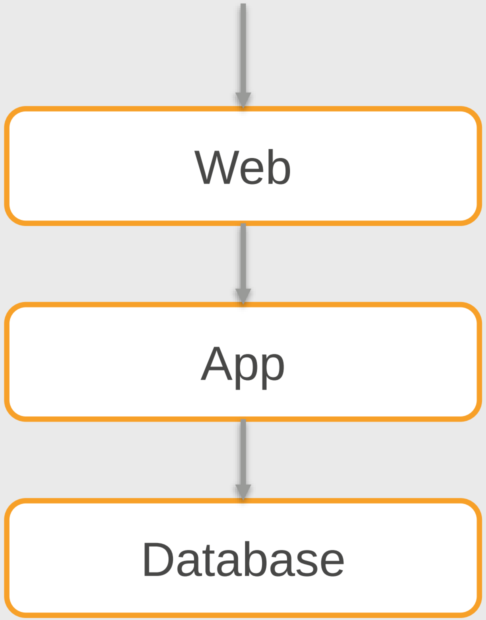
Architecture Walk-thru

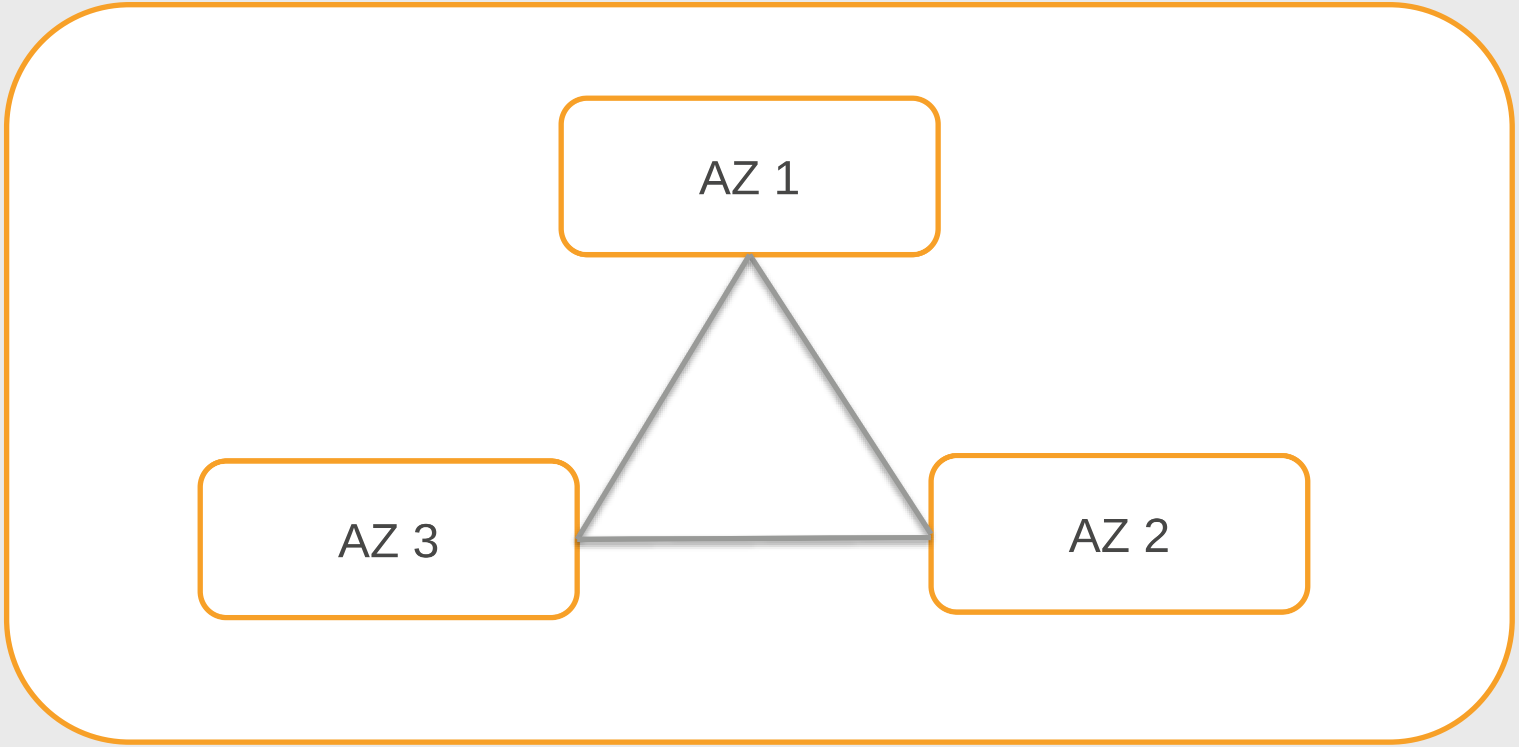
- Server based
- Serverless

Online Order Processing Application



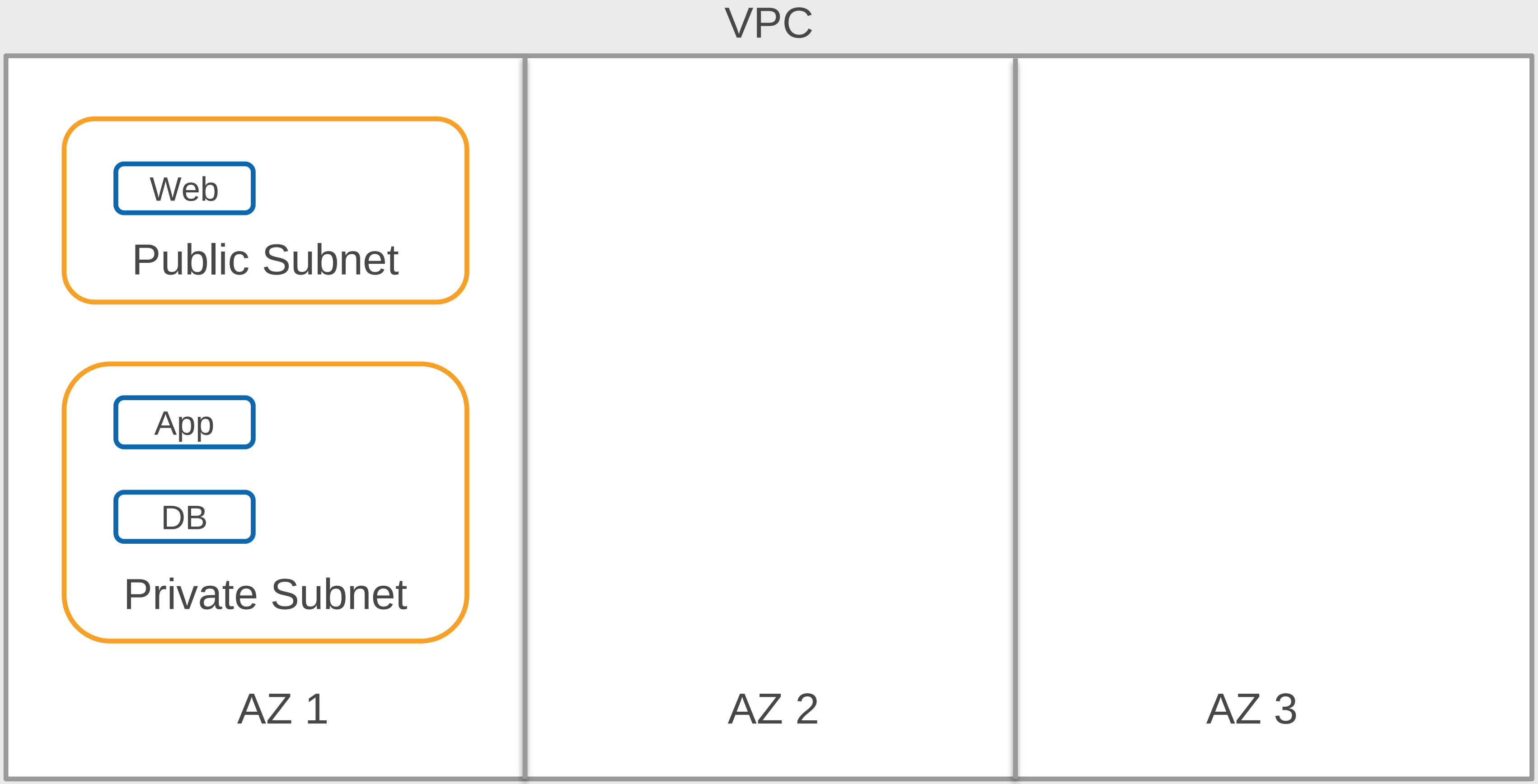
- Resilient
- Scaling
- Security
- Cost

Region

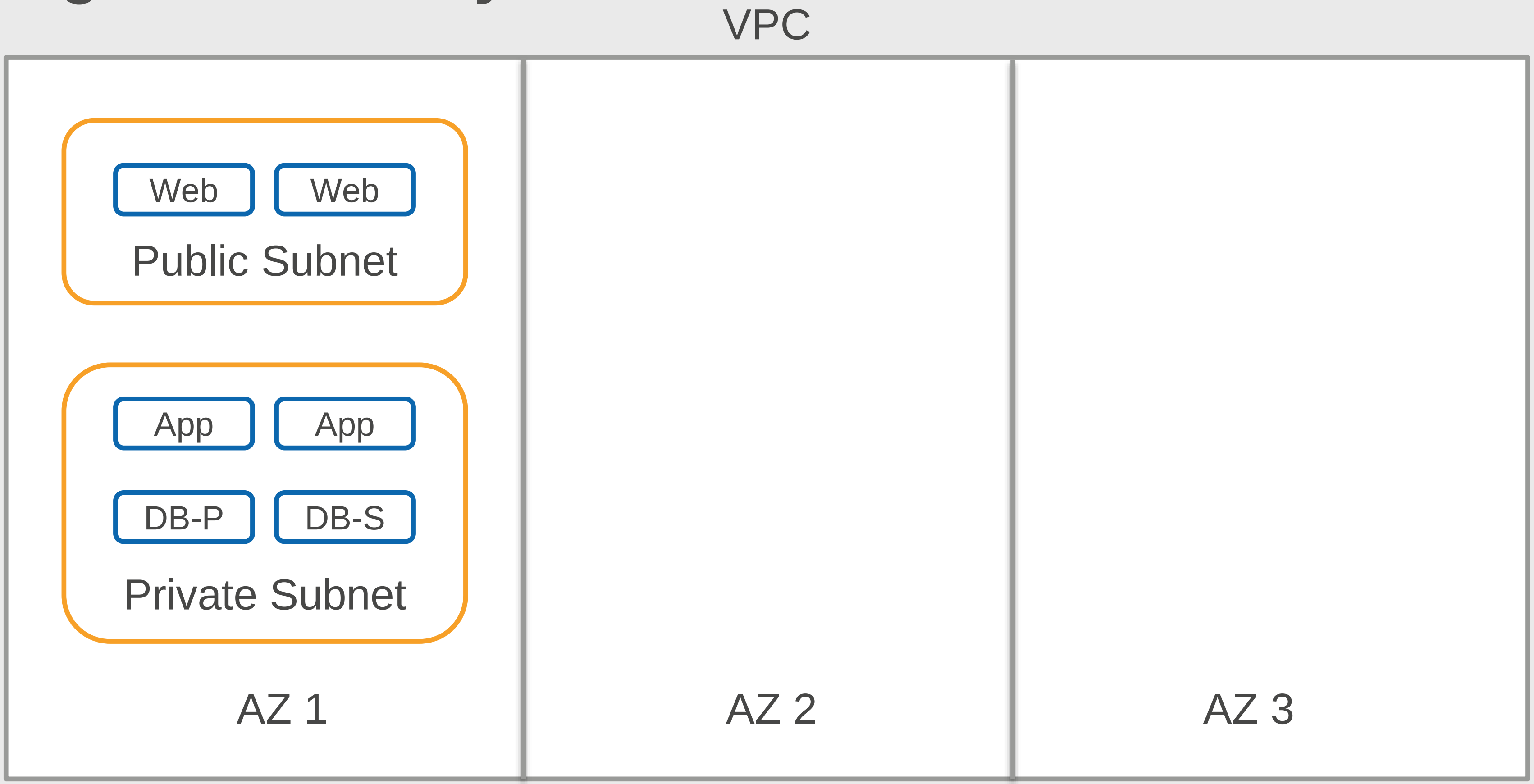


Application should be spread across two or more availability zones

