS name
            - Publicly available
        - Internal load balancers
            - Used in VPCs with private subnets
        - HTTPS load balancers
            - SSL is terminated on the load balancer (SSL offload)
            - Certificate must be installed on the ELB
            - Does not support *Server Name Indication* (SNI)
                - Causes issues if multiple sites/domains are hosted behind the ELB
                - SSL cert must contain *Subject Alternative Name* (SAN) for each domain behind the ELB
    - Listeners
        - ELB only balances on specified methods
        - Supported methods
            - HTTP
            - HTTPS
            - TCP
            - SSL
    - Configuration
        - Idle connection timeout
            - 60 seconds by default
            - With HTTP/HTTPS the *keep-alive* option can be used
            - Otherwise, idle connections may reconnect to a different instance
        - Cross-zone Load Balancing
        - Connection Draining
            - Allows the ELB to stop sending connections to instances that are deregistering or unhealthy
            - Connections are held open for 300 seconds by default
            - Attempts to complete transaction if the instance becomes healthy again
        - Proxy protocol
            - Human readable header is attached to inbound traffic
            - Contains IP and port information from original traffic
            - Not usually necessary and can even cause issues
        - Sticky sessions
            - Also called *session affinity*
            - Normally, all connections are routed independantly
            - Sticky sessions route all requests from a client to the same instance
            - Uses a cookie to track the sessions
        - Health check
            - Testing method to determine if an instance is healthy
            - Methods
                - Ping
                - Connection
                - Web page
- Amazon CloudWatch
    - Monitoring service hosted inside of AWS
    - Provides tracking and alerting
    - Actions can be extended with AWS Simple Notification Service, Lambda, etc.
    - Used in conjunction with auto scaling
- Auto Scaling
    - Allows scaling instances in/out based on performance data gathered from CloudWatch
    - Auto Scaling Plans
        - Maintain current levels
            - Enforces a minimum # of instances

        - Manual scaling
        - Scheduled scaling
        - Dynamic scaling
    - Limits
        - `aws autoscaling describe-account-limits`
        - 100 launch configurations per region
        - 20 EC2 instances per region
    - Components (Diagram)
        - Launch configuration
            - Name
            - AMI
            - Instance Type
            - Security Group
            - Instance Key Pair
        - Auto Scaling Group
            - Name
            - Launch configuration
            - Availability Zones
            - Minimum Size
            - Desired Capacity
            - Maximum Capacity
            - Load Balancers
        - Scaling Policy (2 diagrams)
            - Associates CloudWatch alarms with scaling operations
    - Scale out quickly, scale in slowly