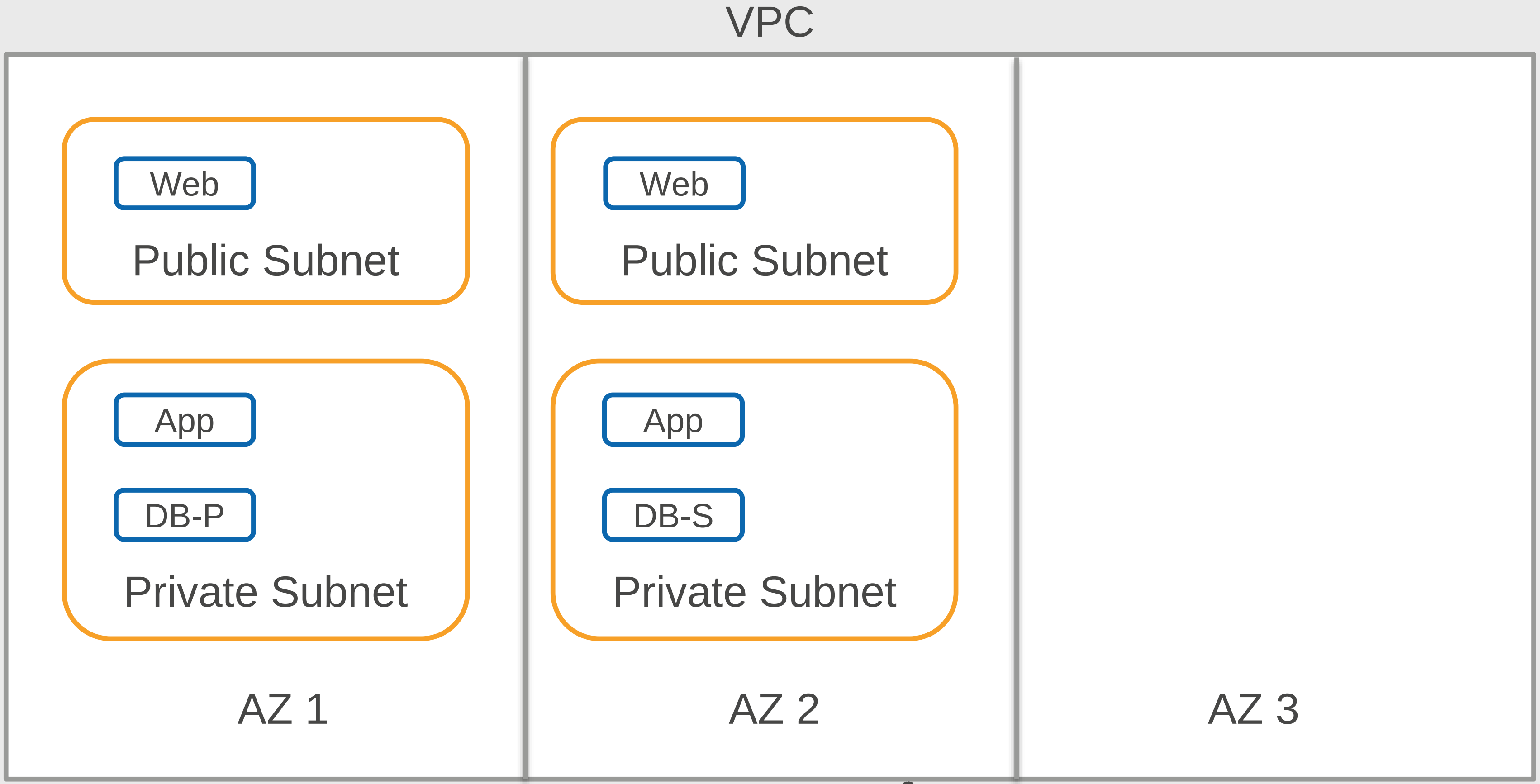
Network



High Availability

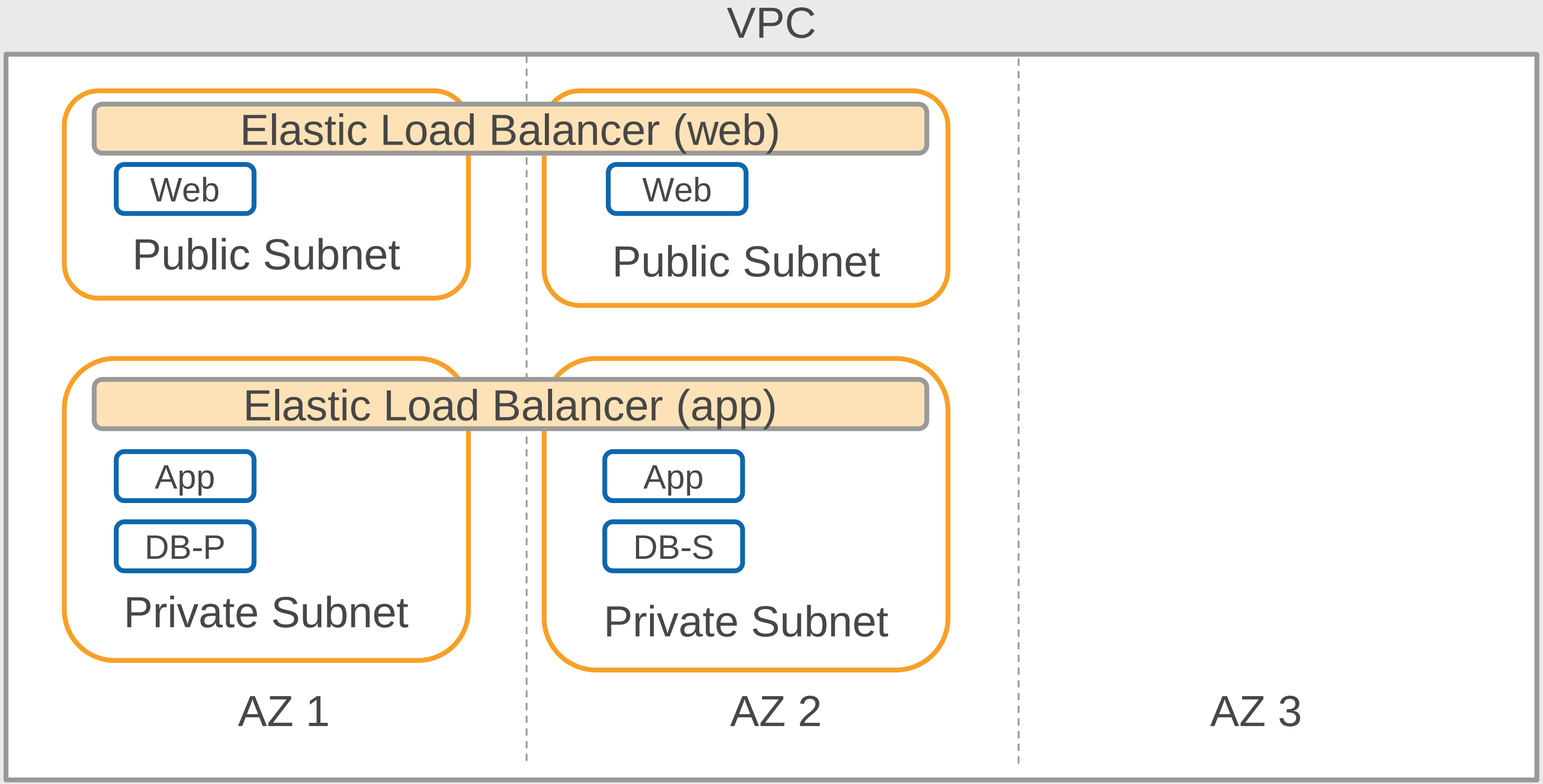


High Availability – Multi-AZ

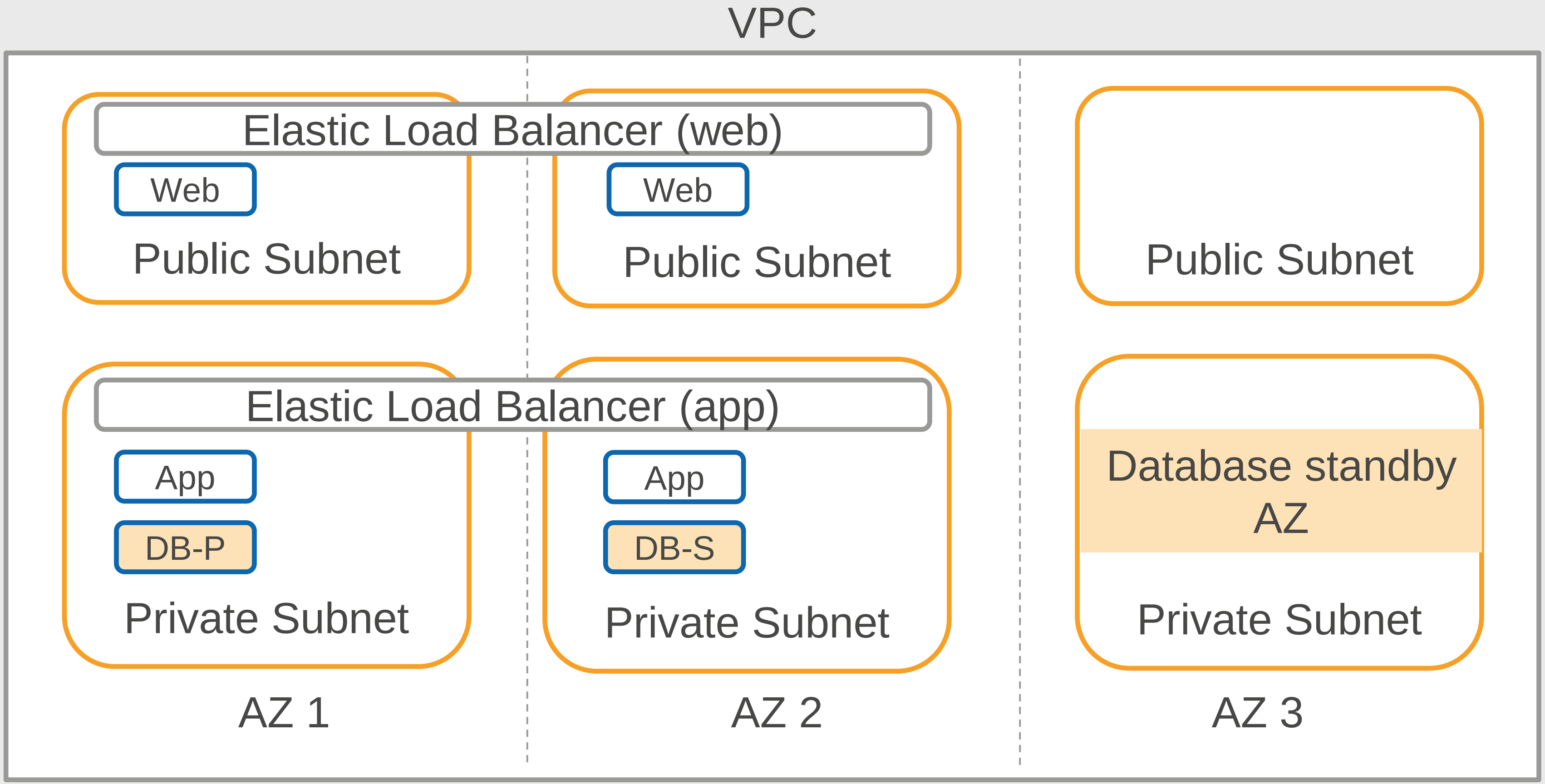


No single point of entry

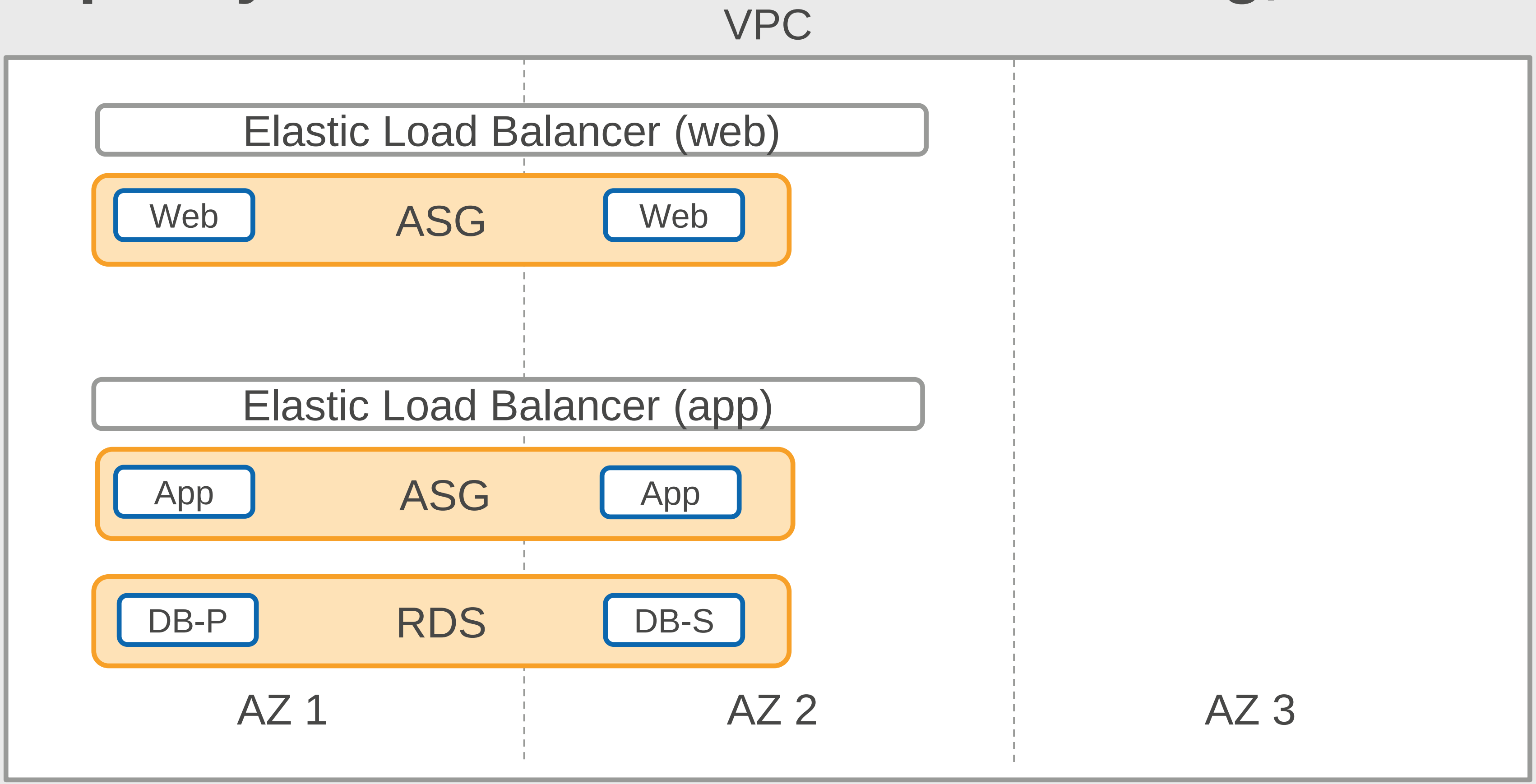
With ELB


VPC
Elastic Load Balancer (web)
Web
Public Subnet
Web
Public Subnet
Elastic Load Balancer (app)
App
DB-P
Private Subnet
App
DB-S
Private Subnet
AZ 1
AZ 2
AZ 3

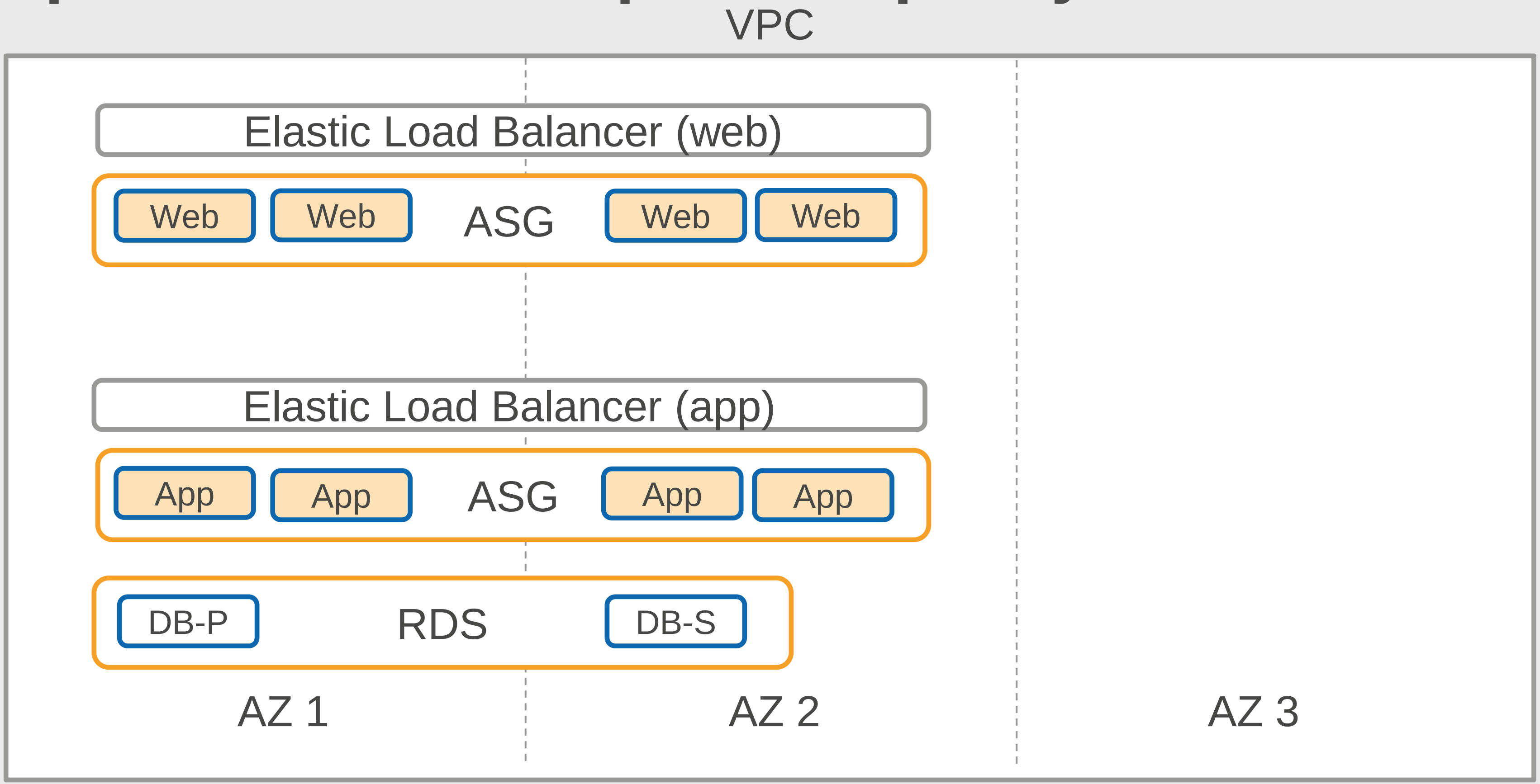
Database


VPC
Elastic Load Balancer (web)
Web
Web
Public Subnet
Public Subnet
Public Subnet
Elastic Load Balancer (app)
App
DB-P
App
DB-S
Database standby AZ
Private Subnet
Private Subnet
Private Subnet
AZ 1
AZ 2
AZ 3

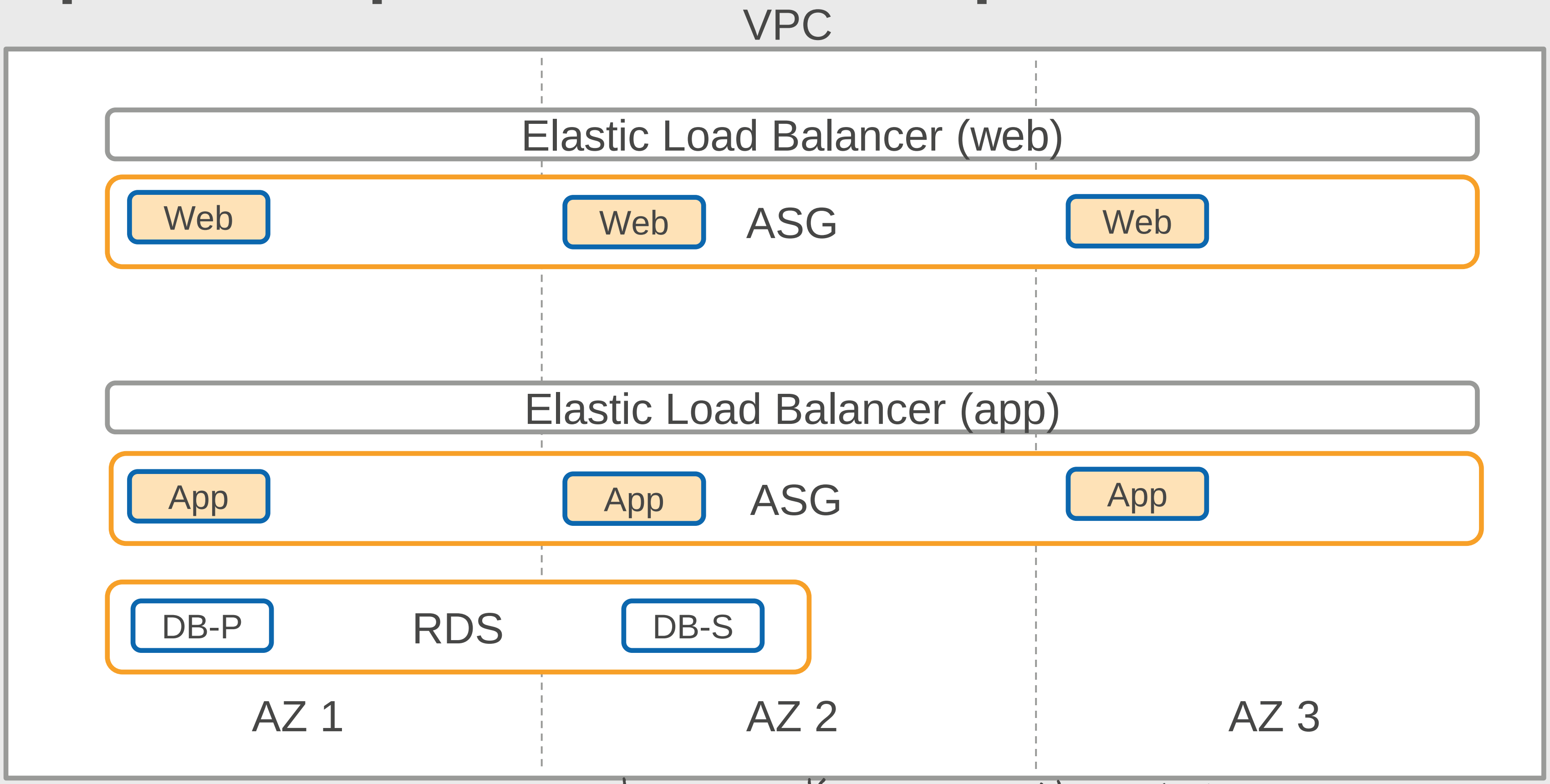
Capacity Maintenance with AutoScaling, RDS



Option 1: 2X the required capacity



Option 2: Spread across multiple AZs



Fewer resources by spreading across three AZs

Scaling Policies

- Dynamic
- Maintain
- Scheduled

Dynamic

- Target tracking
- Step scaling
- Simple scaling

Target Tracking

Specify a scaling metric and target utilization

AWS creates alarm and adjusts server count to maintain target utilization

Step Scaling

You need to create alarm and specify what action to take at each step

Continuously monitors for new breaches and responds

Simple Scaling

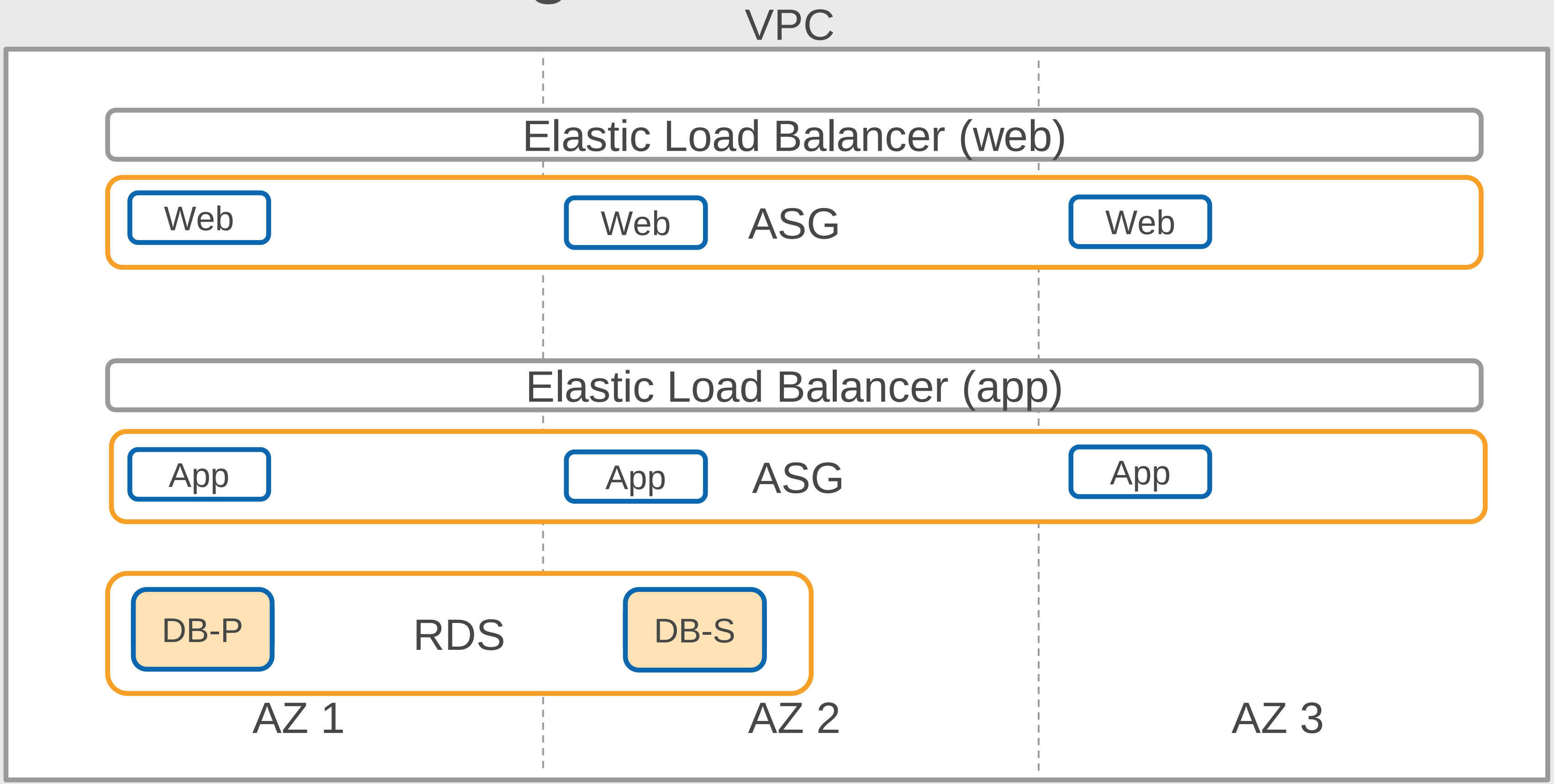
You need to create alarm and specify what action to take

After every scaling action, policy pauses for cooldown period to expire before taking another scaling action

Dynamic

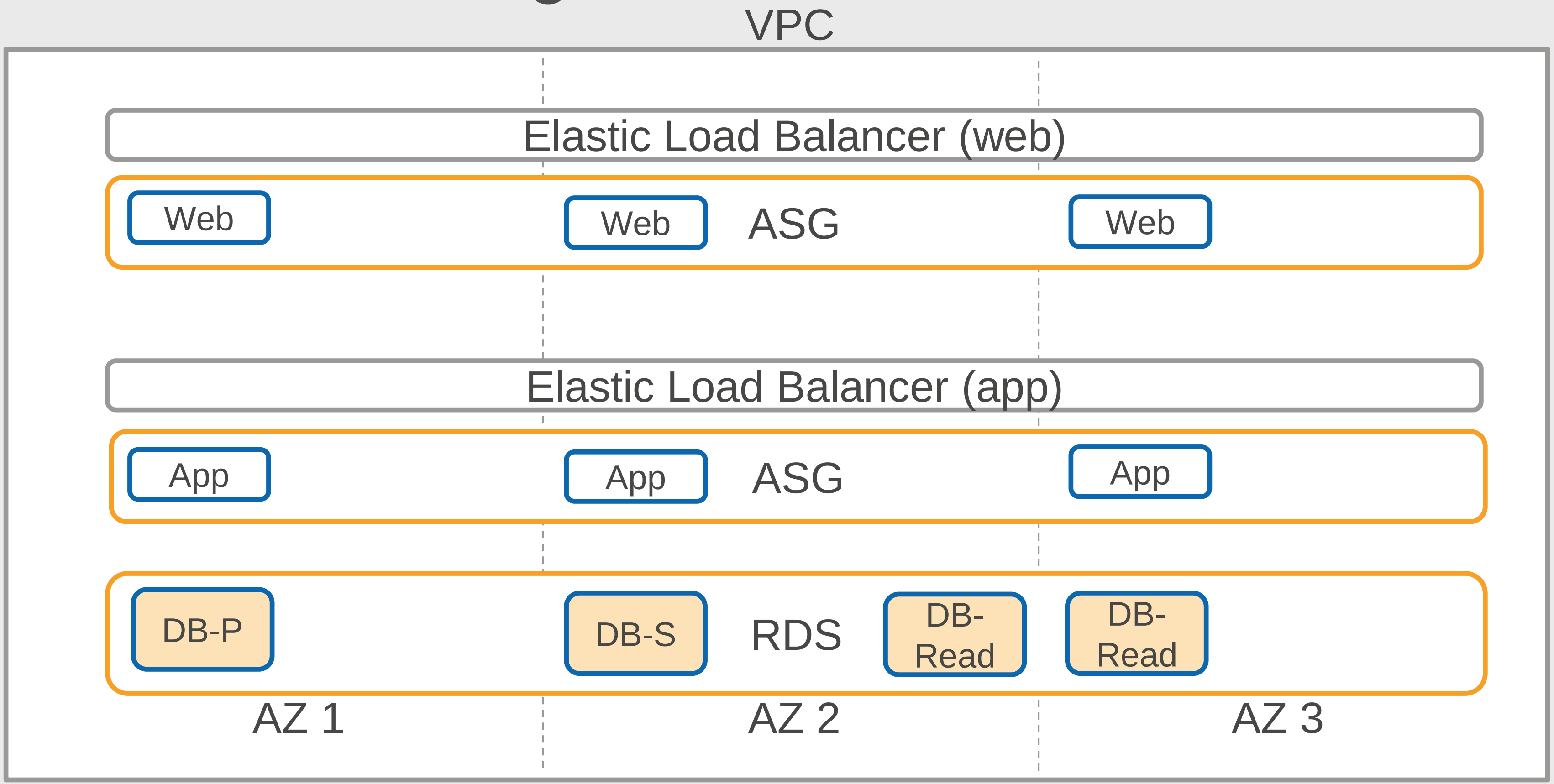
- Target tracking
- Step scaling
- Simple scaling

Database Scaling – Read/Write Bottleneck



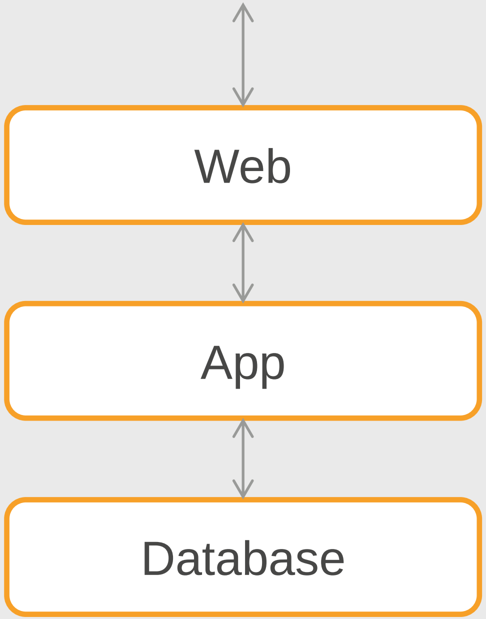
DB instances – scaleup or scale down

Database Scaling – Offload Reads



Database Read Replicas – offload read traffic

Application for Online Orders



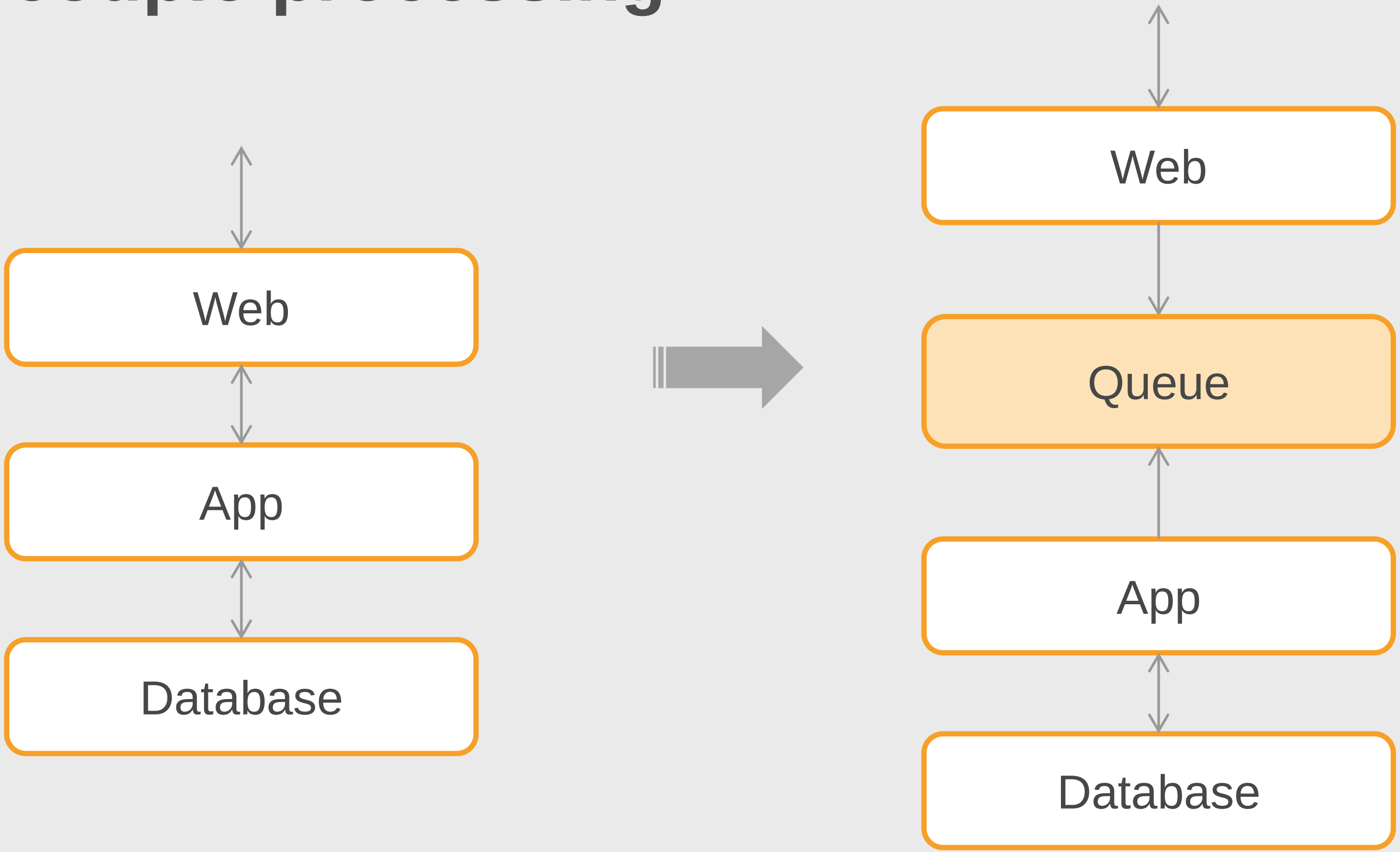
- All layers are tightly coupled
- Need to scale all layers to handle changes in traffic
- Issue in one-layer impacts other layers

Example

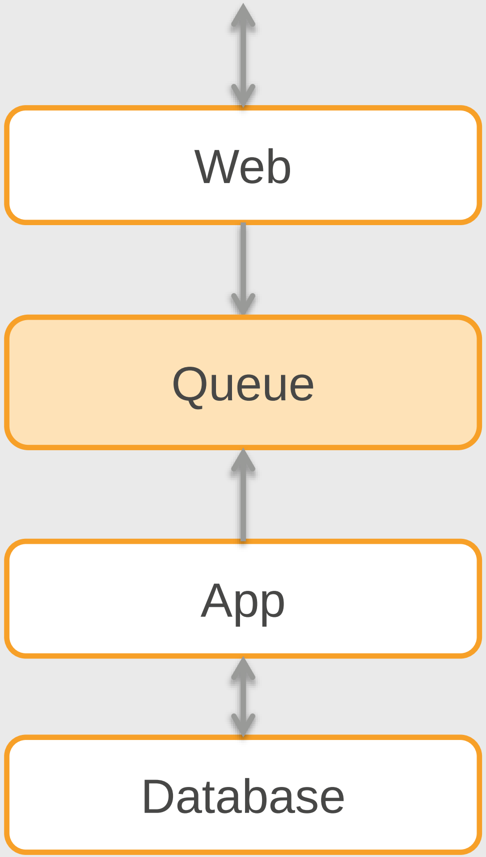
During DB failover, web and app layer are impacted

A burst of new orders can overwhelm app layer and DB layer

Decouple processing

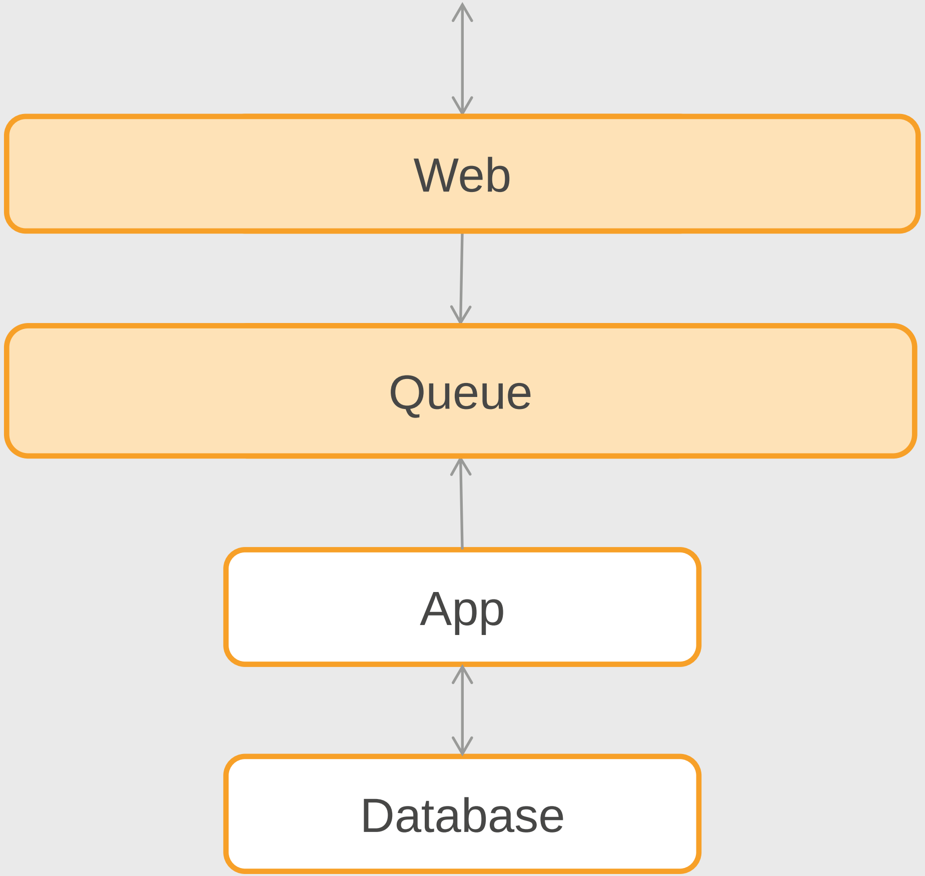


Decouple Layers using a Queue



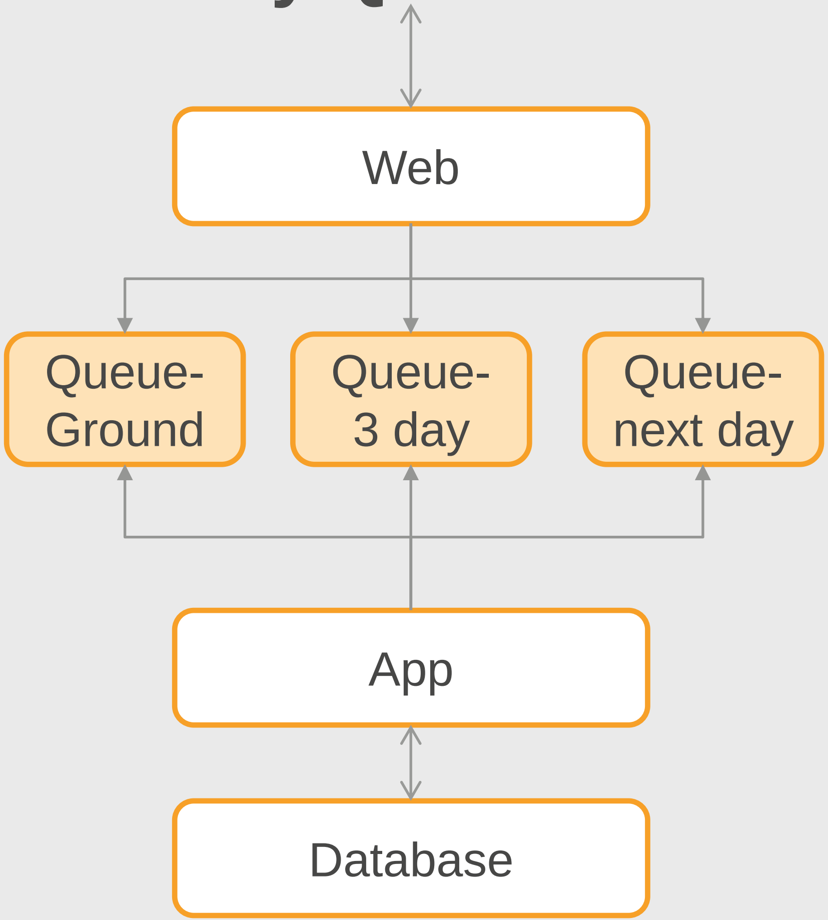
- Web layer accepts orders and stores them safely in an SQS queue
- Customer is acknowledged that order was accepted
- App layer processes items in the queue
- Queue buffers spikes in orders - shielding App layer
- Order can be accepted even during database failover event

Decouple Layers using a Queue - Scaling



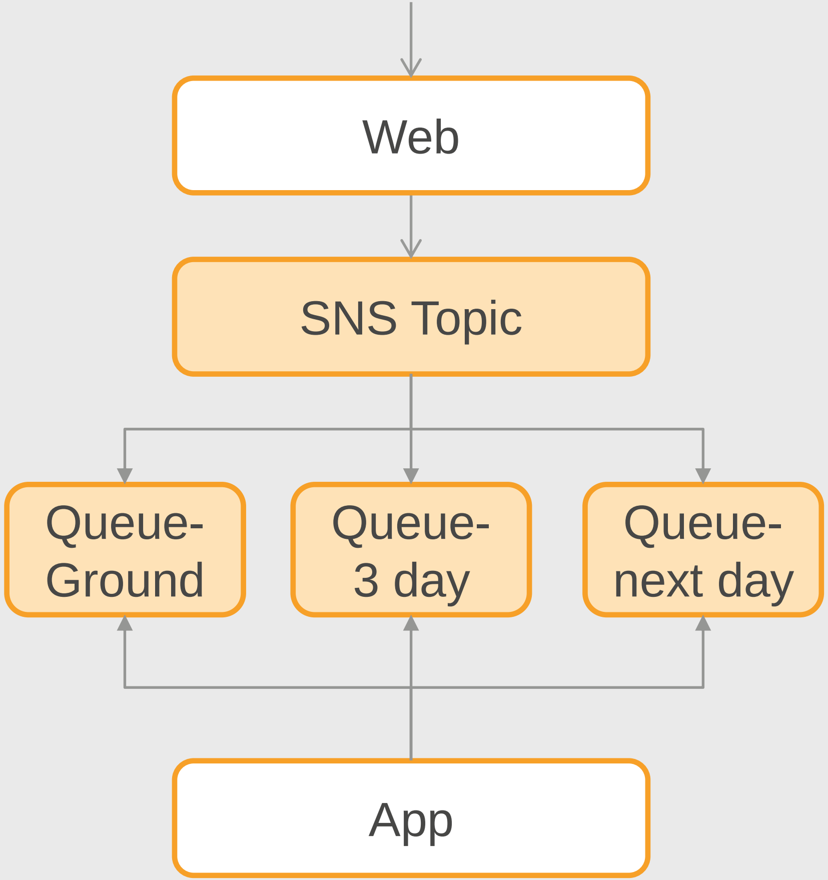
- Scale web layer to handle traffic increase
- SQS Queue automatically scales
- SQS redundantly stores data across multiple availability zones
- App layer can scale based on pending queue items and/or limit imposed by database

Priority Queue



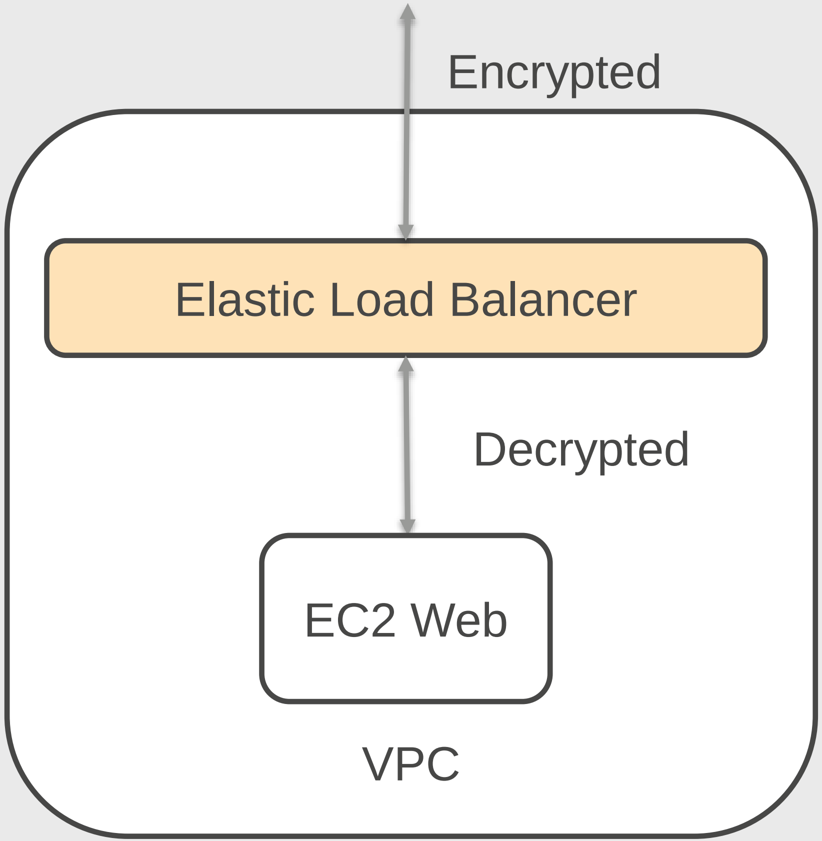
- Order is placed in appropriate queue
- App layer processes order based on shipping priority

Priority Queue using SNS Fanout



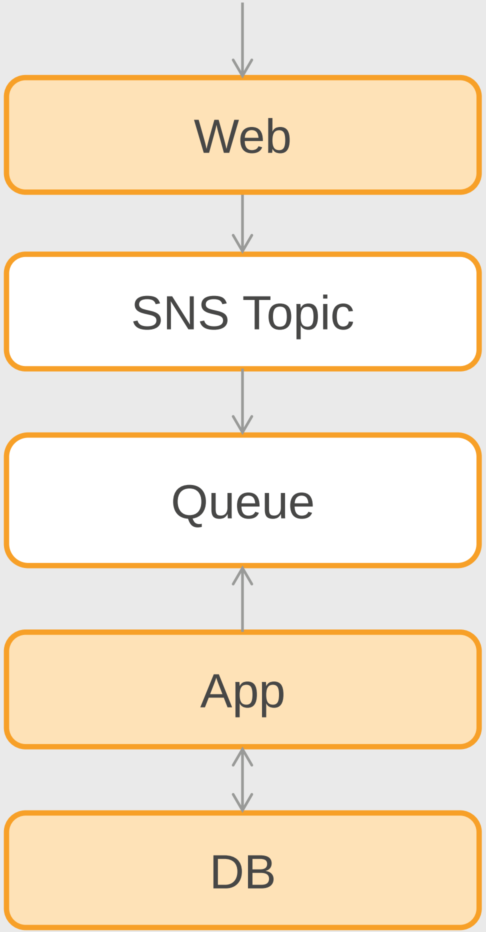
- SNS broadcasts message to all subscribers
- Filter message in SNS - Implement Priority Queues
- Order is added to appropriate queue based on shipping priority
- App layer supports long running workflows

Elastic Load Balancing – Security



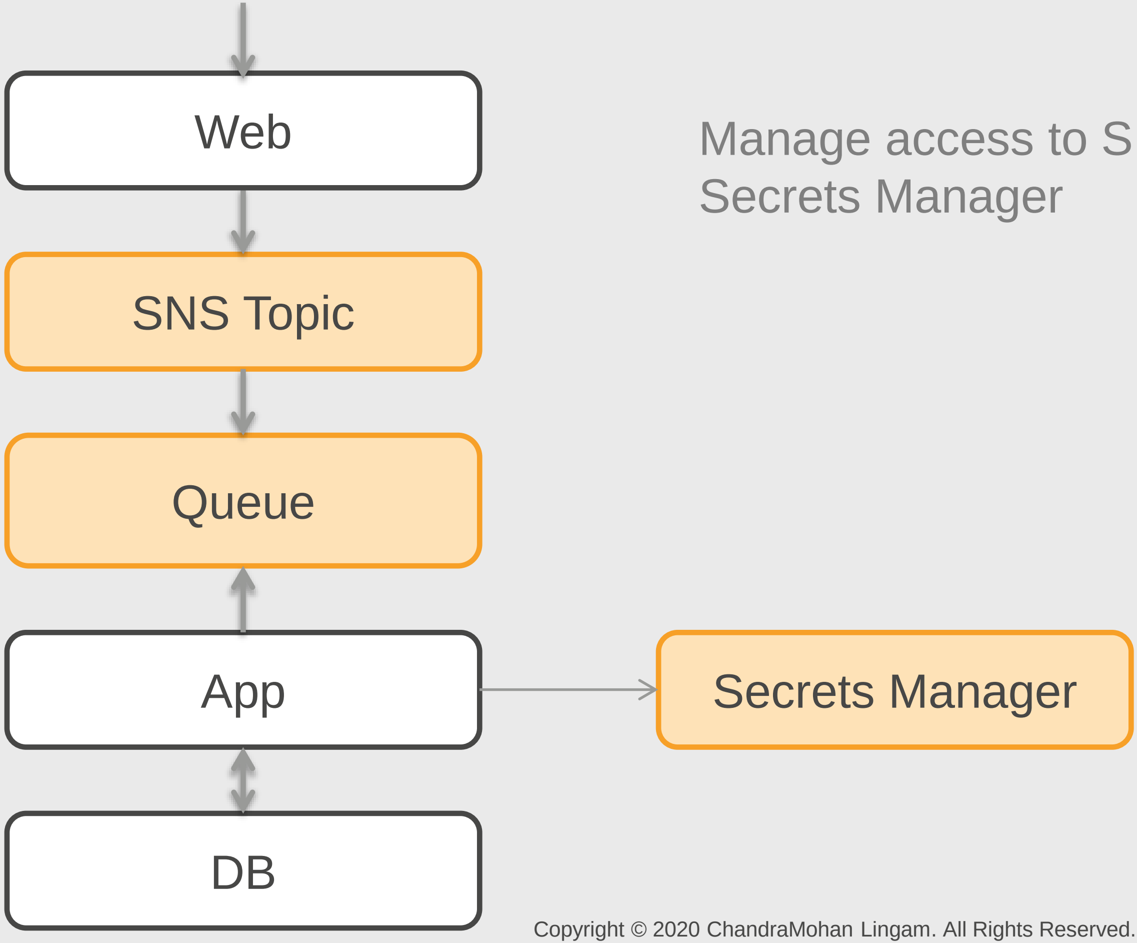
- Offload SSL/TLS at ELB
- Integrated Certificate Management
- User Authentication – Cognito (Application Load Balancer)
  - Internet Identity Providers
  - SAML
  - OpenID Connect
  - Cognito User pools
- WAF with ALB

Security Group and NACL



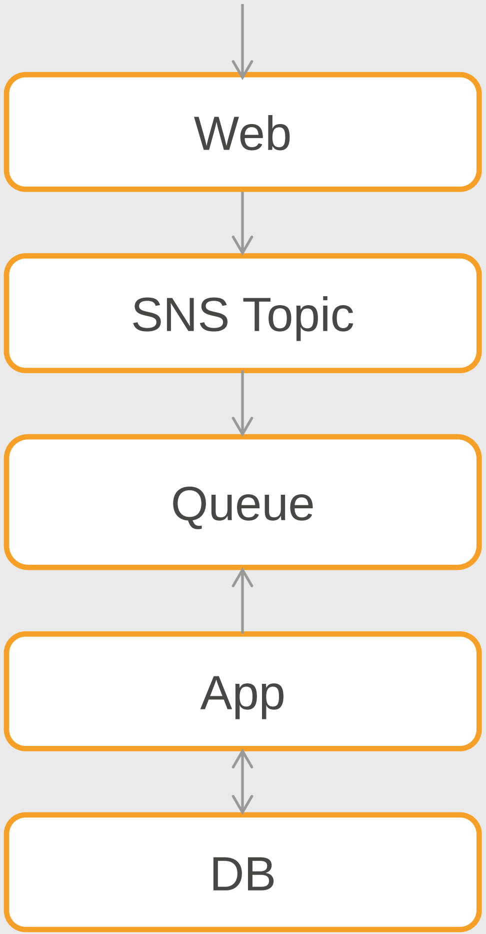
Control network traffic to VPC and Servers

IAM Roles



Manage access to SNS, SQS and Secrets Manager

Continuous Monitoring and Protection



- Configuration drift - Config
- Server vulnerabilities - Inspector
- AWS best practices - Trusted Advisor
- Patching - Systems Manager
- Monitoring - CloudWatch, CloudWatch Log
- Audit trail - CloudTrail

Cost

EC2:

- Hourly (on-demand, reserved, spot, scheduled)
- Data Transfer and Storage

ELB – Hourly, number of load balancer compute units (LCU)

RDS – Hourly, storage, backup, data transfer out

You need to pay hourly charges even if your application is idling

Cost

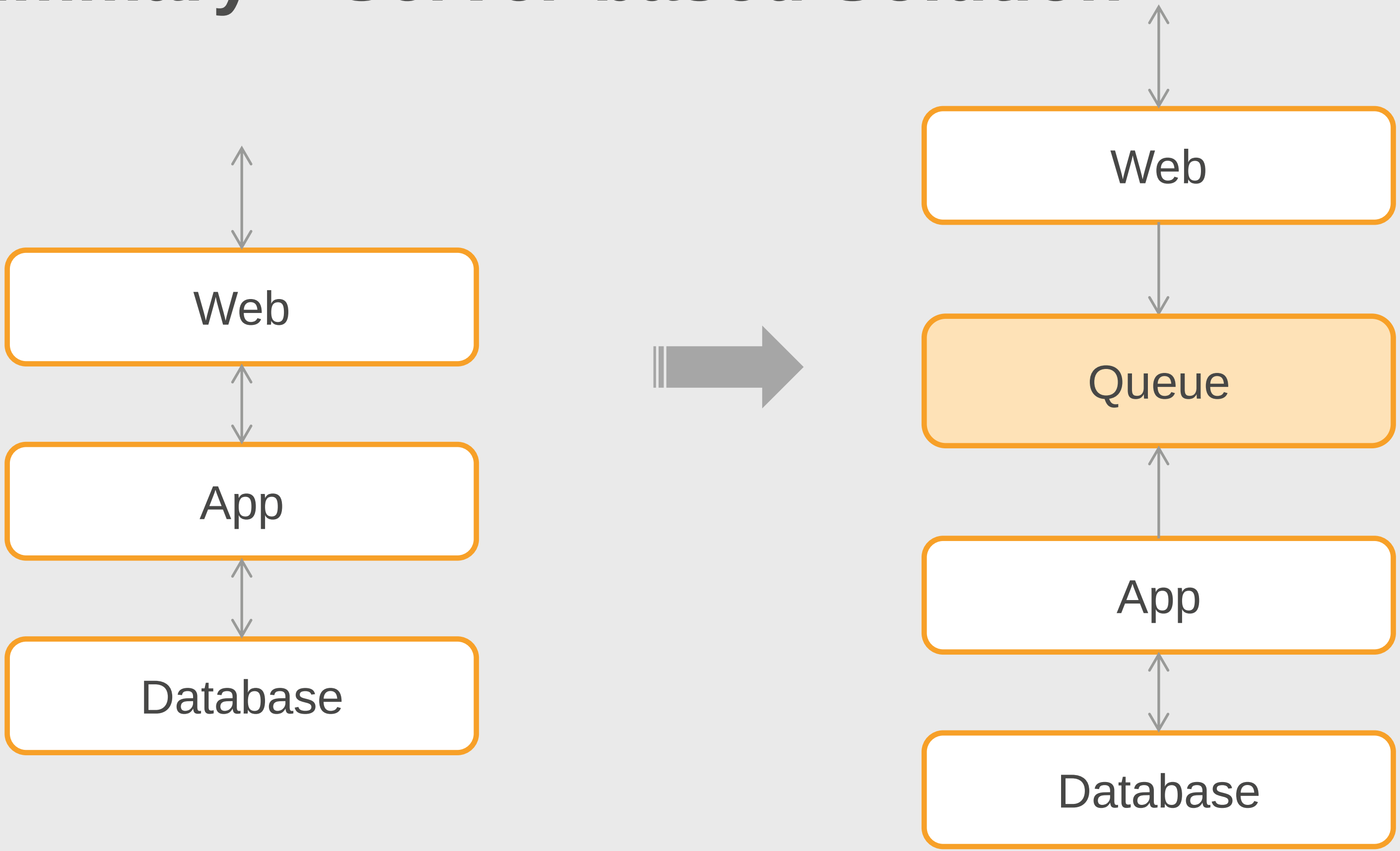
SNS:

- No. of requests (in 64 KB chunks)
- One API call with 256 KB payload is counted as 4 requests

SQS:

- No. of requests (in 64 KB chunks)
- One API call with 256 KB payload is counted as 4 requests
- Data transfer out

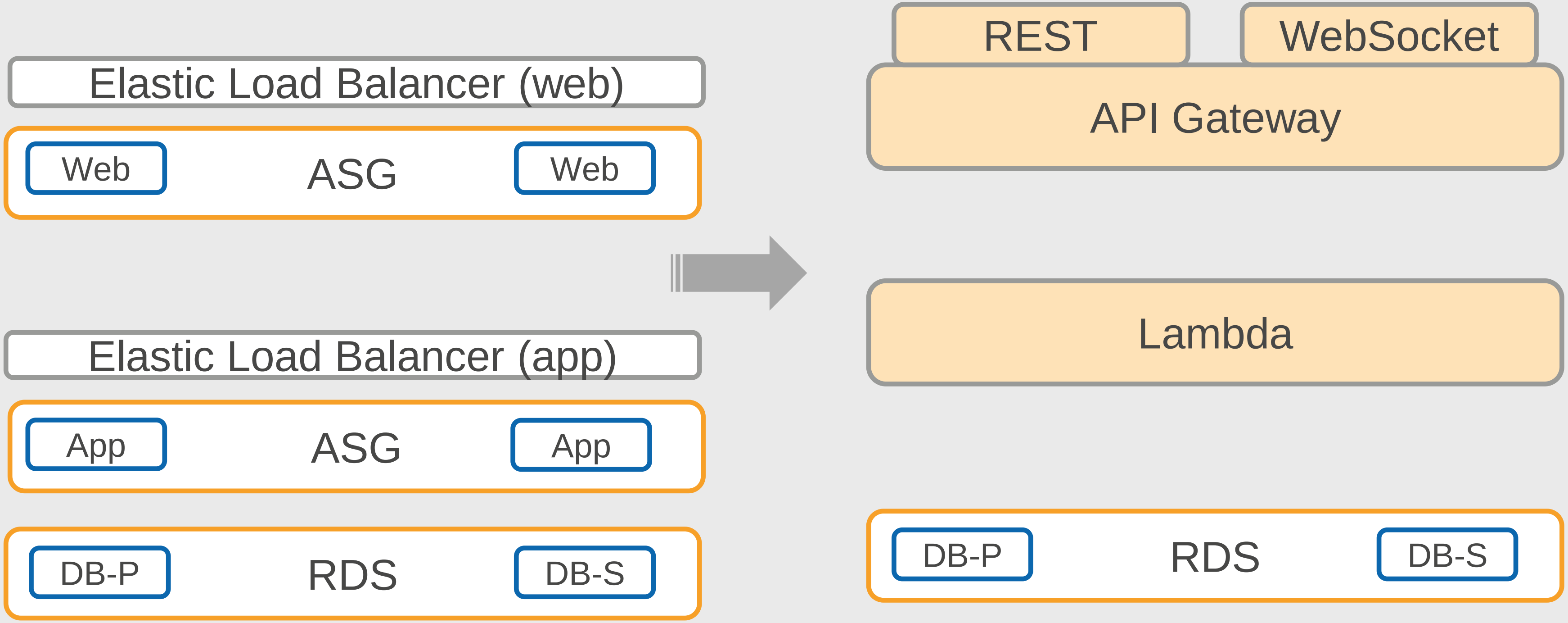
Summary – Server-based Solution



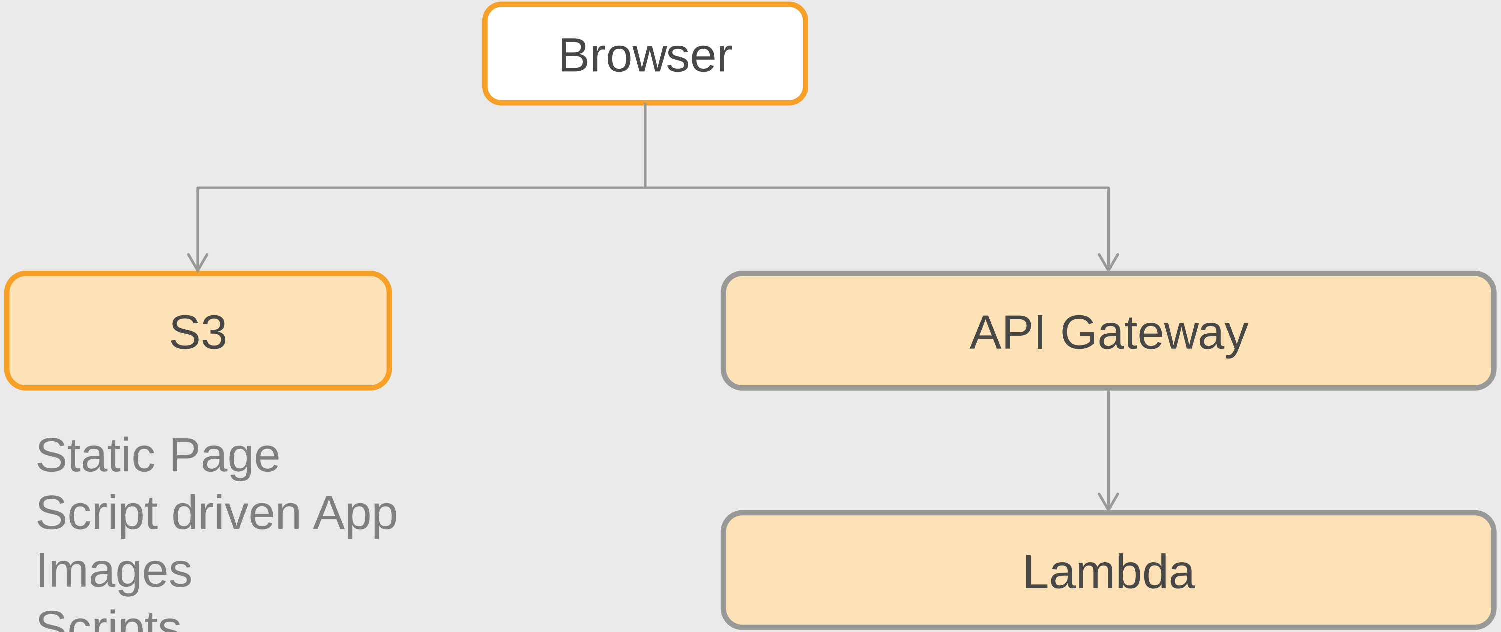
Resiliency, Scaling, Security, Cost

Serverless Implementation

Serverless Implementation

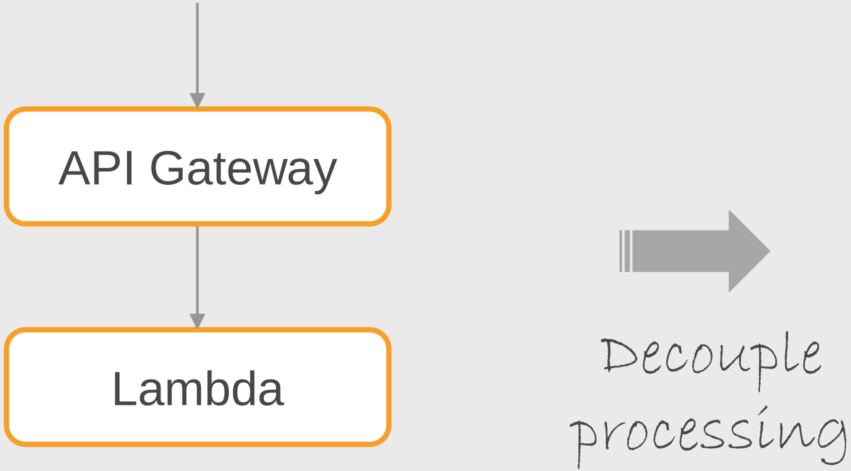


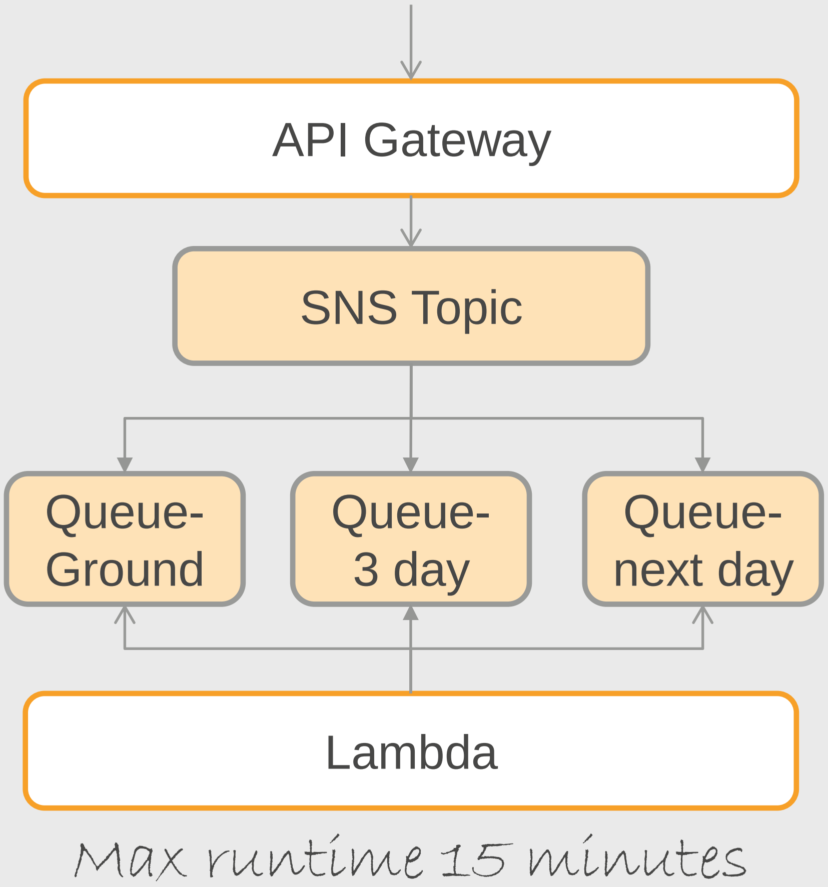
Serverless Implementation



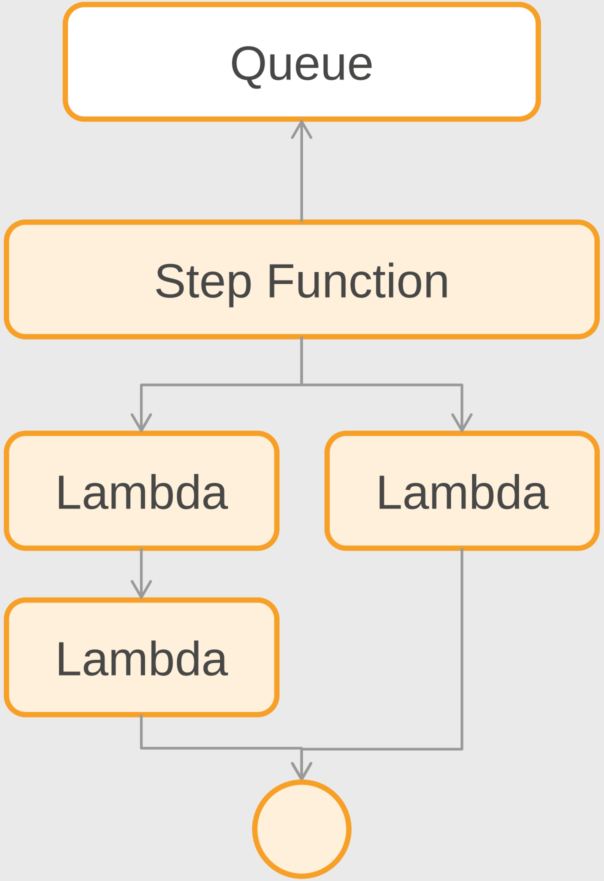
- Static Page
- Script driven App
- Images
- Scripts

Serverless - Decoupling





App Layer with Step Function



Step Function – Orchestrate workflows

Stich together services (Lambda, Containers, DynamoDB, SNS, SQS,..)

Workflow is made up of steps – with each step acting as input to the next

Sequential, Parallel, Branching, Error-handling Steps

Support for long running workflows (up to 1 year)

Serverless - Resiliency

API Gateway, Lambda, Step Functions, SNS, SQS

- Multi-AZ

SNS, SQS:

- Redundant copies of a message are stored across multiple AZs

API Gateway Scaling

- Default scaling up to 10,000 requests/second
- Throttle requests
  - Limit requests/second by API keys
  - Usage Plan based on API Keys
  - Customize based on your requirement

Lambda Scaling

Allocate Memory (and proportional compute is provided)

Initial burst 500 concurrency (number of lambda function instances)

Scale by an additional 500 instances per minute (up to concurrency limit)

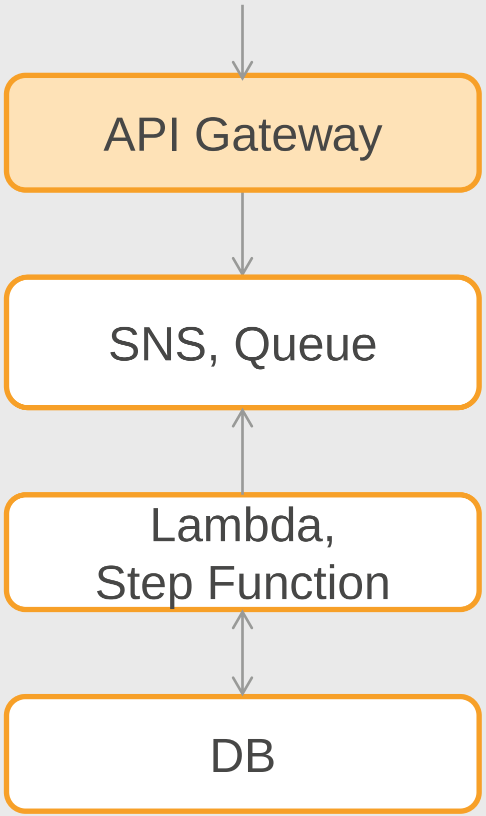
Default concurrency limit is 1000 instances

Lambda Scaling

Reserved concurrency – how much concurrency to allocate for a function. For example, limit app tier to 100 instances

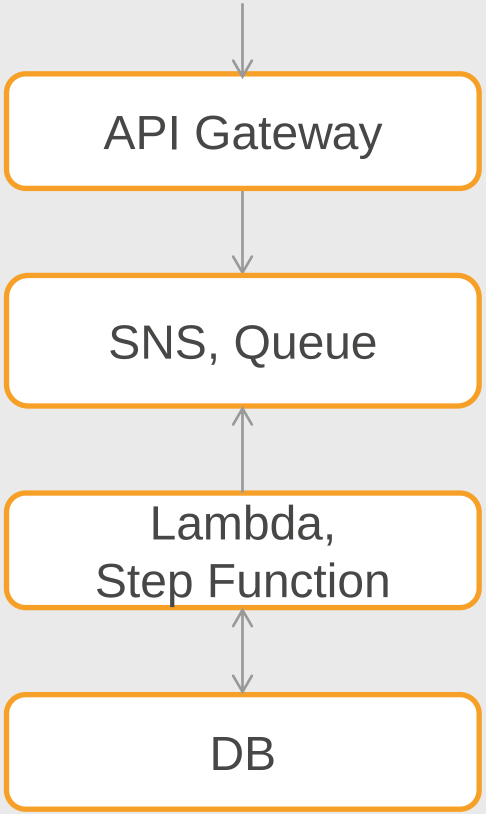
Retries and Dead-Letter-Queue (async invocation)

Security and Protection



- TLS/SSL encryption at API Gateway
- User authentication and authorization – API Gateway
- IAM, Cognito, OAuth, Lambda Authorizer
- Protection against vulnerabilities – WAF for API Gateway

Security and Protection



- Service access - IAM Roles
- Logging - CloudWatch Log
- Monitoring - CloudWatch
- Audit trail – CloudTrail
- Configuration drift - Config
- AWS best practice - Trusted Advisor

No need to use Systems Manager or Inspector

Cost

API Gateway:

- Number of requests
- Amount of data transferred out

Lambda:

- Number of requests
- Duration
- Memory allocated

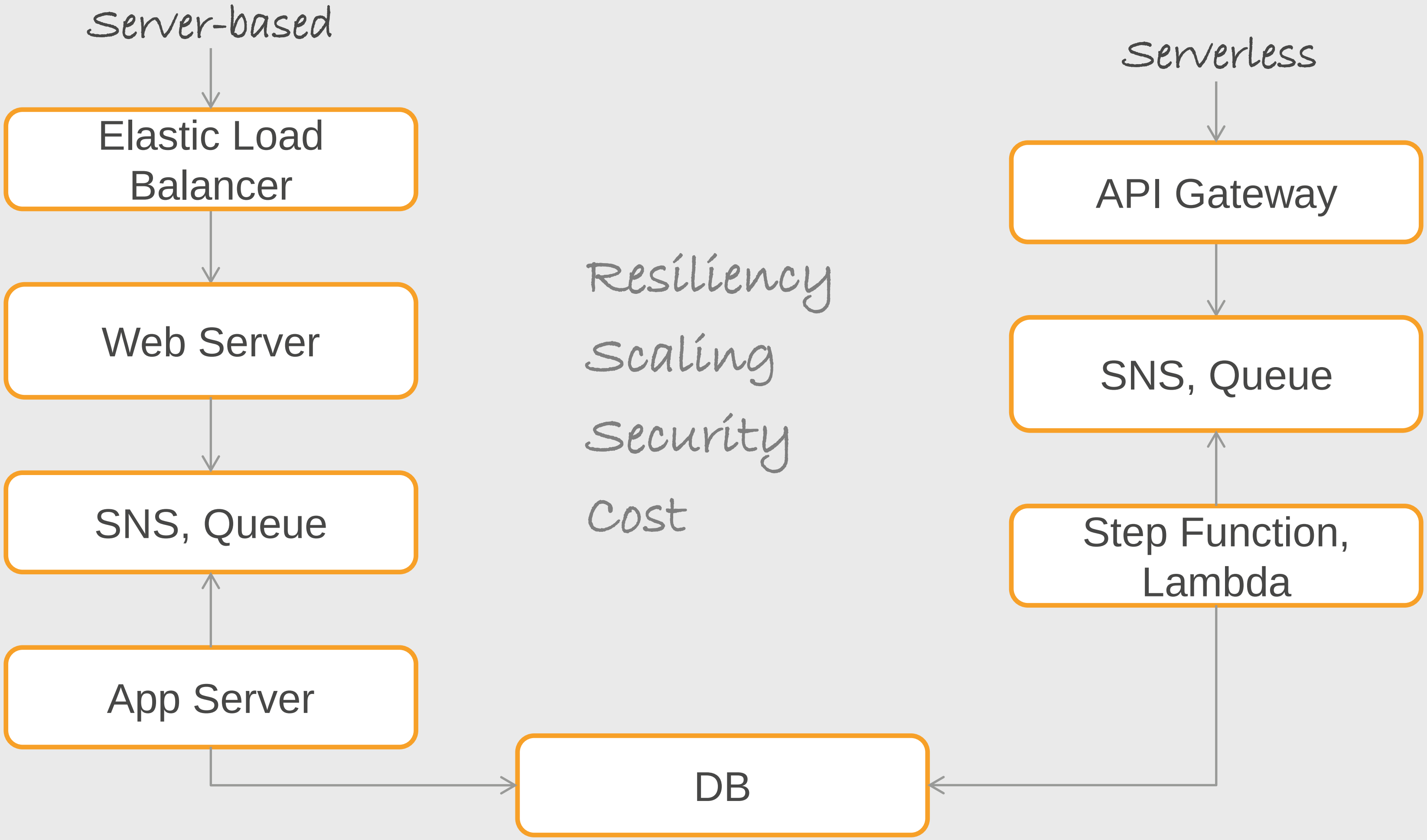
Built-in resiliency, scaling and no need to pay for idle infrastructure

Global

Route 53 – Custom domain and routing

Global Accelerator – region-specific origins, traffic flow control

CloudFront – Edge Caching



Server-based

Elastic Load Balancer

Web Server

SNS, Queue

App Server

Resiliency

Scaling

Security

Cost

Serverless

API Gateway

SNS, Queue

Step Function, Lambda

DB

Chandra Lingam

57,000+ Students

For AWS self-paced video courses, visit:

https://www.cloudwavetraining.com/

Cloud Wave LLC